10th Workshop on Speech and Language Technology in Education (SLaTE)
22-24 August 2025, Nijmegen, Netherlands

A Novel Data Augmentation Approach for Automatic Speaking Assessment on Opinion Expressions

Chung-Chun Wang, Jhen-Ke Lin, Hao-Chien Lu, Hong-Yun Lin, Berlin Chen

[1]Dept. Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
{takala, jacob, howchien, buffett, berlin}@ntnu.edu.tw

Abstract

Automated speaking assessment (ASA) on opinion expressions is often hampered by the scarcity of labeled recordings, which restricts prompt diversity and undermines scoring reliability. To address this challenge, we propose a novel training paradigm that leverages a large language models (LLM) to generate diverse responses of a given proficiency level, converts responses into synthesized speech via speaker-aware text-to-speech synthesis, and employs a dynamic importance loss to adaptively reweight training instances based on feature distribution differences between synthesized and real speech. Subsequently, a multimodal large language model integrates aligned textual features with speech signals to predict proficiency scores directly. Experiments conducted on the LTTC dataset show that our approach outperforms methods relying on real data or conventional augmentation, effectively mitigating low-resource constraints and enabling ASA on opinion expressions with cross-modal information.

Index Terms: Automated speaking assessment, Opinion Expression, Data Augmentation

1. Introduction

In recent years, technologies for computer-assisted language learning (CALL), such as automated speaking assessment (ASA), have made significant strides to meet the growing demand for scalable and objective evaluation of second-language (L2) speaking proficiency in both academic and professional contexts [1, 2, 3]. In L2 speaking examinations, test-takers are often required to express opinions on a given topic or to participate in open-ended discussions, tasks that are crucial for assessing learners' communicative competence and critical thinking in languages such as English [4]. ASA systems that provide immediate feedback and support large-scale scoring are becoming commonplace in both CALL applications and formal testing scenarios.

However, traditional human scoring is time-consuming, labor-intensive, and prone to inconsistency among raters, a problem that is especially acute in high-stakes opinion-expression tasks such as those found in TOEFL, IELTS, and GEPT [5, 6, 7]. CALL platforms, through their immediate feedback and interactive instruction, have evolved from early academic prototypes into mature intelligent classroom assistants and robust online testing systems, effectively alleviating many of these challenges. By reducing the burden on human raters and improving scoring consistency, effective ASA for opinion-expression tasks has figured prominently in diverse use cases of language teaching and assessment.

Early ASA approaches combine automatic speech recognition (ASR) with hand-crafted acoustic and linguistic features, such as pronunciation accuracy [8, 9], fluency [10], prosody [11], rhythm [12], and others. These features play a fundamental role in many automated scoring tasks and became the cornerstone of the field [6, 13, 14, 15, 16]. A few pioneering systems extract measures of pronunciation fluency, lexical richness, grammatical accuracy, content relevance, and the like from L2 speech, and then draw on statistical models to predict a composite proficiency score.

More recently, multi-task and multimodal architectures have been introduced to capture complementary information inherent in both the text content and speech signal of a spoken response. The emergence of self-supervised speech encoders (e.g., wav2vec 2.0 [5, 7, 17]), Transformer-based text models [18], and multimodal large language model (MLLM) such as Phi-4 multimodal [19] gives rise to a variety of end-to-end modeling frameworks that learn response representations jointly from raw audio and textual transcripts, markedly improving scoring accuracy and generalization. Nevertheless, these models still rely excessively on large amounts of annotated training datasets that are frequently scarce for ASA on opinion expressions, thus often leading to overfitting on limited prompts.

To alleviate data scarcity, prior studies explore combining text-to-speech (TTS) techniques with LLM–driven data augmentation [20]. These approaches have shown some success on various speech and language processing tasks, such as automatic speech recognition (ASR) [21, 22], emotion recognition [23], text classification [24, 25], and intent recognition [26]. For the purpose of spoken content generation, these methods typically use LLMs to generate diverse texts, thereby increasing content variety, and subsequently employ TTS systems to synthesize speech from the generated texts, enriching speaker characteristics, prosody, and content. Because synthesized and real texts and audio waves differ in their acoustic feature distributions, there is good reason to apply dynamic instance reweighting or metric-learning techniques to adaptively adjust the importance of each synthesized instance based on its proximity to real counterparts in a embedding space, thus narrowing the domain gap [24]. However, to our knowledge, there is still a dearth of work investigating the synergistic effect of combining the aforementioned methods and their extensions to automated scoring of spoken opinion expressions.

In view of this, we present a unified framework for spoken opinion-expression assessment under the constraint of limited resources, which combines generative data augmentation with neural multimodal modeling. First, we develop a two-step augmentation pipeline in which an LLM generates proficiency-conditioned text responses [27] and a high-fidelity TTS system synthesizes these text responses into spoken utterances while retaining non-native prosody and natural disfluencies [22]. Next,

10.21437/SLaTE.2025-40

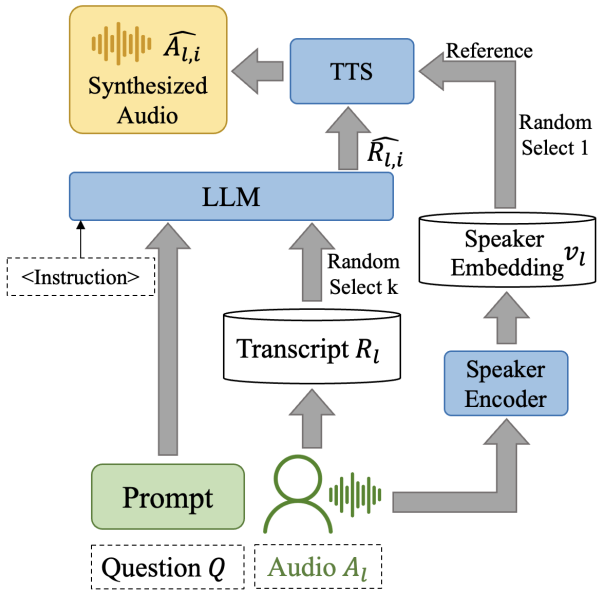


Figure 1: *Pipeline for synthesized speech generation. We build in-context prompts from k real transcripts and condition a multi-speaker TTS model on LLM-generated text $\hat{R}_{l,i}$ and a speaker embedding v_l to produce learner-like audio.*

we introduce a novel dynamic importance loss that builds on a recently proposed data-weighting principle [24] for training of our scoring model. This training loss automatically increases the contribution of synthesized instances, so as to bridge domain gaps or proficiency distribution imbalance and meanwhile curb overreliance on the generated data. Finally, the backbone of our scoring model is a Phi-4 multimodal [19] finetuned with LoRA adapters [28], which directly fuses text prompt, as well as raw audio and ASR transcript of a spoken response, to predict the corresponding proficiency score. Our approach delivers robust performance on both seen and unseen prompts despite that only limited labeled instances are made available. Our approach not only attains superior performance on low-resource speaking assessment but also offers a replicable paradigm for other multimodal, low-resource speaking assessment tasks. In summary, our key contributions are at least three-fold:

- **Unified Low-resource Modeling Framework:** Combining texts generated from a large language model and their corresponding high-fidelity synthesized speech utterances to substantially expand L2 training data.
- **Dynamic Importance Loss:** Leveraging an adaptive loss mechanism that effectively mitigates domain and distribution mismatches between synthesized and real data.
- **Multimodal Scoring:** Developing an efficient end-to-end scoring model that fuses speech and text embeddings for robust opinion-expression scoring.

2. Methodology

Our approach enhances opinion expression assessment in low resource settings through three tightly integrated innovations. First, a proficiency aware text synthesis module uses a large language model, prompted with randomly sampled and permuted transcripts of authentic test taker responses, to generate answers aligned with each proficiency level and faithful to real language use. Second, speaker conditioned voice cloning, driven by speaker embeddings extracted from real recordings, guides a multi speaker TTS model to convert these texts into speech while preserving non native prosody, pauses, and disfluencies; Figure 1 illustrates the complete speech synthesis pipeline. Finally, the proposed dynamic importance reweighting loss, shown in Figure 2, compares the confidences that a frozen quality model and a trainable target model assign to the correct label and then dynamically increases the weight of highly realistic synthetic samples while keeping genuine data predominant. By combining rich text augmentation, authentic speech cloning, and principled loss reweighting, our method yields a training corpus that balances diversity and authenticity and achieves robust scoring performance under strict data constraints. Implementation details and experimental results appear in the following sections.

2.1. LLM-Based Text Synthesis

To simulate responses from test-takers of different proficiency levels and enrich the diversity of our synthesized training instances, we leverage the in-context learning capability of an LLM. For each target level l, we first randomly select k transcribed responses from the genuine training dataset, denoted as $\{R_{l,1}, R_{l,2}, \ldots, R_{l,k}\}$. We then concatenate the task prompt Q with these examples in order to form an input prompt:

$$P_l = \big[Q;\, R_{l,1};\, R_{l,2};\, \ldots;\, R_{l,k}\big], \tag{1}$$

accompanied by the instruction, "Please generate your response in the style of the above examples." During decoding, we set a relatively high temperature τ so that

$$\hat{R_{l,i}} \sim \mathrm{LLM}\big(P_l;\, \tau\big), \quad i = 1, \ldots, n, \tag{2}$$

where $\hat{R_{l,i}}$ is the i-th synthesized response and n is the number of instances generated per level. By shuffling the order of the examples and repeating this generation process across multiple rounds, we obtain a synthesized dataset of text responses that remains aligned with the prompt's content while exhibiting rich variation in vocabulary, syntax, and viewpoints.

2.2. Voice Cloning Synthesis

When converting a synthesized texts $\hat{R_{l,i}}$ into the corresponding realistic speech of a test-taker, we observed that a standard TTS model tends to produce over-perfect audio which lacks natural characteristics of non-native speakers, such as disfluencies, mispronunciations, and prosodic variability, since it is typically trained on corpora of native speakers. To remedy this, we employ a multi-speaker TTS conditioned on disparate speaker embeddings rather than a single fixed voice.

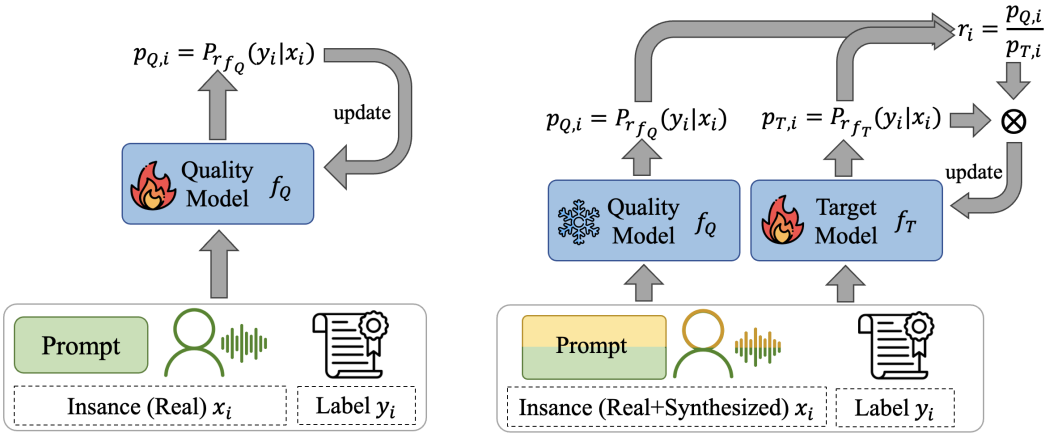
For each generated text $\hat{R_{l,i}}$, we randomly pick up a short reference audio clip $A_{l,j}$ from a real speaker at the same proficiency level l. A speaker encoder E is then used to map this clip to an embedding:

$$v_{l,j} = E(A_{l,j})\,, \tag{3}$$

which encodes voice identity, speaking rate, rhythm, and typical learner disfluencies. We feed both the text and embedding into the multi-speaker TTS model G to produce a cloned waveform:

$$\hat{A_{l,i}} \sim G\big(\hat{R_{l,i}},\, v_{l,j}\big)\,, \quad i = 1, \ldots, n. \tag{4}$$

By randomly pairing each $g_{l,i}$ with different $v_{l,j}$, we ensure the synthesized audio faithfully reproduces learner-specific



Stage 1. Quality Model Fine-Tuning

Stage 2. Target Model Training with DIMP

Figure 2: *Two-stage training workflow. Stage 1 fine-tunes a Quality Model f_Q on real data. Stage 2 trains the Target Model f_T on mixed real and synthesized instances using the ratio r_i and weight w_i to dynamically reweight each instance's loss.*

traits, in contrast to the uniform, overly clean output generated by a conventional single-speaker TTS model.

2.3. End-to-End Multimodal Training with Dynamic Importance Reweighting

After preparing synthesized texts and cloned speech, we introduce a new dynamic importance reweighting loss that automatically adjusts the weigh of each training instance according to its fidelity to real learner behavior. This loss computes the ratio of confidence scores produced by a fixed "quality model" and by our trainable "target model," then uses that ratio to amplify high-quality synthesized instances and suppress poorly matched ones. At the same time, real instances receive an elevated base weight to ensure the model remains grounded in authentic speech distributions. We integrate this loss into a neural scoring model built upon the Phi-4 multimodal [19] backbone and fine-tuned with LoRA adapters [28] for maximal parameter efficiency. The training phase unfolds in two stages:

Stage 1. Quality Model Calibration We first fine-tune the Phi-4 multimodal [19] backbone on real prompt–audio–score triplets alone, using standard cross-entropy to obtain a *Quality Model* f_Q. This model establishes a reliable reference distribution over genuine responses of test-takers.

Stage 2. Target Model Optimization Next, we train a separate *Target Model* f_T on the union of real and synthesized instances. For each instance x_i with true label y_i, we compute

$$p_{Q,i}=\Pr_{f_Q}(y_i|x_i) \quad \text{and} \quad p_{T,i}=\Pr_{f_T}(y_i|x_i), \tag{5}$$

which denote the confidence that each of these two model assigns to the correct label. We then define the *importance ratio*:

$$r_i = \frac{p_{Q,i}}{p_{T,i}}, \tag{6}$$

which will be employed to amplify those synthesized instances that most closely match real-data behavior. To prevent overreliance on generated data—whose fluency often exceeds natural learner speech—we introduce a real-instance weight $\alpha > 1$ with learning rate η. The combined per-instance loss is

$$\ell_i = -\,\eta\, r_i\, w_i\, \log\big(p_{T,i}\big), \quad w_i = \begin{cases} \alpha, & x_i \in \mathcal{D}_{\text{real}}, \\ 1, & x_i \in \mathcal{D}_{\text{syn}}. \end{cases} \tag{7}$$

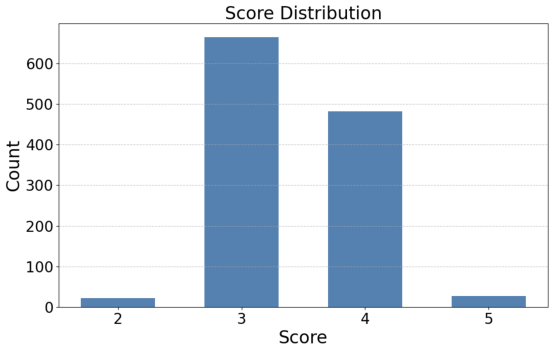


Figure 3: *Score distribution for real data.*

We optimize the average of these losses over all N training instances:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \ell_i. \tag{8}$$

By introducing this novel training loss, we ensure that the augmentation of synthesized instances enriches the training set without distorting the model's prediction capability on the authentic speech. By doing so, the estimated multimodal scoring model is anticipated to achieve robust performance under low-resource constraints.

3. Experiments and Results

3.1. Dataset

We used the LTTC GEPT Intermediate to Advanced opinion expression dataset, which contains 1 200 prompt–response–score triples evenly distributed across four prompt types within the discussion subtask. Of these, 300 triples belonging to the fourth prompt type were set aside as an unseen test set in order to simulate domain transfer. The remaining 900 triples were divided in an 8:1:1 ratio into training, validation, and seen-test sets. Each spoken response received a 1-to-5 score from two expert raters. Following the official GEPT standard, we averaged the two scores, then applied floor rounding (no response

Table 1: *Comparison of baseline models and our approach. End-to-End models are marked with "✓".*

Model	End-to-End	Seen Test		Unseen Test	
		Overall Acc.	Pass–Fail Acc.	Overall Acc.	Pass–Fail Acc.
Wav2vec2 [17]	✓	61.80%	65.10%	55.33%	57.52%
BERT [18]		75.28%	78.65%	63.00%	67.89%
Wav2vec2+BERT		73.33%	77.78%	64.00%	67.89%
Phi-4mm [19] (only real)	✓	73.33%	77.78%	63.33%	68.00%
Our Model	✓	**76.67%**	**81.11%**	**64.00%**	**68.33%**

ended with a score of 1). Scores of 4 or 5 were labeled passing and scores of 1 to 3 were labeled failing; this binary label reflects the GEPT Intermediate to Advanced speaking threshold. All spoken responses were manually transcribed for model input. To alleviate data scarcity, we generated a synthesized corpus that matches the size and score distribution of the real training set. Figure 3 shows the final score distributions for real data.

3.2. Experimental Setup

We generate synthesized texts with OpenAI o4-mini [1] using up to ten shuffled in-context examples per prompt and decoding at temperature $\tau = 1.5$. For each text, we sample a real reference clip at the same proficiency level, extract its 512-D speaker embedding via Coqui-ai XTTSv2 [2], and synthesize learner-like speech with a multi-speaker TTS. Our scoring backbone is a Phi-4 multimodal [19] equipped with LoRA adapters [28]. We first fine-tune a Quality Model on real prompt–speech–score triplets for three epochs using AdamW ($\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 10^{-7}$), gradient clipping at 1.0, learning rate $\alpha = 2$, batch size 16. We then train the Target Model on the combined real and synthesized corpus under the same settings, applying a real-sample weight $\alpha = 2$ in our dynamic importance loss.

3.3. Experiment Results

Table 1 compares our proposed model with strong single-modality and dual-modality baselines on both the seen and unseen prompts. The baselines include an acoustics-only wav2vec2 [17] system, a text only BERT [18] model, and a Phi-4 multimodal [19] finetuned on the real data alone.

On the test set of seen prompts, our model achieves the highest overall accuracy and pass–fail accuracy among all methods. By combining speech and text through LLM-driven augmentation and dynamic reweighting, it alleviate the downsides the systems that rely either on purely acoustic or purely textual information. It notably outperforms wav2vec in grading consistency and robustness, while surpasseing BERT despite BERT's strong contextual encoding of transcripts. Although fine-tuning Phi-4 multimodal model on real data alone yields competitive results, our pipeline with synthesized augmentation and the dynamic importance reweighting loss delivers the best performance, confirming the practical effectiveness of our contributions.

On the the test of unseen prompt, our model maintains its advantage and incurs a smaller performance drop than any baseline. The synergy of diverse synthesized data and adaptive loss weighting enables more dependable scoring of novel prompts, alleviating under- or over-scoring on unfamiliar content. These results show that our unified training framework not only boosts accuracy on familiar prompts but also enhances robustness across different prompts.

3.4. Ablation Studies

To isolate the effects of synthesized augmentation and real-instance weighting, we compare three variants in Table 2: training on synthesized data alone, mixing real and synthesized with equal weight, and mixing with an elevated real-sample weight of $\alpha = 2$. Training on synthesized data by itself performs poorly, demonstrating that synthesized speech cannot substitute entirely for authentic speech. Introducing real training instances alongside synthesized ones recovers almost all of the performance lost in the first variant, validating the value of mixed-data training. Finally, increasing the weight on real instances yields the best results on both seen and unseen prompts, confirming that our dynamic importance reweighting effectively balances synthesized diversity against fidelity to genuine speech.

Table 2: *Ablation results on the seen and unseen test sets.*

Variant	Seen Acc.	Unseen Acc.
Only synthesized	42.22%	46.67%
Mix ($\alpha = 1$)	72.22%	63.67%
Mix ($\alpha = 2$)	**76.67%**	**64.00%**

4. Conclusion

In this paper, we have proposed a unified, low-resource training framework for opinion-expression assessment that integrates LLM-driven text synthesis, speaker-conditioned voice cloning, and a novel dynamic importance reweighting loss within an end-to-end multimodal model. Empirical results on the LTTC GEPT Intermediate–Advanced dataset have shown that our approach outperforms strong single- and dual-modality baselines and generalizes robustly to both seen and unseen prompts, demonstrating the promise of generative augmentation and adaptive loss reweighting for scalable, accurate automated speaking assessment under data-scarcity scenarios.

Limitations and Future Work Despite these gains, accent and hesitation mismatches persist and our experiments are limited to one language. The skewed mid-range score distribution also impairs extreme-level discrimination. Future work will adapt the framework to additional languages and domains, address class imbalance, and explore lightweight on-device models alongside advanced augmentation strategies.

[1] OpenAI o4-mini https://platform.openai.com/docs/models/o4-mini
[2] Coqui-ai XTTSv2 https://github.com/coqui-ai/TTS

5. Acknowledgement

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

6. References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech communication*, vol. 51, no. 10, pp. 832–844, 2009.

[2] B. Lin and L. Wang, "Attention-based multi-encoder automatic pronunciation assessment," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7743–7747.

[3] K. Zechner and X. Xi, "Towards automatic scoring of a test of spoken language with heterogeneous task types," in *Proceedings of the third workshop on innovative use of NLP for building educational applications*, 2008, pp. 98–106.

[4] S. Bannò, K. M. Knill, M. Matassoni, V. Raina, and M. Gales, "Assessment of l2 oral proficiency using self-supervised speech representation learning," in *9th Workshop on Speech and Language Technology in Education (SLaTE)*, 2023, pp. 126–130.

[5] T.-H. Lo, F.-A. Chao, T.-i. Wu, Y.-T. Sung, and B. Chen, "An effective automated speaking assessment approach to mitigating data scarcity and imbalanced distribution," in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 1352–1362. [Online]. Available: https://aclanthology.org/2024.findings-naacl.86/

[6] 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.

[7] H.-C. Lu, C.-C. Wang, J.-K. Lin, and B. Chen, "Development of an english oral assessment system with the gept dataset," in *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2024, pp. 1–6.

[8] L. Chen, K. Evanini, and X. Sun, "Assessment of non-native speech using vowel space characteristics," in *2010 IEEE Spoken Language Technology Workshop*. IEEE, 2010, pp. 139–144.

[9] K. Takai, P. Heracleous, K. Yasuda, and A. Yoneyama, "Deep learning-based automatic pronunciation assessment for second language learners," in *HCI International 2020-Posters: 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*. Springer, 2020, pp. 338–342.

[10] H. Strik and C. Cucchiarini, "Automatic assessment of second language learners' fluency," 1999.

[11] E. Coutinho, F. Hönig, Y. Zhang, S. Hantke, A. Batliner, E. Nöth, and B. Schuller, "Assessing the prosody of non-native speakers of english: Measures and feature sets," 2016.

[12] K. Kyriakopoulos, K. M. Knill, and M. J. Gales, "A deep learning approach to automatic characterisation of rhythm in non-native english speech." ISCA, 2019.

[13] 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.

[14] 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.

[15] L. Chen, K. Zechner, S.-Y. Yoon, K. Evanini, X. Wang, A. Loukina, J. Tao, L. Davis, C. M. Lee, M. Ma *et al.*, "Automated scoring of nonnative speech using the speechrater sm v. 5.0 engine," *ETS Research Report Series*, vol. 2018, no. 1, pp. 1–31, 2018.

[16] J. Park and S. Choi, "Addressing cold start problem for end-to-end automatic speech scoring," in *Interspeech 2023*, 2023, pp. 994–998.

[17] 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.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[19] A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, A. Benhaim, M. Cai, V. Chaudhary, C. Chen *et al.*, "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," *arXiv preprint arXiv:2503.01743*, 2025.

[20] K. Feng and T. Chaspari, "A knowledge-driven vowel-based approach of depression classification from speech using data augmentation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[21] K. Yang, T.-Y. Hu, J.-H. R. Chang, H. S. Koppula, and O. Tuzel, "Text is all you need: Personalizing asr models using controllable speech synthesis," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[22] A. Fazel, W. Yang, Y. Liu, R. Barra-Chicote, Y. Meng, R. Maas, and J. Droppo, "Synthasr: Unlocking synthetic data for speech recognition," in *Interspeech 2021*, 2021, pp. 896–900.

[23] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Emomix: Emotion mixing via diffusion models for emotional speech synthesis," in *Interspeech 2023*, 2023, pp. 12–16.

[24] H.-Y. Kuo, Y.-H. Liao, Y.-C. Chao, W.-Y. Ma, and P.-J. Cheng, "Not all LLM-generated data are equal: Rethinking data weighting in text classification," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=oI5tZaWkF9

[25] K. M. Yoo, D. Park, J. Kang, S.-W. Lee, and W. Park, "GPT3Mix: Leveraging large-scale language models for text augmentation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2225–2239. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.192/

[26] G. Sahu, P. Rodriguez, I. Laradji, P. Atighehchian, D. Vazquez, and D. Bahdanau, "Data augmentation for intent classification with off-the-shelf large language models," in *Proceedings of the 4th Workshop on NLP for Conversational AI*, B. Liu, A. Papangelis, S. Ultes, A. Rastogi, Y.-N. Chen, G. Spithourakis, E. Nouri, and W. Shi, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 47–57. [Online]. Available: https://aclanthology.org/2022.nlp4convai-1.5/

[27] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language models with self-generated instructions," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 13 484–13 508. [Online]. Available: https://aclanthology.org/2023.acl-long.754/

[28] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.