



Multimodal and Multitask Learning for Predicting Multiple Scores in L2 English Speech

Sehyun Oh¹, Sunhee Kim², Minhwa Chung^{1,3}

¹Interdisciplinary Program in AI, Seoul National University, South Korea

²Department of French Language Education, Seoul National University, South Korea

³Department of Linguistics, Seoul National University, South Korea

ohsehyun12@snu.ac.kr, sunhkim@snu.ac.kr, mchung@snu.ac.kr

Abstract

This study presents a novel multimodal and multitask learning model for predicting five proficiency scores of L2 English speeches. The proposed approach integrates speech and text embeddings using multimodal transformer blocks with cross-modal attention to refine features dynamically between modalities, capturing complementary information. A joint loss function, combining MSE and a Trait-Aware (TA) loss, enhances the model by leveraging relationships among proficiency traits. Experiments with different combinations of four embeddings (MFCCs, GloVe, wav2vec 2.0, and BERT) revealed that the proposed model with wav2vec 2.0 and BERT embeddings achieved the best performance, with a mean PCC of 0.734 and a standard deviation of 0.0129 across five criteria. This approach significantly outperforms unimodal and baseline multimodal models, demonstrating the potential of advanced multimodal architectures and task-aware optimization in automated speech assessment systems.

Index Terms: Automatic speech assessment, multitask learning, non-native spontaneous speech, sub-level speech scores

1. Introduction

Automated Spoken Language Assessment (SLA) has become an increasingly important component of Computer-Assisted Language Learning (CALL) systems. Its primary objective is to provide an automated evaluation of oral proficiency, producing a proficiency score that is closely aligned with those given by expert human raters. Traditional approaches relied heavily on the extraction of hand-crafted features derived from methods such as natural language processing, prosodic analysis, and signal processing [1, 2]. The development of Automated Speech Recognition (ASR) systems integrated with feature extraction methods has been widely adopted for speech scoring tasks. For example, systems like SpeechRaterSM from the Educational Testing Service® (ETS) [2] utilized a combination of signal processing techniques and prosodic analysis to extract features like fluency, pronunciation, and vocabulary use, which were then fed into statistical models to predict proficiency scores [2, 3, 4].

The advent of deep learning has shifted the paradigm toward end-to-end systems that learn predictive features directly from raw data, demonstrating superior performance over conventional methods. Researchers developed Context-Dependent Deep Neural Network and Hidden Markov Models (CD-DNN-HMMs), significantly enhancing ASR performance and enabling more accurate assessments of non-native speech [5, 6]. Subsequently, researchers have used DNNs for ASR and speech scoring tasks, demonstrating improvements in pronunciation evaluation over Gaussian mixture models and other tradi-

tional approaches [7, 8, 9]. Studies in [10, 11] employed advanced neural networks such as Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory Recurrent Neural Networks (BLSTM) to capture temporal dependencies and local contexts in speech data. The approach in [12] utilized an attention-based BLSTM mechanism, enabling models to leverage the utterance history from the speech to improve scoring accuracy for long speeches. More recently, with the advent of transformers [13], researchers have utilized pre-trained models such as wav2vec 2.0 [14], BERT [15], or HuBERT [16] to develop more accurate systems for evaluating the pronunciation, fluency, and comprehensibility of non-native speeches [17, 18, 19].

Despite these advancements, most research has focused on predicting a single sub-level proficiency score at a time. These approaches, while effective in predicting a single score, may overlook the interdependent aspects of spoken language. Only a few studies have introduced multi-task learning (MTL) in speech scoring to predict several scores from speech. MTL allows a single model to optimize multiple tasks simultaneously, leveraging shared representations to enhance performance. Using MTL for spoken language assessment, a study showed the efficacy of jointly learning from each sub-level pronunciation score [20]. Furthermore, [21] presented a model that used speech and text embeddings from pre-trained models to predict four scores simultaneously and demonstrated the effectiveness of the MTL approach, achieving an average Pearson correlation coefficient (PCC) of 0.720. However, this model from [21] demonstrated relatively higher PCC scores on two evaluation criteria, while showing lower PCC scores for the other two criteria. This resulted in performance discrepancies across different scoring criteria, suggesting that while MTL helped optimize some traits, it led to inconsistent performance in predicting multiple scores simultaneously.

One possible way to improve performance is by incorporating additional sources of information, such as text embeddings. While some studies have explored this multimodal approach for speech scoring, their effectiveness remains limited by how they integrate speech and text embeddings. Some studies, such as [12, 19, 21], have leveraged multimodal fusion to improve prediction accuracy by simply concatenating extracted embeddings or applying independent processing pipelines for each modality. However, such approaches often fail to capture intricate cross-modal dependencies, as they treat speech and text features as largely separate information. Another study has sought to improve this by incorporating attention mechanisms for fluency scoring [22]. However, among these multimodal approaches, only [21] has attempted to combine both MTL and multimodal learning for predicting multiple scores simultaneously, while the others have focused on single-score prediction. Further-

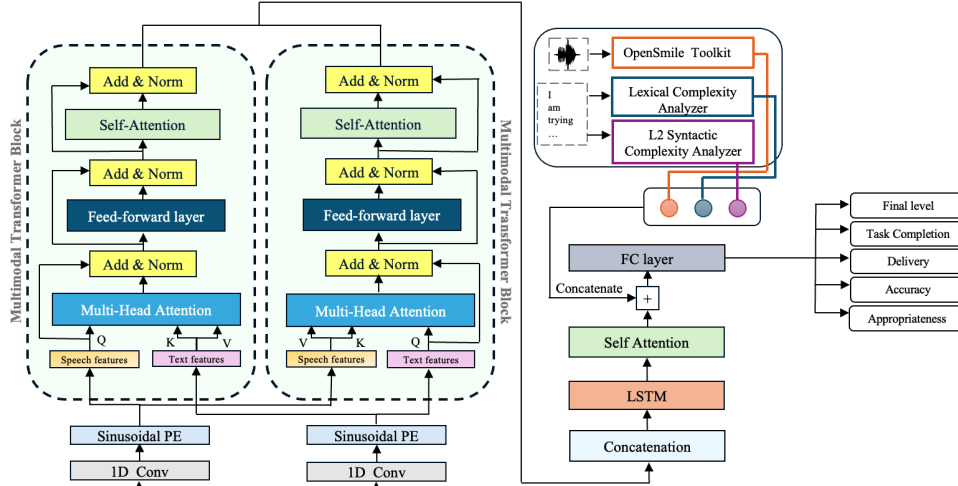


Figure 1: Proposed multimodal model architecture with two transformer blocks.

more, previous multimodal approaches often lack mechanisms for iterative refinement of interactions between the two modalities across multiple processing stages. Moreover, none of them incorporated a task-aware optimization strategy for MTL that can capture relationships among different scoring dimensions, leading to potential inconsistencies in multitask learning.

To address these limitations, we propose an effective multimodal and multitask learning model that predicts five distinct proficiency scores for non-native English speech. Our model employs two multimodal transformer blocks, where multi-head attention in each block captures cross-modal interactions by iteratively refining each modality’s features based on cues from the other. In addition, we incorporate a trait-aware loss to strengthen MTL training, allowing each scoring dimension to be more effectively and accurately predicted. With these, our method not only predicts a broader set of scoring criteria but also demonstrates enhanced robustness and generalizability over simpler unimodal or multimodal approaches in the application of automated speaking assessment.

2. METHOD

The proposed multimodal model architecture depicted in Figure 1 includes two key parts: (i) two multimodal transformer blocks and (ii) the inclusion of additional features. The model is trained using a joint loss function, which includes a Trait-Aware (TA) loss along with Mean Squared Error (MSE) loss to improve performance across multiple scoring criteria.

2.1. Multimodal Transformer Blocks

Building on the transformer-based fusion architectures explored in emotion analysis [23] and speech emotion recognition [24], we focus on the unique challenges of English speech assessment. The multimodal transformer blocks lie at the core of our architecture, designed to address the challenges of integrating and refining information from speech and text modalities. Each block utilizes a multi-head attention (MHA) mechanism as the first step to iteratively enhance features from one modality using cues from the other. Unlike simple concatenation-based approaches explored in previous studies, our MHA-driven block dynamically exchanges information across modalities. This enables the model to selectively focus on relevant speech or text features for more comprehensive proficiency scoring.

2.2. Additional Features

To enhance the model’s ability to capture detailed characteristics from both modalities, we also incorporate additional linguistic and acoustic features. These additional features are essential because transformer blocks, while capturing high-level representations for accurate predictions, may overlook finer linguistic and acoustic details. We address this by incorporating more granular speech characteristics. Additional acoustic features are extracted using the openSMILE Toolkit [25], which extracts low-level descriptors for speech analysis. L2 Syntactic Complexity Analyzer [26] extracts syntactic features, which compute the syntactic complexity of a written language. Finally, lexical features are extracted using the Lexical Complexity Analyzer [27], which is a computational system designed to determine dimensions of lexical richness.

2.3. Joint Loss

Most speech scoring systems are trained with MSE loss. However, using only MSE loss is not sufficient for the MTL model to understand the relationships between the traits. Therefore, we integrate trait-similarity into the loss function, which is called the TA loss. For TA loss, we first calculate the similarity of ground-truth trait scores with PCC scores. If the similarity is beyond a certain threshold, the model learns the cosine similarity of its predicted trait scores. The formula for calculating the similarity between the traits is determined as follows:

$$TA = \begin{cases} 1 - \cos(\hat{y}_j, \hat{y}_k), & \text{if } r(\mathbf{y}_j, \mathbf{y}_k) \geq \delta \\ 0, & \text{else} \end{cases} \quad (1)$$

Here, \cos , r , δ represent cosine similarity, PCC, and the threshold respectively. It is important to note that our TA loss of calculating similarity between every trait is distinct from [28], which excluded *Overall* trait due to its relatively low correlation with other traits in the task of essay scoring. On the other hand, we chose to include the *Final Level* trait in calculating similarities between the traits. Our rationale is that a higher score in the *Final Level* trait could correlate with higher scores in other traits as well, potentially aiding the model in better understanding the relationships between traits. Therefore, the total loss of combining TA and MSE losses is formulated as follows:

$$L_{\text{total}}(y, \hat{y}) = (1 - \alpha) \cdot L_{\text{TA}}(y, \hat{y}) + \alpha \cdot L_{\text{MSE}}(y, \hat{y}) \quad (2)$$

where the MSE loss is formulated as,

$$L_{\text{MSE}}(y, \hat{y}) = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \hat{y}_{ij})^2 \quad (3)$$

for predicting M number of traits for N speeches. For TA loss, we used 0.7 for both δ and α .

3. Experiments

We conducted experiments with four embedding types for speech (MFCCs, wav2vec 2.0) and text (GloVe, BERT) to identify the optimal pair for our multimodal model. We compared our model with (i) unimodal (speech-only or text-only) models to emphasize the advantages of our multimodal approach, (ii) two multimodal baseline models from [12] and [21] that used concatenated speech and text embeddings and treated it as an integrated input, and (iii) an independent multimodal processing approach from [19], originally proposed for single-score automatic pronunciation assessment tasks.

3.1. Data

The Evaluation Data of Topic-adaptive English Speaking Test by Korean Learners of English [29] is an open-source dataset designed for automatic speech evaluation in English proficiency tests. The participants, aged between 20-40, responded to questions related to self-introduction, topics from a pre-survey, and unrelated topics to ensure a wide range of responses. Each response is evaluated by two human expert raters who assign scores on a scale of 1-5 across four categories: *Task Completion*, *Delivery*, *Accuracy*, and *Appropriateness*. The final score for each category is the average of the scores given by the two raters. The *Final Level* is then determined by averaging the scores from these four categories.

Each utterance was transcribed by our in-house ASR system, which was built by using Kaldi [30]. To optimize the system for non-native English speakers' pronunciation, it integrates both native English and non-native Korean-accented English speech to train the acoustic model. As a result, the ASR model showed 14.2% word error rates (WER). We restricted the dataset to include only those samples in which at least one word was spoken. This resulted in a total of 29,857 speech samples. The data was partitioned into three datasets: train (N=20,965, 567.76 hours), validation (N=4,501, 121.66 hours), and test (N=4,391, 121 hours).

3.2. Embeddings

MFCCs are traditional acoustic features that capture the short-term power spectrum of sound, often used in automatic pronunciation and speech assessment tasks [10, 31]. In this study, frames of 39-dimensional MFCC features are extracted using a 25 ms frame size with a 10 ms shift. In addition, pre-trained GloVe word embeddings with 300-dimensional vectors are used for text embeddings. GloVe vectors can capture the meaning of words by analyzing the co-occurrence statistics of words in a corpus [32]. Each word in the text transcript is mapped to its corresponding GloVe embedding vector. If a word is not found in the GloVe vocabulary, a zero vector serves as a placeholder.

Transformer-based embeddings from wav2vec 2.0-base for speech and BERT-base for text, which are kept frozen, are also adopted, leveraging their robust representations that have shown promising results in evaluating the speech of English learners [19, 23]. Audio embeddings from wav2vec 2.0-base model are

obtained from the last hidden state. For text representation, each sentence in the transcript is tokenized using the BERT-base tokenizer and processed through the model to obtain the last hidden state, thereby capturing the semantic meaning of the text.

3.3. Experimental Setup

We used four embedding combinations in all the tested models. For unimodal experiments, MFCCs, wav2vec 2.0, GloVe, and BERT are used individually in our proposed model with only one multimodal transformer block, in which the MHA mechanism in this case operates on a single modality.

We also compare our proposed model with three multimodal baseline models that can employ both speech and text embeddings. The first baseline is the multitask model from [21], which employs separate linear regression layers to predict multiple scores simultaneously from the concatenated embeddings. The second baseline model, based on [12], employs a BLSTM-RNN layer with 200 hidden units per direction followed by a ReLU activation and a final FC layer to predict single score at a time. These two baseline models concatenate speech and text embeddings into a single input vector, treating them as an integrated representation. Although they utilize embeddings from two distinct modalities, they do not handle each modality separately. Therefore, we refer to these models as *Integrated*_multimodal. The third baseline model, from [19], processes speech and text embeddings separately using BLSTM layers to independently encode both acoustic and linguistic features. Each output is refined through a linear layer and global average pooling (GAP), and the combined features are processed through a final linear layer to produce a single score prediction. We refer to this model as *Independent*_multimodal. These three multimodal baselines were trained using the Adam optimizer [33] with a learning rate of 1e-4 and a batch size of 64, using only MSE loss.

In our proposed independent multimodal setup, two transformer blocks are employed as Figure 1 to independently handle each modality, mixing speech and text embedding dynamically. Unimodal and our proposed multimodal models were trained with the Adam optimizer at a learning rate of 1e-5 and with a batch size of 128, using the combination of MSE and TA losses.

4. RESULTS

The results from Table 1 demonstrate that our proposed multimodal model outperforms unimodal, integrated multimodal, and independent multimodal baselines in each embedding combinations. Notably, the combination of wav2vec 2.0 and BERT embeddings with our proposed multimodal model yielded the highest performance, with an average PCC of 0.734 and the lowest standard deviation of 0.0129, highlighting the model's accuracy and generalization ability across the scoring criteria.

4.1. Performance Comparison with Baseline Models

Unimodal models demonstrated the lowest performance overall, emphasizing the limitations of relying on a single data type. Among them, the model using wav2vec 2.0 embeddings performed the best, achieving an average PCC of 0.682, while models using MFCCs, BERT, and GloVe embeddings achieved average PCCs of 0.665, 0.658, and 0.636, respectively.

The *Integrated*_multimodal baselines, based on [21] and [12], showed improvements over unimodal configurations but were still outperformed by our proposed model. The highest-performing model, from [12], achieved an average PCC of

Table 1: *PCC results of our proposed independent multimodal model against unimodal and integrated multimodal models depending on its embeddings. TC: Task Completion, D: Delivery, ACC: Accuracy, APP: Appropriateness, FL: Final Level, AVG: Average, SD: Standard Deviation.*

	Embeddings	TC	D	ACC	APP	FL	AVG	SD
Unimodal	MFCCs	0.681	0.675	0.657	0.648	0.664	0.665	0.013
	Wav2vec 2.0	0.698	0.696	0.672	0.663	0.682	0.682	0.0139
	GloVe	0.652	0.629	0.612	0.657	0.632	0.636	0.0183
	BERT	0.677	0.642	0.651	0.658	0.663	0.658	0.0131
[21]	MFCCs & GloVe	0.614	0.630	0.603	0.578	0.554	0.596	0.0301
	MFCCs & BERT	0.617	0.602	0.589	0.575	0.556	0.588	0.0236
	Wav2vec 2.0 & GloVe	0.603	0.611	0.587	0.563	0.566	0.586	0.0215
	Wav2vec 2.0 & BERT	0.708	0.701	0.675	0.668	0.687	0.688	0.0169
[12]	MFCCs & GloVe	0.672	0.678	0.658	0.632	0.680	0.664	0.0198
	MFCCs & BERT	0.678	0.660	0.644	0.634	0.639	0.651	0.0179
	Wav2vec 2.0 & GloVe	0.641	0.683	0.652	0.628	0.633	0.648	0.0219
	Wav2vec 2.0 & BERT	0.724	0.720	0.698	0.684	0.706	0.706	0.0163
[19]	MFCCs & GloVe	0.677	0.673	0.668	0.653	0.690	0.680	0.0154
	MFCCs & BERT	0.685	0.670	0.664	0.643	0.692	0.671	0.0192
	Wav2vec 2.0 & GloVe	0.688	0.698	0.673	0.664	0.689	0.683	0.0136
	Wav2vec 2.0 & BERT	0.723	0.730	0.705	0.695	0.720	0.714	0.0143
Proposed Model	MFCCs & GloVe	0.735	0.737	0.728	0.703	0.722	0.725	0.0137
	MFCCs & BERT	0.709	0.673	0.702	0.703	0.690	0.695	0.0143
	Wav2vec 2.0 & GloVe	0.691	0.714	0.711	0.661	0.695	0.694	0.0211
	Wav2vec 2.0 & BERT	0.746	0.747	0.729	0.716	0.734	0.734	0.0129

0.706 using wav2vec 2.0 and BERT embeddings, which is significantly lower than our model’s performance (0.734).

The *Independent* multimodal baseline from [19], which concatenates speech and text embeddings after separately processing them through BLSTM layers, exhibited better performance than the *Integrated* models. With wav2vec 2.0 and BERT embeddings, this model achieved an average PCC of 0.714, outperforming both the *Integrated* baselines and unimodal models. However, all these baseline models still fell short of our proposed model, demonstrating the added value of our cross-modal attention mechanism and the use of the TA loss in capturing inter-trait dependencies.

4.2. Proposed Model Performance Across Embedding Combinations

Among all embedding combinations, wav2vec 2.0 and BERT consistently outperformed the others, achieving the highest average PCC of 0.734 and the lowest SD of 0.0129. This underscores their ability to effectively capture both acoustic nuances and semantic richness, making them an effective choice for L2 English speech assessment. Other combinations, such as MFCCs & GloVe, also performed well but exhibited higher variability in SD and lower average PCC scores, indicating less consistent predictions across scoring criteria. This further confirms the effectiveness of transformer-based embeddings in evaluating L2 English speech.

While our model exhibited strong performance across all scoring criteria, it showed relatively weaker performance in the *Appropriateness* criterion compared to other traits. This limitation is likely attributed to the lack of contextual information about the specific questions being answered, which is a key component of *Appropriateness* evaluation.

4.3. Ablation Studies

To explore the impact of each component of our proposed multimodal model, we conducted an incremental analysis as our ab-

lation studies using the best-performing embeddings (wav2vec 2.0 and BERT), as shown in Table 2. The results showed that predicting scores with audio embeddings only led to the lowest performance (average PCC of 0.682), indicating that relying solely on audio embeddings is insufficient for optimal performance. Adding text embeddings and additional features progressively improved the model’s performance. The inclusion of TA loss resulted in the best results, significantly reducing performance variance across traits. This highlights the critical role of multimodal integration in capturing cross-modal dependencies and the effectiveness of MTL in leveraging inter-trait relationships.

Table 2: *PCC results of ablation studies for the individual components of our proposed model.*

Model	TC	D	ACC	APP	FL	AVG
Audio Only	0.698	0.696	0.672	0.663	0.682	0.682
+ Text	0.728	0.730	0.704	0.693	0.702	0.711
+Add. features	0.736	0.739	0.721	0.701	0.715	0.722
+TA Loss	0.746	0.747	0.729	0.716	0.734	0.734

5. CONCLUSION

This study proposed a multimodal and trait-aware model for predicting five proficiency scores for L2 English speech using speech and text embeddings. The results demonstrated that the proposed model, particularly with wav2vec 2.0 and BERT embeddings, significantly outperformed baseline models, highlighting the effectiveness of combining cross-modal attention with Trait-Aware loss. By capturing nuanced relationships between proficiency traits, this model aligns well with human expert ratings, showcasing the potential of advanced multimodal approaches in automated speech assessment systems.

6. Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)].

7. References

- [1] M. Eskenazi, "An overview of spoken language technology for education," *Speech communication*, vol. 51, no. 10, pp. 832–844, 2009.
- [2] 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.
- [3] J. Bernstein, A. Van Moere, and J. Cheng, "Validating automated speaking tests," *Language Testing*, vol. 27, no. 3, pp. 355–377, 2010.
- [4] K. Evanini and X. Wang, "Automated speech scoring for non-native middle school students with multiple task types." in *INTERSPEECH*, 2013, pp. 2435–2439.
- [5] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *Sixth European Conference on Speech Communication and Technology*, 1999.
- [6] G. Deville, O. Deroo, H. Leich, S. C. Gielen, and J. Vanparys, "Automatic detection and correction of pronunciation errors for foreign language learners: the demosthenes application." in *EUROSPEECH*, 1999, pp. 843–846.
- [7] W. Hu, Y. Qian, and F. K. Soong, "A new dnn-based high quality pronunciation evaluation for computer-aided language learning (call)." in *Interspeech*, 2013, pp. 1886–1890.
- [8] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.
- [9] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children english language learners," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [10] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, X. Wang, K. Zechner, L. Chen, J. Tao, A. Ivanou, and Y. Qian, "Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 338–345.
- [11] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, "End-to-end neural network based automated speech scoring," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 6234–6238.
- [12] Y. Qian, P. Lange, K. Evanini, R. Pugh, R. Ubale, M. Mulholland, and X. 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)*. IEEE, 2019, pp. 8112–8116.
- [13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *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.
- [14] 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.
- [15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [16] 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.
- [17] S. Park and J. Culnan, "Automatic perceptual judgment using neural networks," *The Journal of the Acoustical Society of America*, vol. 146, no. 4-Supplement, pp. 2957–2957, 2019.
- [18] —, "Automatic proficiency judgments: Accentedness, fluency, and comprehensibility," *The Journal of the Acoustical Society of America*, vol. 150, no. 4-Supplement, pp. A357–A357, 2021.
- [19] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," in *INTERSPEECH*, 2022, pp. 1411–1415.
- [20] X. Wang, K. Evanini, Y. Qian, and M. Mulholland, "Variations of multi-task learning for spoken language assessment," in *INTERSPEECH*, 2022, pp. 74 456–4460.
- [21] 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.
- [22] J. Liu, A. Wumaier, C. Fan, and S. Guo, "Automatic fluency assessment method for spontaneous speech without reference text," *Electronics*, vol. 12, no. 8, p. 1775, 2023.
- [23] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.
- [24] K. Kim and N. Cho, "Focus-attention-enhanced crossmodal transformer with metric learning for multimodal speech emotion recognition," in *Proceedings of INTERSPEECH 2023*, 2023, pp. 2673–2677.
- [25] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.
- [26] X. Lu, "Automatic analysis of syntactic complexity in second language writing," *International Journal of Corpus Linguistics*, vol. 15, no. 4, pp. 474–496, 2010.
- [27] —, "The relationship of lexical richness to the quality of esl learners' oral narratives," *Modern Language Journal*, vol. 96, no. 2, pp. 190–208, 2012.
- [28] H. Do, Y. Kim, and G. G. Lee, "Prompt- and trait relation-aware cross-prompt essay trait scoring," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 1538–1551.
- [29] "Evaluation data of topic-adaptive english speaking test by korean learners of english," <https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=71418>, Solution for Learning Innovation (SLI) Edu, 2023.
- [30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proceedings of the 2014 IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.
- [31] Z. Yu, D. Gerritsen, A. Ogan, A. W. Black, and J. Cassell, "Automatic prediction of friendship via multi-modal dyadic features," in *Proceedings of the SIGDIAL 2013 Conference*, 2013, pp. 51–60.
- [32] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.
- [33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.