Analysis of Various Self-Supervised Learning Models for Automatic Pronunciation Assessment

Haeyoung Lee* Sunhee Kim*† and Minhwa Chung*‡
* Interdisciplinary Program in Cognitive Science, Seoul National University, Seoul, Republic of Korea
† Department of French Language Education, Seoul National University, Seoul, Republic of Korea
‡ Department of Linguistics, Seoul National University, Seoul, Republic of Korea
E-mail: {haeylee, sunhkim, mchung}@snu.ac.kr

***Abstract*—The advancement of Automatic Pronunciation Assessment (APA) systems has been significantly improved by Self-supervised Learning (SSL) models. However, despite these performance gains, there remains a lack of systematic research on effective utilization of SSL models and the explainability of their behavior in APA. This study aims to evaluate pronunciation with high accuracy using SSL models and to provide explanations for the scoring outcomes. To achieve this, we fine-tune various SSL models using multiple strategies, comparing their performance through extrinsic analysis to identify the key factors influencing performance improvements. Furthermore, intrinsic analysis is conducted using Principal Component Analysis (PCA) to gain insights into the model's scoring patterns. Extrinsic analysis highlights the importance of strategic fine-tuning and acoustic similarity between fine-tuning and pre-training datasets. Intrinsic analysis reveals that different SSL models focus on distinct pronunciation features, with the Wav2Vec2.0 model capturing more advantageous information for APA. This study presents the first in-depth analysis of SSL models in APA, proposing a novel intrinsic analysis method based on feature distribution manifolds. We provide model-specific fine-tuning guidelines for APA tasks and recommend appropriate SSL models based on specific pronunciation assessment goals. This research significantly contributes to the future development of explainable APA systems based on SSL models.**

I. Introduction

Recent Automatic Pronunciation Assessment (APA) research has focused on improving performance using the Speechocean762 dataset [1]. Advancements have been made by incorporating word and sentence-level features alongside traditional phoneme-level features, and by introducing multi-aspect scoring systems that evaluate fluency, prosody, and completeness through parallel structure models that simultaneously predict all scores [2]. Further improvements include hierarchically enhancing the parallel model structure to effectively capture the linguistic hierarchy of pronunciation [3] and implementing L1-L2 aware word-level modeling that reflects Chinese tonal characteristics [4]. Recent integration of Self-Supervised Learning (SSL) model features has significantly enhanced performance [4]–[8]. In terms of the Pearson Correlation Coefficient (PCC) metric, the integration of SSL models has had the most significant impact on enhancing performance [4], highlighting their substantial capabilities for APA. However, research fully exploiting SSL features' potential in APA remains insufficient, with current approaches facing three primary limitations.

Firstly, existing studies have primarily relied on Automatic Speech Recognition (ASR) tasks, overlooking APA-specific characteristics. Traditional APA systems have utilized ASR results to predict pronunciation scores, and SSL-based research has also followed this approach by either fine-tuning models for speech or phoneme recognition tasks [5], [6], or extracting general features without fine-tuning [8]. However, optimizing SSL features for APA requires score-based fine-tuning specific to the APA task, rather than relying on ASR-oriented approaches.

Secondly, recent studies often employ the latest SSL models without comprehensive evaluation. Despite each SSL model's distinctive processing of speech data due to varying training datasets and algorithms, studies have used Wav2Vec2.0 XLSR without comparison [4] or conducted limited experiments with Wav2Vec2.0 and HuBERT [5]. A more thorough approach with diverse pre-trained models and fine-tuning methods is needed to identify the optimal SSL model for APA.

Thirdly, ensuring the reliability of APA systems requires not only accurate evaluation but also the ability to provide justification for assessment scores. Unlike traditional systems with manually designed features [9], SSL models' black-box nature hinders understanding of their evaluation mechanisms. While some studies have analyzed layer-wise representations of ASR fine-tuned models [5] or extracted frame-level features to assess Wav2Vec2.0's phoneme discrimination on non-native datasets [7], these analyses remain relatively superficial. Moreover, combining SSL with other features [4] further complicates identifying performance-influencing elements.

In this study, we address these limitations by conducting diverse experiments to construct a high-accuracy APA model and identifying performance-enhancing features via extrinsic analysis. Motivated by recent NLP studies which analyze word embedding structures [10], [11], we utilize Principal Component Analysis (PCA) for our intrinsic analysis. By examining the feature distribution manifolds constructed from PCA on embedding vectors, our intrinsic analysis provides insights into the model's scoring mechanism.

II. Method

A. Score-based Fine-Tuning of SSL models

This study proposes a score-based fine-tuning approach to optimize SSL models for APA tasks. Fig. 1 provides the

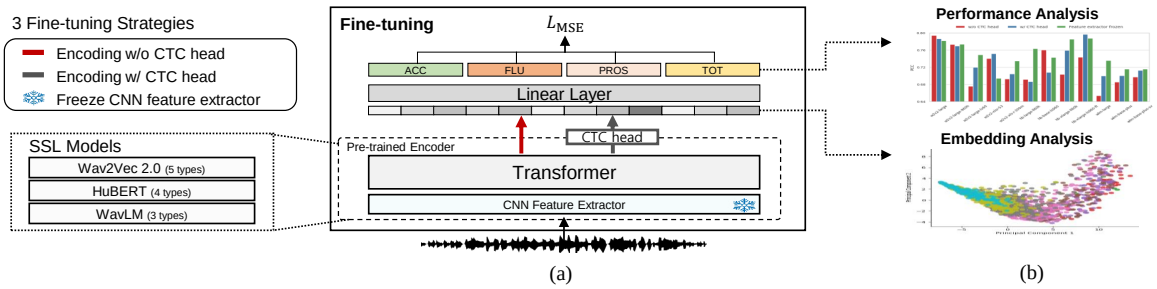

Fig. 1. The overview diagram of the proposed method. (a) the fine-tuning process, (b) the extrinsic and intrinsic analysis of the fine-tuned models.

overview of our method, illustrating (a) the fine-tuning process and (b) the subsequent extrinsic and intrinsic analysis. The fine-tuning process involves adding a linear layer to the pre-trained SSL encoder to output logits for four pronunciation scores, with the model optimized using Mean Squared Error (MSE) loss between predicted and ground truth scores. This approach ensures that the SSL model is optimized for generating scores, thereby providing robust SSL feature-based interpretability.

1) SSL Pre-Trained Models: We select Wav2Vec2.0 and HuBERT for their superior performance in diverse speech tasks [12], and include WavLM for its adaptability to non-ASR tasks. Wav2Vec2.0 utilizes contrastive loss for masked speech prediction, while HuBERT and WavLM employ cross-entropy loss for pseudo-label sequences. Our study evaluates 12 pre-trained models, each trained on distinct unlabeled datasets.

2) Fine-Tuning Strategies: We implement three fine-tuning strategies to optimize the utilization of pre-trained knowledge: a) Freezing CNN feature extractor (FE) parameters to preserve the capabilities of the pre-trained models. b) Adding a Connectionist Temporal Classification (CTC) head to the encoder output to encode information favorable for ASR. c) Extracting general audio features without using a CTC head. The first strategy mitigates catastrophic forgetting [13] by freezing CNN layers. This technique is particularly effective for large-scale datasets like SSL models, but requires validation for APA tasks due to its varied efficacy across different speech processing domains [14]–[16]. The second and third strategies involve introducing a CTC head, which automatically aligns speech and text in ASR tasks, allowing a comparison of its efficacy against general acoustic features for APA. While previous studies have focused on ASR-based phonetic features, prosodic features like rhythm and intonation, shown to be effective in predicting fluency and prosody scores [17], may also be beneficial and require evaluation.

B. Extrinsic and Intrinsic Analysis

This study employs extrinsic probing to compare the APA performance of SSL models under various conditions, identifying optimal adaptation settings for each model. However, models with similar PCC patterns may utilize different intrinsic evaluation factors. For instance, models highly correlated with fluency scores might prioritize prosodic or pitch information differently. Extrinsic analysis alone is insufficient to elucidate these intrinsic mechanisms. Therefore, we propose an intrinsic probing methodology using Principal Component Analysis (PCA), based on the manifold hypothesis [18]. PCA reduces embedding vectors' dimensionality while preserving key variability. Focusing on the last hidden state, presumed to contain critical information for APA, as deep learning models typically handle prediction in the final layer, we apply PCA to reduce high-dimensional embedding vectors. By categorizing these reduced features according to true pronunciation score labels and analyzing their distribution patterns, we gain insights into the inherent operational mechanisms of different model configurations. This dual approach enables a comprehensive understanding of SSL models in APA, revealing both performance outcomes and underlying evaluation factors.

III. Experiments

A. Dataset

This study employs Speechocean762 [1], a benchmark dataset for APA comprising 5,000 English utterances from 250 non-native Mandarin speakers. Five experts manually evaluated each utterance across phoneme, word, and sentence levels. We focus on four utterance-level scores: accuracy, fluency, prosodic, and total, each rated on a 0-10 scale.

B. Experimental Setup

This study examines Wav2Vec2.0, HuBERT, and WavLM, selecting 12 pre-trained variants based on diverse unlabeled datasets. These include Librispeech (LS) [19], comprising 960 hours of clean, clearly pronounced speech, and Libri-light (LL) [19], offering an extensive 60k hours of less refined public domain audiobook data. We also evaluate multilingual models (XLSR [20] and XLS-R [21]), and fine-tune ASR-pretrained models for APA to assess cross-task knowledge transfer. The pre-trained models, sourced from Huggingface [22], are as follows, with the number of parameters in brackets:

- **wav2vec2-large**: 960 hrs of LS [315.4M]
- **wav2vec2-large-960h**: ASR fine-tuned of wav2vec2-large on 960 hrs of LS [315.4M]
- **wav2vec2-large-lv60**: 53k hrs of LL [315.4M]
- **wav2vec2-large-xlsr-53**: 56k hrs of 53 langs [315.4M]
- **wav2vec2-xls-r-300m**: 436k hrs of 128 langs [315.4M]
- **hubert-large-ll60k**: 60k hrs of LL [316.6M]

TABLE I
PCCs OF SSL FINETUNED MODELS. THE BEST MODEL AMONG THE THREE STRATEGIES FOR EACH ASPECT IS INDICATED IN BOLD, AND THE BEST MODEL FOR EACH STRATEGY IS MARKED WITH AN ASTERISK(*).

Finetuned Model	Accuracy			Fluency			Prosodic			Total		
	Train all		Freeze	Train all		Freeze	Train all		Freeze	Train all		Freeze
	w/o CTC	w/ CTC	FE	w/o CTC	w/ CTC	FE	w/o CTC	w/ CTC	FE	w/o CTC	w/ CTC	FE
w2v2-large	0.691	0.688	**0.694**	***0.794**	0.787	0.782	***0.786**	0.785	0.776	**0.728**	0.718	0.723
w2v2-large-960h	*0.706	**0.708**	0.702	0.773	0.770	**0.774**	0.773	0.771	**0.775**	***0.734**	0.729	0.727
w2v2-large-lv60	0.623	**0.666**	0.649	0.676	0.720	**0.749**	0.672	0.730	**0.742**	0.642	**0.686**	0.679
w2v2-xlsr-53	0.678	**0.691**	0.645	0.740	**0.752**	0.694	0.734	**0.751**	0.691	0.694	**0.706**	0.664
w2v2-xls-r-300m	0.633	0.649	**0.661**	0.693	0.705	**0.735**	0.681	0.692	**0.727**	0.647	0.663	**0.679**
hb-large-ll60k	0.620	0.616	**0.698**	0.692	0.687	**0.763**	0.683	0.681	**0.760**	0.633	0.633	**0.716**
hb-base-ls960	0.673	0.626	**0.674**	**0.760**	0.708	0.743	**0.759**	0.693	0.739	**0.704**	0.649	0.698
hb-xlarge-ll60k	0.631	0.686	**0.702**	0.704	0.759	**0.786**	0.693	0.761	**0.783**	0.646	0.705	**0.728**
hb-xlarge-ls960-ft	0.670	*0.719	***0.722**	0.743	***0.797**	*0.788	0.741	***0.788**	*0.784	0.693	*0.734	***0.745**
wlm-large	0.613	0.649	**0.656**	0.654	0.700	**0.736**	0.644	0.695	**0.726**	0.620	0.659	**0.680**
wlm-base-plus	0.603	0.636	**0.653**	0.686	0.701	**0.716**	0.681	0.696	**0.708**	0.632	0.653	**0.673**
wlm-base-plus-sv	0.649	0.641	**0.656**	0.697	0.713	**0.716**	0.687	0.698	**0.714**	0.667	0.664	**0.680**

- **hubert-base-ls960**: 960 hrs of LS [94.3M]
- **hubert-xlarge-ll60k**: 60k hrs of LL [962.5M]
- **hubert-xlarge-ls960-ft**: ASR fine-tuned of hubert-xlarge-ll60k on 960 hrs of LS [962.5M]
- **wavlm-large**: mix 94k hrs data [315.4M]
- **wavlm-base-plus, wavlm-base-plus-sv**: mix 94k hrs data [94.3M] (sv for speaker verification)

Fine-tuning is conducted using the Hugging Face toolkit. For configurations employing a CTC head, we utilize the Hugging Face class to attach the CTC head to the transformer encoder output, encoding each sequence into logits representing probabilities of 32 alphabet tokens through linear transformation. For configurations without a CTC head, the SSL model's encoding approach is used, producing 768- or 1024-dimensional embeddings for base and larger models, respectively. Models are fine-tuned on 2,500 Speechocean762 training samples using MSE loss, with a batch size of 8, for 30 epochs, employing AdamW optimizer with a learning rate of 1e-5. Results are averaged over two runs with different random seeds. Hyperparameter optimization yields 1e-5 as the optimal learning rate. Notably, simultaneous prediction of all four pronunciation scores outperforms single-score prediction.

C. Evaluation Metric

The Pearson Correlation Coefficient (PCC) is used to measure the correlation between model-predicted scores and human-annotated scores. A PCC value close to 1 indicates a strong positive relationship between the model's predictions and the human ratings.

IV. RESULTS

A. Performance Comparison of Fine-Tuned Models

Table I presents the APA performance (PCC) of 36 fine-tuned models. Wav2Vec2.0 and HuBERT exhibit higher PCC performance compared to WavLM. Specifically, within the Wav2Vec2.0 models, w2v2-large-960h excels in accuracy and total scores, while w2v2-large shows superior performance in fluency and prosodic scores. Among the HuBERT models, hb-xlarge-ls960-ft demonstrates the best PCC performance.

Models trained on Librispeech outperform those trained on the larger but less refined Libri-light dataset. Notably, the Librispeech-trained models w2v2-large and hb-base-ls960 outperform their Libri-light counterparts w2v2-large-lv60 and hb-large-ll60k, even when the Librispeech models have smaller architectures. Conversely, multilingual models trained on diverse linguistic environments generally show lower performance.

TABLE II
EXPLAINED VARIANCE RATIOS OF THE PCS

	Pretrained		Finetuned	
	PC1	PC2	PC1	PC2
w2v2-large	0.34	0.13	**0.74**	0.16
hb-large-ll60k	0.07	0.07	**0.85**	0.12
wlm-large	0.08	0.07	**0.83**	0.10

Fine-tuning strategy efficacy varies across models. Freezing the CNN FE yields optimal results for most models. HuBERT, in particular, exhibits consistent performance improvement with increasing model size. Conversely, the Librispeech-trained w2v2-large and hb-base-ls960 perform comparably or better when fully fine-tuned, outperforming their Libri-light-trained counterparts, w2v2-large-lv60 and hb-large-ll60k.

CTC head integration generally enhances performance, but with model-specific variations. Notably, the impact of CTC head integration is significantly influenced by dataset differences. Wav2Vec2.0 and HuBERT exhibit contrasting patterns: multilingual Wav2Vec2.0 models and the Libri-light-trained w2v2-large-lv60 benefit from the CTC head. In Contrast, the Libri-light-trained hb-large-ll60k performs poorly regardless of the CTC head, while the Librispeech-trained hb-base-ls960 shows improved performance without it. This suggests that HuBERT may not fully leverage the benefits of the CTC head under certain conditions. Furthermore, the Librispeech-trained w2v2-large and the ASR-fine-tuned w2v2-large-960h demonstrate excellent performance irrespective of CTC head integration. Conversely, the HuBERT x-large model, known for its superior ASR performance, shows significant improvement when incorporating the CTC head.

B. PCA-based Analysis of SSL Model Embeddings

Table II presents the explained variance ratios of the principal components (PCs) from PCA applied to the final hidden states of large-sized models encoded without a CTC head

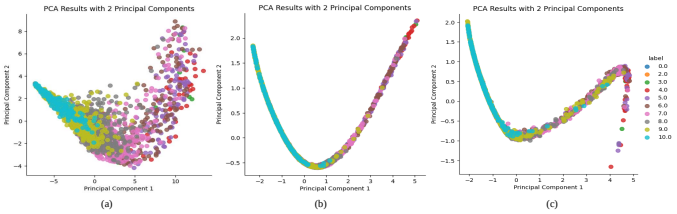

Fig. 2. 2D PCA visualization of (a) Wav2Vec2.0 (b) HuBERT (c) WavLM

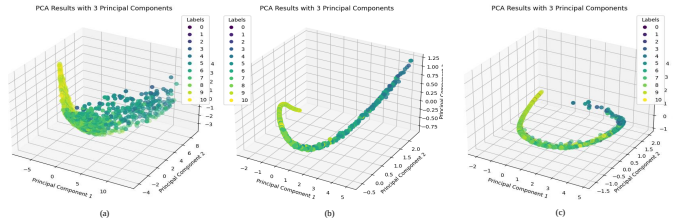

Fig. 3. 3D PCA visualization of (a) Wav2Vec2.0 (b) HuBERT (c) WavLM

and labeled with correct score labels. The explained variance ratios of PC1 for the fine-tuned models correspond to 0.74, 0.85, and 0.83, respectively, whereas for the non-fine-tuned models, they are 0.34, 0.07, and 0.08. This indicates that the PC1 alone accounts for approximately 75% to 85% of the primary variability in the original data, suggesting that fine-tuning adjusts the model's feature vectors to be more suitable for pronunciation evaluation.

Figure 2 visualizes the 2D PCA results for fluency scores of the models in Table II, with PC1 and PC2 as axes. Figure 3 adds PC3 for 3D PCA visualization. The manifold shapes vary by model type: (a) Wav2Vec2.0 forms a cone shape, (b) HuBERT forms a V shape, and (c) WavLM forms an S shape. In 2D, HuBERT and WavLM do not show a continuous score distribution, but in 3D, lower scores become more distinct with the newly introduced z-axis.

The most notable model is Wav2Vec2.0, which demonstrates the largest variance in the mid-score range and shows greater variance in the lower score range compared to other models. High scores are sharply modeled, while lower scores tend to be more dispersed, indicating a continuous distribution of scores.

Figure 4 shows PCC results measured for various combinations of principal component (PC) vectors for each model. This analysis aims to verify whether the dimension-reduced PC vectors can accurately predict the actual pronunciation evaluation scores. Remarkably, all three models achieve similar or better performance using only PC1 compared to the original 1024-dimensional vectors. Notably, the HuBERT and WavLM models show improved performance with multiple PCs. Additionally, the PC combinations affecting pronunciation evaluation vary across different aspects for each model.

Figure 5 visualizes the fluency scores for different HuBERT models, categorized by score. The PCC performance for models (a), (b), and (c) are 0.692, 0.763, and 0.797, respectively. Notably, higher-performing HuBERT models exhibit patterns similar to the manifold observed in Wav2Vec2.0 models.

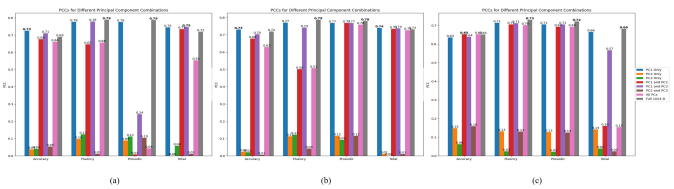

Fig. 4. PCCs for Different PC Combinations of (a) Wav2Vec2.0 (b) HuBERT (c) WavLM

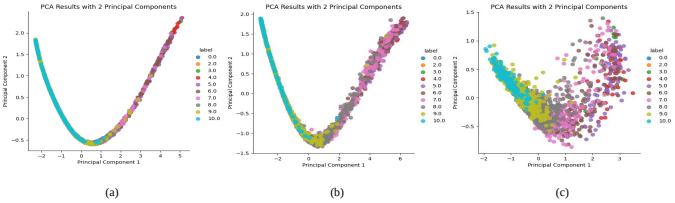

Fig. 5. PCA visualization of (a) hb-large-ll60k w/o CTC (b) hb-large-ll60k with frozen FE (c) hb-xlarge-ls960-ft w/ CTC

V. Discussion

A. Extrinsic Analysis of Factors Contributing to Performance Improvement

Table III compares the performance of our best model, based on fluency and prosodic scores with the latest APA baseline model. On the Speechocean762 dataset, our model achieves PCC values up to 0.048 and 0.037 higher in fluency and prosodic aspects, respectively. Although it does not surpass the state-of-the-art performance reported in Ref. [4], it is noteworthy that simply fine-tuning a pre-trained SSL model can yield performance comparable to the latest pronunciation assessment models.

a) Performance of ASR-adapted settings: ASR-adapted settings perform excellently in APA tasks. Among the three models, Wav2Vec2.0 and HuBERT demonstrate the best performance when fine-tuned for ASR, while the noise-resistant WavLM model showed relatively poorer performance. Additionally, models trained on the cleaner Librispeech dataset outperform those trained on noisier datasets.

b) Impact of Pre-Training Data: The type of data used for pre-training significantly influences APA performance. Models trained on the clean Librispeech dataset outperform those trained on the noisier Libri-light dataset, despite the quantitative disadvantage. Interestingly, the smaller hb-base-ls960 model using Librispeech data outperforms the larger hb-

TABLE III
PERFORMANCE COMPARISON WITH BASELINE

Model	Utterance Score (PCC) ↑			
	Accuracy	Fluency	Prosody	Total
Kim et al. [5]	-	0.780	0.770	-
GOPT [2]	0.714	0.753	0.760	0.742
HiPAMA [3]	0.730	0.749	0.751	0.754
3MH [4]	0.782	0.843	0.836	0.811
w2v2-large	0.691	**0.794**	**0.786**	0.728
hb-xlarge-ls960-ft	0.719	**0.797**	**0.788**	0.734
wlm-large	0.656	0.736	0.726	0.680

large-ll60k model, suggesting that dataset quality has a greater impact on performance than model size.

Moreover, models trained on the clean, English ASR dataset Librispeech outperform those trained on multilingual datasets, particularly for APA tasks involving non-native speakers. This is likely due to the acoustic similarity between the pre-training and fine-tuning data, as demonstrated in Ref. [23], emphasizing the importance of dataset relevance over volume and diversity. Overall, we find that the optimal tuning methods vary depending on the characteristics of the pre-trained models. Therefore, strategically tuning the models while considering their inherent properties significantly impacts APA performance. Below are the details of our discussion.

c) Role of low-level feature extraction: Maintaining the low-level feature extraction ability of pre-trained models is crucial for APA performance. Freezing CNN FE, which are closely related to the unlabeled datasets used during pre-training, often results in superior performance. However, fine-tuning the entire model proves more beneficial for models using Librispeech, aligning with findings in Ref. [13], which suggest that fine-tuning all layers with a low learning rate is effective when the source and target data are acoustically similar.

d) CTC head influence: Learning information between speech and text through a CTC head generally has a positive impact on APA. However, Wav2Vec2.0 and HuBERT models exhibit contrasting improvements depending on the ASR environment. For Wav2Vec2.0, the CTC head is particularly beneficial in less favorable ASR conditions, such as when using the Libri-light or multilingual datasets, which have shown lower performance in English pronunciation assessment. However, it has little impact in more favorable settings like using the Librispeech dataset or ASR fine-tuned models. Conversely, HuBERT struggles even with the CTC head in challenging conditions. Notably, in the hb-xlarge-ll60k model, which demonstrates strong ASR performance [24], incorporating the CTC head significantly enhances performance. This suggests that while HuBERT can learn effectively with the CTC head in favorable conditions, it performs poorly without it, even in these optimal settings. In contrast, Wav2Vec2.0 models perform well without the CTC head in favorable ASR environments, likely due to their effective learning of phoneme-level information during pre-training, as supported by several studies [7], [25]. These findings highlight the different mechanisms by which the two models learn phonetic information, emphasizing that optimal tuning strategies should consider the specific characteristics of each pre-trained model to maximize pronunciation assessment performance.

B. Intrinsic Analysis of Feature Representation

The high variance explained by PC1 in Table II, along with the superior APA performance observed when using PC1 alone in Fig. 4, indicates that SSL models fine-tuned for APA tasks effectively capture key high-dimensional features critical for score prediction. Each SSL model evaluates scores based on these principal features.

As illustrated in Fig. 2 and Fig. 3, the varying manifold shapes in the feature space composed of PCs indicate different pronunciation assessment patterns across models. Specifically: (a) The conical shape of the Wav2Vec2.0 manifold suggests a primary reliance on PC1 for evaluation. (b) The V-shaped manifold of the HuBERT model implies the use of two main criteria. (c) The S-shaped manifold of the WavLM model indicates an assessment based on multiple criteria. Notably, unlike Wav2Vec2.0, HuBERT and WavLM demonstrate a clearer score distribution in three-dimensional space, as evidenced by the performance improvement with additional PCs in Fig. 4. This suggests that while Wav2Vec2.0 primarily relies on a single PC for evaluation, HuBERT and WavLM utilize a broader range of evaluation criteria.

As shown in Fig. 2(a), the Wav2Vec2.0 model uniquely captures a continuous score distribution and exhibits significant variance in the middle score range, unlike the other models. In the Speechocean762 dataset, the middle score range is predominant, and the model's ability to recognize various error patterns within this range underscores its superior performance. Moreover, the data distribution in Wav2Vec2.0, with most data concentrated in the middle and high scores and scarcely any data in the low scores, aligns with the actual distribution of the Speechocean762 dataset. For real-world APA applications, it is crucial that models are well-calibrated to realistic scenarios, which Wav2Vec2.0 accomplishes most effectively. Additionally, Fig. 5 demonstrates that the higher-performing HuBERT model exhibits a manifold similar to that of the Wav2Vec2.0, further emphasizing its excellence. In summary, Wav2Vec2.0 learns the most advantageous information for APA and can be considered the best model for this task.

Different fine-tuning strategies and datasets impact model performance but do not alter the intrinsic features the models learn. Figures 2 and 3 show noticeable differences between models but minimal changes due to dataset or fine-tuning variations. This suggests these factors affect performance without fundamentally changing the intrinsic features.

VI. Conclusion

This study explores the characteristics of SSL models that enhance APA performance through various experiments and explains their performance results. Additionally, we propose an intrinsic analysis methodology that visually confirms the distribution and variability of data via manifold shapes, reflecting the criteria and complexity used in pronunciation assessment. This approach helps in understanding the intrinsic operational mechanisms of each model. Future research should aim to analyze PC vectors in detail to clarify the features they represent, as the exact scoring criteria of SSL models were not clearly identified. The diverse analytical results of this study offer crucial guidelines for utilizing SSL models in APA and enhance understanding of the previously opaque SSL models, thereby contributing to the development of explainable SSL model-based pronunciation assessment models in the future.

References

[1] J. Zhang, Z. Zhang, Y. Wang, *et al.*, "Speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[2] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7262–7266.

[3] H. Do, Y. Kim, and G. G. Lee, "Hierarchical pronunciation assessment with multi-aspect attention," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[4] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "A hierarchical context-aware modeling approach for multi-aspect and multi-granular pronunciation assessment," *arXiv preprint arXiv:2305.18146*, 2023.

[5] 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.

[6] A. Zahran, A. Fahmy, K. Wassif, and H. Bayomi, "Fine-tuning self-supervised learning models for end-to-end pronunciation scoring," *IEEE Access*, 2023.

[7] W. Liu, K. Fu, X. Tian, *et al.*, "An asr-free fluency scoring approach with self-supervised learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[8] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "3m: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, pp. 575–582.

[9] 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.

[10] R. Bommasani, K. Davis, and C. Cardie, "Interpreting pretrained contextualized representations via reductions to static embeddings," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4758–4781.

[11] T. Musil, "Examining structure of word embeddings with pca," in *Text, Speech, and Dialogue: 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings 22*, Springer, 2019, pp. 211–223.

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

[13] M. Iman, H. R. Arabnia, and K. Rasheed, "A review of deep transfer learning and recent advancements," *Technologies*, vol. 11, no. 2, p. 40, 2023.

[14] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[15] Y. Gao, C. Chu, and T. Kawahara, "Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with asr and gender pretraining," in *Proc. Interspeech*, 2023.

[16] H. Aronowitz, I. Gat, E. Morais, W. Zhu, and R. Hoory, "Towards a common speech analysis engine," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 8162–8166.

[17] Y. E. Kheir, A. Ali, and S. A. Chowdhury, "Automatic pronunciation assessment–a review," *arXiv preprint arXiv:2310.13974*, 2023.

[18] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[19] J. Kahn, M. Rivière, W. Zheng, *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7669–7673.

[20] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[21] A. Babu, C. Wang, A. Tjandra, *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[22] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[23] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, *et al.*, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023.

[24] 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.

[25] 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.