



# Phonetic and Prosody-aware Self-supervised Learning Approach for Non-native Fluency Scoring

*Kaiqi Fu, Shaojun Gao, Shuju Shi, Xiaohai Tian, Wei Li, Zejun Ma*

ByteDance

kaiq.fu@gmail.com, gaoshaojun123@163.com,  
{shuju.shi, xiaohai.tian, liwei.speech, mazejun}@bytedance.com

## Abstract

Speech fluency/disfluency can be evaluated by analyzing a range of phonetic and prosodic features. Deep neural networks are commonly trained to map fluency-related features into the human scores. However, the effectiveness of deep learning-based models is constrained by the limited amount of labeled training samples. To address this, we introduce a self-supervised learning (SSL) approach that takes into account phonetic and prosody awareness for fluency scoring. Specifically, we first pre-train the model using a reconstruction loss function, by masking phones and their durations jointly on a large amount of unlabeled speech and text prompts. We then fine-tune the pre-trained model using human-annotated scoring data. Our experimental results, conducted on datasets such as Speechocean762 and our non-native datasets, show that our proposed method outperforms the baseline systems in terms of Pearson correlation coefficients (PCC). Moreover, we also conduct an ablation study to better understand the contribution of phonetic and prosody factors during the pre-training stage.

**Index Terms:** Computer Assisted Pronunciation Training (CAPT), Non-native Fluency Scoring, Phonetic and Prosody-aware, Self-supervised Learning

## 1. Introduction

The ability to speak fluently is a significant aspect when evaluating a learner’s language proficiency [1]. It is characterized by the seamless and effortless production of speech with minimal pauses, hesitation, or corrections [2–5]. L2 learners typically exhibit slower speech and more frequent unnecessary pauses compared to native speakers. Automatic scoring of fluency, serves as an essential module in computer-aided language learning (CALL) systems. It has been extensively studied in both “read aloud” [6–11] and “open response” [12–15] scenarios. In the “read aloud” scenario, L2 learners are required to read a provided prompt text, whereas the “open response” requires them to express their opinions freely based on a given question.

In this paper, we focus on “read aloud” scenario, where forced-alignment model is first applied to a pair of non-native speech and prompt text to generate time stamps of speech segments, such as phonemes, words and etc. Fluency related features are then extracted and fed into subsequent fluency scorers. Although recent end-to-end neural network based fluency scorers have achieved satisfactory results [7–11, 15], their performances heavily rely on the size of labeled scoring samples. In fact, the non-native data labeling process is costly and has scalability issues [16]. Take the recently released public free dataset Speechocean762 [17] for example, only 5,000 sentences have been assigned with human fluency scores. The comparison in [18] shows that the largest nonnative corpus only contains

90,841 utterances, but it is not publicly available.

To overcome the challenge of limited labeled data, many researchers are using pre-training and fine-tuning paradigms to leverage large amounts of unlabeled data [19,20]. In the field of natural language processing (NLP), masked language modeling (MLM) has become a popular method for pre-training models such as BERT [21], RoBERTa [22], and ERNIE [23]. MLM involves masking a subset of tokens in a sequence and training the model to predict these masked tokens, which enables the model to learn high-level contextual representations that can be beneficial for downstream tasks.

Recently, a new multi-stream transformer language model (MS-TLM) was proposed to jointly model phonetic content and prosody [24], which demonstrated the effectiveness of prosody prediction. In this paper, we propose a self-supervised learning approach that incorporates phonetic and prosodic information to improve non-native fluency scoring. The pre-trained model is used to predict masked phones and durations, which enhances the model’s ability to represent long-range phonetic and prosodic information. Specifically, we use an automatic speech recognition (ASR) system to generate phone-level raw sequential features, e.g. acoustic features, phone sequences, and duration, for pairs of non-native speech and prompt text. We then randomly mask 15% of these phone-level features and train our fluency scorer to predict the masked phone and duration. Finally, the pre-trained model is fine-tuned using limited human-annotated fluency scores. Experimental results show that the proposed approach can significantly improve fluency scoring in various configurations. An ablation study is also conducted to analyze the effect of different loss functions used in our pre-training stage on fluency scoring.

## 2. Related work

Over the past few decades, extensive research has been conducted on spoken fluency scoring. Traditionally, handcrafted features such as the statistics of speech break [6], speech rate [6, 7, 12–14], filled pause, and goodness of pronunciation (GOP) [7–9] were collected based on phone boundaries and fed into various fluency scorers such as SVM [12, 14], and multiple linear [6]. Recent works have employed sequence models to directly learn utterance-level fluency scores from phone-level raw features, including phonetic features (e.g., phone sequence [7–10]), prosodic features (e.g., energy [9], pitch [7] and phone duration [10]). Bi-directional Long Short Term Memory (BLSTM) [7, 10, 11, 15] and Transformer models [8, 9] have been used to capture the dynamic changes of phone-level pronunciation-related features for better modeling the evolution of local fluency over time.

More recently, self-supervised learning (SSL)-based speech

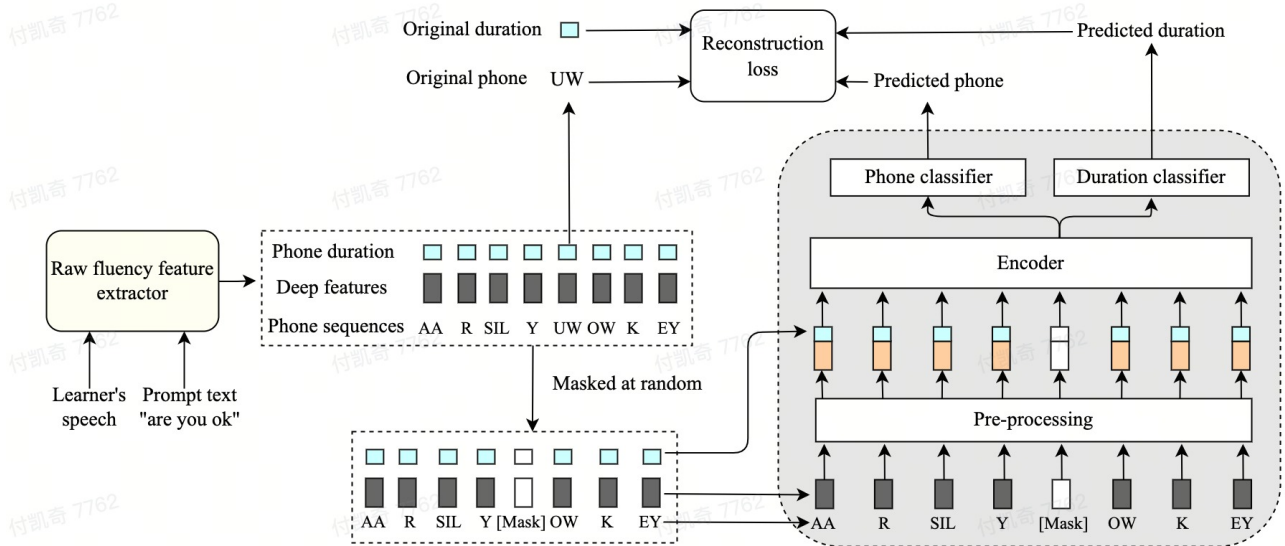


Figure 1: An overview of the phonetic and prosody-aware pre-training process.

models such as wav2vec2 [25] have been shown to be effective in learning meaningful representations from raw speech signals in various downstream tasks [26]. Inspired by this success, researchers used pre-trained SSL models like wav2vec2 [25], HuBERT [27], and WavLM [28] to extract features directly and feed them into fluency scorers [9, 11, 15]. Due to the promising performance, we consider the two SSL-based models [9, 15] as strong baselines of this work.

### 3. Method

This section details our approach for fluency scoring. Initially, we outline our phonetic and prosody-aware pre-training technique that employs self-supervised learning to reconstruct the masked phone and duration in each pre-training sample. The process flow for pre-training is illustrated in Figure 1. Subsequently, we elaborate on how we employ the pre-trained model for fluency scoring in downstream tasks.

#### 3.1. Phonetic and prosody-aware pre-training

##### 3.1.1. Phone-level fluency feature extraction

In [10], phone-level raw features were shown to be effective for assessing learners' speech fluency. Followed by the previous study, three segmental features (deep features, phone, and its duration) were extracted in this study. Initially, forced alignment is carried out to obtain time boundaries at the phone level, such as the beginning and end time of each phone and pause. The boundary information is then used to derive the phone sequence and its duration, which are denoted as  $\mathbf{e} \in \mathbf{R}^{1 \times N}$  and  $\mathbf{t} \in \mathbf{R}^{1 \times N}$ , respectively.  $\mathbf{t}$  refers to the number of frames within the phone. Following that, frame-level deep features (also known as bottleneck features) extracted from the acoustic model are averaged by the duration of each phone to obtain phone-level deep features, represented as  $\mathbf{X} \in \mathbf{R}^{N \times D}$ , where  $D$  and  $N$  represent the acoustic feature dimension and the number of phones, respectively. Finally, the phone-level raw features are inputted into the model for pre-training and fine-tuning.

##### 3.1.2. Masking strategy

During the pre-train stage, we utilize random masking to randomly replace 15% of phone-level raw features in each sample with a special mask token. Among the selected positions, 90% will be replaced with the special mask token and the remaining 10% will be kept unchanged. The model takes three input features, which correspond to three different mask marker methods: 1) the selected phonemes are replaced with the mask token, 2) the selected phone duration is set to zero, and 3) the selected deep features are replaced with zero vectors. To provide the model with duration ground-truth, we set the duration label to a range of 1-100. If the phone duration exceeds 100 frames, we cap the duration label at 100.

##### 3.1.3. Multitask based reconstruction loss

The pre-processing steps takes phone-level deep features  $\mathbf{X}$  as input, which are initially transformed into a condensed feature space  $\mathbf{X}'$  using a fully connected layer. Next, the phone sequence  $\mathbf{e}$  is converted into phone embeddings  $\mathbf{E}$ . The sum of  $\mathbf{E}$  and  $\mathbf{X}'$  output is then concatenated with phone duration  $\mathbf{t}$  and utilized as a sequence of input features for the SSL encoder, which generates the phone-level hidden representations  $\mathbf{H}$ .

$$\mathbf{H} = \mathcal{E}([\mathbf{X}' + \mathbf{E}; \mathbf{t}]), \quad (1)$$

where  $\mathcal{E}$  is presented the SSL encoder.

The phone-level hidden representations  $\mathbf{H}$  are then passed through two classifiers for phoneme and duration prediction, respectively. The pre-training model is optimized jointly by utilizing a multitask approach that minimizes the cross-entropy loss between the predicted and ground truth phonemes and durations. The loss function of  $i$ -th masked token is described as follows:

$$\mathcal{L}_i = \mathcal{L}_{ce}(y_i^p, \mathcal{P}_p(\mathbf{h}_i)) + \mathcal{L}_{ce}(y_i^d, \mathcal{P}_d(\mathbf{h}_i)), \quad (2)$$

where  $y_i^p$  and  $y_i^d$  are presented the ground truth phoneme and duration, respectively.  $\mathbf{h}_i$  is the phone-level hidden representations in  $i$ -th masked position. The phoneme and duration classifiers are denoted as  $\mathcal{P}_p(\cdot)$  and  $\mathcal{P}_d(\cdot)$ , respectively. It

Table 1: *Data splitting for fluency scorer*

		Train	Dev	Test
<b>Pre-train</b>	Unlabeled data	203,206	2,000	-
<b>Fine-tune</b>	ByteRead	10,000	2,000	2,000
	Speechocean762	2,500	-	2,500

should be noted that the loss function is calculated exclusively based on the masked phonemes and duration. The total loss is calculated by summing the loss values of all the masked tokens across sentences.

### 3.2. Fine-tuning for fluency scoring

In this phase, our objective is to fine-tune the pre-trained model for fluency scoring. The fluency scoring model comprises an encoder and a scorer. Initially, we utilize the pre-trained weights to initialize the encoder of the scoring model. Subsequently, the scorer performs average pooling on a sequence of encoder outputs  $\mathbf{H}$  (as illustrated in Eq. (1)), resulting in an utterance-level fluency representation. This representation is then fed into a linear layer to generate machine score. Mean square error (MSE) calculated between predicted and human-annotated fluency scores are used as the objective for entire network fine-tuning.

## 4. Experimental setup

### 4.1. Speech corpora

The acoustic model was trained on a total of 5,000 hours of English speech data, including 960 hours of native speech from the LibriSpeech [29] and 4,000 hours of non-native private recordings from Bytedance. Additionally, we collected approximately 436 hours (about 200,000 utterances) of reading speech by Chinese L2 adult learners and prompt text for MLM pre-training. To evaluate fluency scoring, we performed experiments on two additional datasets: ByteRead, an internal dataset of 14,000 English utterances collected from Bytedance’s education product (described in detail in [10]), and Speechocean762, an open-source speech assessment corpus consisting of 5,000 utterances collected from 250 speakers [17]. The data statistics were detailed in Table 1.

### 4.2. Feature extraction

Raw fluency features were extracted using the deep feedforward sequential memory network-hidden Markov models (DFSMN-HMM) acoustic model, as described in [30]. The model architecture includes 2 convolutional layers, 24 FSMN layers, a bottleneck layer, and a feedforward layer. The input features were 39-dimensional Mel-frequency cepstral coefficients (MFCCs). The bottleneck layer extracts frame-level deep features with a dimensionality of 512. A HMM-based force-aligner is employed to obtain the phone sequence along with the corresponding start and end time boundaries for each phone.

### 4.3. Setup of proposed and baseline systems

#### 4.3.1. Proposed systems setup

Given the L2 learner’s speech and prompt text, phone-level raw features (deep features, phone sequence, and duration) can be first obtained in the fluency feature extraction module.

The phone sequence is projected into a 32-dim phone embedding, while the 512-dim deep features are transformed into 32-dim features and added to the phone embedding to obtain the compact features. These compact features are then concatenated with a 1-dimensional duration feature, resulting in a 33-dimensional output of the pre-processing. This output serves as input to the pretrain model.

- **Transformer-pre:** The proposed Transformer-based pretrain model. A trainable [CLS] token was appended to the processed feature sequence. And a trainable position embedding and the 33-dimensional processed features were summed together. The pre-trained encoder consists of two transformer layers, with the first layer removing the residual connection to increase the input feature dimension to 128. The multi-head attention block employs 4 heads. The output of the transformer encoder for the [CLS] token, with a dimensionality of 128, was used as the corresponding utterance-level representation for predicting the fluency score.
- **BLSTM-pre:** The proposed BLSTM-based pretrain model. The 64-dim phone-level contextual representations output of the BLSTM encoder will be fed into a mean pooling layer and a linear layer to get the final fluency score. According to our empirical study, an 8-layer BLSTM architecture lead to the best results.

The adam optimization algorithm was employed for updating the pre-training and scoring models in all proposed systems. During pre-training, the batch size was set to 256 and the learning rate was 0.001. For fine-tuning, the batch size was set to 32 and the learning rate was 0.002.

#### 4.3.2. Baseline systems setup

- **BLSTM [10]:** The baseline system without pretraining, where the scorer consists of pre-processing, a 2-layer BLSTM encoder, and a fully connected layer. The input feature and the pre-processing steps were the same as our proposed system as described in 4.3.1.
- **3M-Transformer [9]:** The model input comprises multi-view phone-level features, which consist of prosodic features (duration, energy), SSL features (wav2vec2 [25], HuBERT [27], WavLM [28]), and GOP feature [31]. These features are simply concatenated and subsequently fed into the 3-layer transformer-based scorer to get the fluency score. Multi-granularity pronunciation score labels are used to model the association between different scoring tasks.
- **SSL-IDX-BLSTM [15]:** The system takes the frame-level SSL representations extracted from wav2vec2 Large as input. The k-means clustering algorithm is used to generate the clustered index, which is seen as pseudo phonetic information. A linear layer project the SSL feature into a compact feature, which is concatenated with the index embedding through an embedding layer and fed into 2-layer BLSTM to get the fluency score.

## 5. Results and analyses

In our experiments, the system performance was evaluated using the Pearson correlation coefficient (PCC) between the machine-predicted scores and the human scores. A higher PCC value indicates a better system performance.

Table 2: The PCC performance of different systems on ByteRead and Speechocean762 data sets.

	Model	#Param	Pre-train	Speechocean762	ByteRead
(a)	BLSTM [10]	278K	-	-	0.817
(b)	3M-Transformer [9]	-	-	0.828	-
(c)	SSL+IDX+BLSTM [15]	-	-	0.795	0.828
(d)	Transformer-pre	795K	✗	0.784	0.783
			✓	0.802	0.799
(e)	<b>BLSTM-pre</b>	<b>871K</b>	✗	0.797	0.804
			✓	<b>0.835</b>	<b>0.833</b>

Table 3: The PCC performance of BLSTM-based systems on different scales of the scoring training sets. ByteRead(1000) means 1,000 utterances with prompt text was randomly selected to fine-tune the pre-trained model in the ByteRead training set. Phn and dur loss represent phonetic and prosodic loss, respectively.

Model	Pre-train	Loss	ByteRead(1000)	ByteRead(2500)	ByteRead(5000)	ByteRead	Speechocean762
BLSTM-pre	✗	-	0.669	0.773	0.787	0.804	0.797
	✓	phn+dur	<b>0.787</b>	<b>0.807</b>	<b>0.82</b>	<b>0.833</b>	0.835
	✓	dur	0.78	0.8	0.818	0.826	<b>0.838</b>
	✓	phn	0.734	0.784	0.813	0.82	0.822

### 5.1. Main results

This subsection presents a comparison of the proposed method’s performance on various encoder structures using the Speechocean762 and ByteRead datasets. Additionally, we assessed the proposed method’s effectiveness by comparing its results with baselines. The results of the different systems are presented in Table 2.

First, we evaluated the effectiveness of phonetic and prosody-aware pretrain models using both Transformer and BLSTM architectures, as shown in rows (d) and (e) of Table 2. The results demonstrate that the proposed pretrain model methods consistently outperform their counterparts, which were trained from scratch with labeled data. This suggests that phonetic and prosody-aware pretraining can be beneficial for fluency scoring. Furthermore, we observed that the BLSTM pre-train model outperforms its Transformer counterpart. Hence, BLSTM pretrain model is used in the rest of the experiments.

Apart from the self implemented systems, we also conducted performance comparisons between the proposed BLSTM-pre system and the baseline systems presented in Table 2 (a), (b), and (c). We first compared the performance between proposed BLSTM-pre and the BLSTM system reported in [10]. The results show that our BLSTM-pre outperformed the BLSTM baseline, with an improvement in PCC on the ByteRead database from 0.817 to 0.833. This confirms that the pretrain model is effective for fluency scoring. We then conducted comparisons between our proposed BLSTM-pre system and two SSL feature-based approaches, namely 3M-Transformer and SSL-IDX-BLSTM. Our proposed BLSTM-pre system showed better performance than the 3M-Transformer baseline on the Speechocean762 database, resulting in an increase in PCC from 0.828 to 0.835. Similar results were observed in comparison with SSL-IDX-BLSTM, where our proposed BLSTM-pre system consistently achieved better performance on both Speechocean762 and ByteRead datasets. These findings suggest that our proposed system outperforms state-of-the-art systems for fluency scoring on both Speechocean762 and ByteRead datasets.

### 5.2. Ablation studies

In this section, we conducted a series of experiments to determine the relative importance of the phonetic and prosodic components in the proposed method for scoring fluency. We performed ablation studies by testing different loss function configurations and analyzing the performance of each component. The results are presented in Table 3.

Specifically, we first evaluated the pre-training model’s performance by using only the phonetic aspect as the pre-training loss, which involved predicting the masked phone. Our findings revealed that this approach yielded better results than the no pre-training system. Moreover, we discovered that the prosodic aspect’s contribution to the improvement was more substantial than that of the phonetic aspect. This could be attributed to the duration factor’s significant role in assessing speech fluency. Finally, we combined both phonetics and prosody to optimize the pre-training model, resulting in a more significant improvement, highlighting the effectiveness of the proposed SSL method in fluency scoring.

## 6. Conclusion

This article introduces a self-supervised learning technique that is phonetic and prosody-aware for assessing the fluency of L2 learners’ speech. The method involves masking the phone and duration of input features and then reconstructing them by utilizing a vast amount of unlabeled non-native data during the pre-training phase. To predict the fluency score, a small amount of scoring data was utilized to fine-tune the pre-trained model. Results based on the Speechocean762 datasets and our non-native dataset indicate that the proposed approach outperforms the baseline systems. Our future research aims to explore the benefits of our approach for scoring at various levels (such as phone, and word) and granularities (such as accuracy, and proficiency). Additionally, we plan to explore the impact of using the L1 dataset when pre-training.

## 7. References

- [1] A. Housen and F. Kuiken, "Complexity, accuracy, and fluency in second language acquisition," *Applied linguistics*, vol. 30, no. 4, pp. 461–473, 2009.
- [2] P. Lennon, "Investigating fluency in EFL: A quantitative approach," *Language learning*, vol. 40, no. 3, pp. 387–417, 1990.
- [3] R. Ellis and R. R. Ellis, *The study of second language acquisition*. Oxford University, 1994.
- [4] S. Götz, "Fluency in native and nonnative English speech," *Fluency in Native and Nonnative English Speech*, pp. 1–262, 2013.
- [5] E. Guz, "Establishing the fluency gap between native and non-native-speech," *Research in Language*, vol. 13, no. 3, pp. 230–247, 2015.
- [6] S. Mao, Z. Wu, J. Jiang, P. Liu, and F. K. Soong, "NN-based ordinal regression for assessing fluency of ESL speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7420–7424.
- [7] H. Zhang, K. Shi, and N. F. Chen, "Multilingual speech evaluation: case studies on English, Malay and Tamil," in *Proc. Interspeech 2021*, 2021, pp. 4443–4447.
- [8] 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.
- [9] 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.
- [10] K. Fu, S. Gao, X. Tian, W. Li, Z. Ma, and A. Bytedance, "Using Fluency Representation Learned from Sequential Raw Features for Improving Non-native Fluency Scoring," *Proc. Interspeech 2022*, pp. 4337–4341, 2022.
- [11] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic Pronunciation Assessment using Self-Supervised Speech Representation Learning," in *Proc. Interspeech 2022*, 2022, pp. 1411–1415.
- [12] O. D. Deshmukh, K. Kandhway, A. Verma, and K. Audhkhasi, "Automatic evaluation of spoken English fluency," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4829–4832.
- [13] 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.
- [14] H. Deng, Y. Lin, T. Utsuro, A. Kobayashi, H. Nishizaki, and J. Hoshino, "Automatic fluency evaluation of spontaneous speech using disfluency-based features," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 9239–9243.
- [15] W. Liu, K. Fu, X. Tian, S. Shi, W. Li, Z. Ma, and T. Lee, "An ASR-free Fluency Scoring Approach with Self-Supervised Learning," *arXiv preprint arXiv:2302.09928*, 2023.
- [16] A. Lee *et al.*, "Language-independent methods for computer-assisted pronunciation training." Ph.D. dissertation, Massachusetts Institute of Technology, 2016.
- [17] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "Speechocean762: An Open-Source Non-Native English Speech Corpus for Pronunciation Assessment," in *Proc. Interspeech 2021*, 2021, pp. 3710–3714.
- [18] N. F. Chen, D. Wee, R. Tong, B. Ma, and H. Li, "Large-scale characterization of non-native Mandarin Chinese spoken by speakers of European origin: Analysis on iCALL," *Speech Communication*, vol. 84, pp. 46–56, 2016.
- [19] L. Yang, K. Fu, J. Zhang, and T. Shinozaki, "Pronunciation Error-prone Tendency Detection with Language Adversarial Representation Learning," in *Proc. Interspeech 2020*, 2020, pp. 3042–3046.
- [20] L. Peng, K. Fu, B. Lin, D. Ke, and J. Zhan, "A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis," in *Proc. Interspeech 2021*, 2021, pp. 4448–4452.
- [21] 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423>
- [22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [23] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1441–1451.
- [24] E. Kharitonov, A. Lee, A. Polyak, Y. Adi, J. Copet, K. Lakhota, T. A. Nguyen, M. Riviere, A. Mohamed, E. Dupoux, and W.-N. Hsu, "Text-free prosody-aware generative spoken language modeling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8666–8681. [Online]. Available: <https://aclanthology.org/2022.acl-long.593>
- [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.
- [26] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.
- [27] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, 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.
- [28] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [30] S. Zhang, M. Lei, Z. Yan, and L. Dai, "Deep-FSMN for large vocabulary continuous speech recognition," in *ICASSP 2018-2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5869–5873.
- [31] 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.