INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A Joint Model for Pronunciation Assessment and Mispronunciation Detection and Diagnosis with Multi-task Learning

Hyungshin Ryu[1], *Sunhee Kim*[2], *Minhwa Chung*[1]

[1]Department of Linguistics, Seoul National University, Republic of Korea
[2]Department of French Language Education, Seoul National University, Republic of Korea
{rhss10, sunhkim, mchung}@snu.ac.kr

Abstract

Empirical studies report a strong correlation between pronunciation proficiency scores and phonetic errors in non-native speech assessments of human evaluators. However, the existing system of computer-assisted pronunciation training (CAPT) regards automatic pronunciation assessment (APA) and mispronunciation detection and diagnosis (MDD) as independent and focuses on individual performance improvement. Motivated by the correlation between two tasks, we propose a novel architecture that jointly tackles APA and MDD using CTC and cross-entropy criteria with a multi-task learning scheme to benefit both tasks. To leverage additional knowledge transfer, Wav2Vec2-robust finetuned on TIMIT is used for the joint optimization. The integrated model significantly outperforms single-task learning, with a mean of 0.057 PCC increase for APA and 0.004 F1 increase for MDD on Speechocean762, which reveals that proficiency scores and phonetic errors are correlated for both human and model assessments.

Index Terms: computer-assisted pronunciation training, multi-task learning, mispronunciation detection and diagnosis, automatic pronunciation assessment, transfer learning

1. Introduction

The incorporation of speech technology into education has consistently grown and has brought meaningful results [1, 2]. The field of computer-assisted pronunciation training (CAPT) has similarly made rapid progress, with the spread of internet-based applications and the significant development of automatic speech recognition (ASR) technology. The CAPT system serves as a powerful tool for non-native learners, as it provides customized feedback at a low cost. The minimized time and place constraints typical in traditional instructor-based learning bring another advantage to CAPT [3].

The CAPT system generally consists of two major tasks, automatic pronunciation assessment (APA) and mispronunciation detection and diagnosis (MDD). APA task can be seen as a speech classification task, that aims to provide pronunciation proficiency scores highly correlated with those of human evaluators [4, 5, 6, 7, 8, 9]. The scores reflect different aspects of scoring standards (accentedness, fluency, comprehensibility) [10] or different granularity (phones, words, sentences) [11]. MDD task on the other hand is a non-native phone recognition task. It aims to correctly classify and diagnose the realized phones into correct pronunciations and mispronunciations, by comparing them with the annotated phone transcriptions of human experts and the canonical phone sequences [12, 13, 14, 15, 16, 17, 18]. As both aim to assess non-native (L2) speech, the two tasks inevitably share similar methodologies, including the usage of Goodness of Pronunciation (GOP) measure [4, 5, 6, 7, 12], hand-crafted acoustic features [5, 7, 14], the usage of native (L1) data [9, 13, 14, 16, 17, 18], to the recent pre-trained self-supervised learning model [7, 8, 9, 15, 16, 18].

Indeed, empirical studies report that there exists a distinct correlation between pronunciation proficiency scores and mispronunciations that are annotated by human evaluators for non-native speech assessments. Phonetic errors showed a strong correlation with not only overall assessment such as comprehensibility scores of L2 German [10] and holistic scores of L2 Korean [19], but also prosodic assessment such as fluency scores of L2 Mandarin [20], and accent scores of L2 English [21]. This applied to both cases where mispronunciation annotators and score annotators were different [10, 20, 21] or the same [19]. This provides strong linguistic motivation to leverage the correlation between APA and MDD tasks to benefit each other.

However, the current CAPT system has treated the two tasks as independent and separate. One reason lies in the proposals focusing on improving the model performance on different benchmark datasets for the respective task. L2-arctic [22] is often used to test MDD performance, while Speechocean762 [11] is used for APA, as in [6, 7] that propose multi-aspect multi-granular APA, and [8] that adopt self-supervised learning for holistic/fluency/prosodic APA. Other studies use Speechocean762 only as an MDD benchmark, including [16] that shows their data augmentation method improves MDD performance on out-of-domain Speechocean762 test set, and [18] that adopts a non-autoregressive framework for MDD.

A few studies that mention both tasks still regard one as auxiliary or separate. [4] regards mispronunciation detection as an auxiliary, binary phone-level scoring in multi-granular APA. [9] utilizes native(-like) data and the matching canonical phones for auxiliary CTC training to assist holistic/accuracy APA. Phone-level APA is performed in [17] along with MDD, but as two separate tasks that can be achieved with respective APA and MDD datasets for fine-tuning. However, given the significant linguistic correlation between the two tasks, an integrated model of the two tasks is expected to improve performances on both tasks.

This paper presents for the first time a novel architecture that jointly trains pronunciation assessment task and mispronunciation detection and diagnosis task via a multi-task learning perspective, to leverage their correlation. To further enhance the acoustic representation of the model, we adopt transfer learning to fine-tune a self-supervised learning model on phone recognition before multi-task learning. We contribute by verifying that the joint model shows distinct improvement on both APA and MDD tasks, compared to the respective single-task learning on Speechocean762. Additional analyses show how different loss weights, self-supervised learning model, and transfer learning dataset influence the joint model. Moreover, correlation analy-

10.21437/Interspeech.2023-337

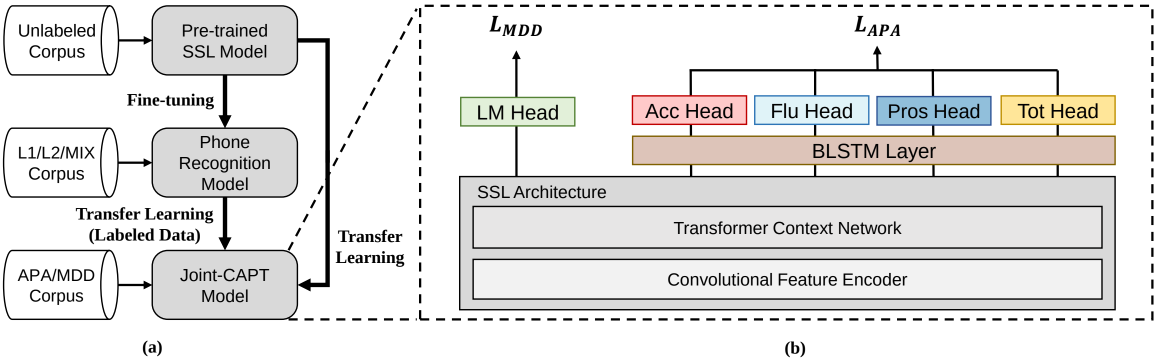

Figure 1: *The overview of our proposed method. (a) the training process, (b) the architecture of the joint model for APA and MDD.*

sis on proficiency scores and phonetic errors for both model predictions and Speechocean762 human annotations reveals that the joint model leveraged their correlation to gain performance improvements, which proves the importance of this study.

2. Proposed method

The proposed method is composed of two steps: **(1) transfer learning** with an auxiliary fine-tuning of self-supervised learning model on phone recognition and the main **(2) multi-task learning** of APA and MDD as shown in Fig. 1 (a). We release the source code for reproducibility. [1]

2.1. Transfer learning

Transfer learning (TL) takes a resource-rich, huge model from another domain to adapt to the target domain. As the CAPT domain suffers from the inherent problem of data scarcity, we utilize a self-supervised learning (SSL) model pre-trained with a vast amount of unlabeled data for the backbone model to leverage its robustness. We explore four different SSL models with the same 300 million parameters but with different datasets and learning schemes:

- **Wav2Vec2-robust** [23]: Trained with 63K hrs. of English, the model consists of a convolutional feature encoder, a Transformer context network, and a quantization module.
- **Wav2Vec2-XLS-R** [24]: A multilingual model trained with 436K hrs. of data from 128 languages, it has the same architecture as the robust model.
- **HuBERT** [25]: Trained with 60K hrs. of English, the architecture is based on Wav2Vec2. Iterative K-means clustering is used for masked prediction.
- **WavLM** [26]: Trained with 94K hrs. of English, the model extends HuBERT. Masked speech denoising and additional gated relative position bias are implemented.

We additionally fine-tune the SSL model on phone recognition before multi-task learning to see if the extra fine-tuning can enhance the model with better acoustic representation. For fine-tuning, a fully-connected layer (language model head) is added on top of the Transformer network of the SSL model to train with Connectionist Temporal Classification (CTC) loss.

2.2. Multi-task learning

Multi-task learning (MTL) simultaneously trains tasks with different objective functions using a shared model. With the increased information, downstream speech tasks including emotion recognition [27] or dysarthria assessment [28] have leveraged the framework to gain more generalized models. Motivated by its effectiveness in various speech domains, we utilize MTL to jointly train APA and MDD. With the joint optimization, we expect the model to learn the correlation between the output pronunciation scores and phone sequences to gain performance increases than respective single-task learning (STL).

Fig. 1 (b) shows the architecture of the joint model. The model utilizes the SSL encoder and its weights. For the raw audio input $x \in \mathbb{R}^L$ with length L, the SSL encoder outputs T sequences of D dimensional latent speech representation $h \in \mathbb{R}^{T\times D}$. For the APA task, the latent speech representation goes through an additional bidirectional long short-term memory (BLSTM) layer shared among four pronunciation assessment tasks to capture the information shared between assessments. The model yields $\bar{h} \in \mathbb{R}^{T\times H}$ where H is the size of the output hidden dimension. The output representation is then passed to each assessment head which consists of a fully connected layer and an average pooling over time dimension, to make $\hat{y}_{\{acc,flu,pros,tot\}} \in \mathbb{R}^C$, which are logits of accuracy, fluency, prosodic, and total score, respectively where C is the number of labels. Logits of each aspect are stacked to make final APA logits $\hat{y} \in \mathbb{R}^{C\times 4}$ to be optimized using cross-entropy criteria (L_{APA}) with the ground-truth score labels after a softmax operation. For the MDD task, the latent speech representation h passes the same fully connected layer used in phone recognition fine-tuning to leverage the fine-tuned weights. The output logit $\hat{z} \in \mathbb{R}^{T\times V}$ is optimized using CTC loss (L_{MDD}) with the ground-truth realized phone annotations after a softmax operation, where V is the size of the vocabulary.

The classification heads and the language model head are then optimized using the joint loss L_{CAPT}, which is a combination of L_{APA} and L_{MDD}:

$$L_{CAPT} = \alpha L_{APA} + \beta L_{MDD} \tag{1}$$

where α and β are used to balance the two losses. α is chosen from the set of $\alpha \in \{0.1, 0.25, 0.5, 1\}$ and β is fixed to 1.0. This is to adjust the weights on L_{APA} as L_{MDD} takes advantage of auxiliary fine-tuning and thus optimizes faster.

3. Experiments

3.1. Datasets

Experiments are conducted using three public datasets, TIMIT [29], L2-arctic [22] for the auxiliary fine-tuning on phone

[1] `https://github.com/rhss10/joint-apa-mdd-mtl`

Table 1: *Experiment results w.r.t transfer learning and multi-task learning. - refers to tasks not performed for STL.*

Model	Pronunciation Scores (PCC)				PER	Correct Pronunciations			Mispronunciations		
	Accuracy	Fluency	Prosodic	Total		Precision	Recall	F1	Precision	Recall	F1
APA-SSL	0.609	0.652	0.650	0.633	-	-	-	-	-	-	-
MDD-SSL	-	-	-	-	**9.89%**	0.997	0.928	0.961	0.267	0.914	0.413
Joint-CAPT-SSL	0.714	0.763	0.767	0.732	9.91%	0.997	**0.929**	**0.962**	**0.268**	0.914	**0.415**
APA-L1	0.629	0.738	0.733	0.680	-	-	-	-	-	-	-
MDD-L1	-	-	-	-	9.90%	0.997	0.927	0.961	0.265	**0.916**	0.410
Joint-CAPT-L1	**0.719**	**0.775**	**0.773**	**0.743**	9.93%	0.997	0.928	**0.962**	0.267	0.914	0.414

Table 2: *Model performances compared w.r.t multi-task learning loss weight.* $\alpha = 0.25$ *is the baseline.*

Loss Weight	Pronunciation Scores (PCC)				PER	Correct Pronunciations			Mispronunciations		
	Accuracy	Fluency	Prosodic	Total		Precision	Recall	F1	Precision	Recall	F1
$\alpha = 0.25$	0.719	0.775	**0.773**	**0.743**	**9.93%**	0.997	**0.928**	**0.962**	0.267	0.914	0.414
$\alpha = 0.1$	0.700	**0.776**	0.771	0.737	10.00%	0.997	0.927	0.961	0.262	0.909	0.407
$\alpha = 0.5$	**0.724**	0.763	0.765	0.741	9.94%	0.997	**0.928**	**0.962**	**0.269**	**0.916**	**0.415**
$\alpha = 1.0$	0.718	0.768	0.764	0.737	10.10%	0.997	0.927	0.961	0.265	0.912	0.411

recognition, and Speechocean762 [11] for the joint APA and MDD training. TIMIT is a native speech dataset that contains recordings of 8 US English dialects and is phonetically transcribed with 61 phone set. L2-arctic v5.0 is a non-native speech dataset that contains English of 6 L1 backgrounds transcribed with 40 phone set. For fine-tuning, we use the original TIMIT train split, and the suggested L2-arctic train split from [13, 17].

SpeechOcean762 contains non-native English recordings of Mandarin speakers, of which we use the original train/test split. For APA task, we use four aspects of sentence scores, accuracy, fluency, prosodic, and total, that range from 0-10. Speechocean762 provides an extra mispronunciation transcription for inaccurate phones using 46 phone set, 39 phones following CMUDict [30], <unk> for unknown phones, and 6 L2 phones. This realized phone transcription is used for MDD task. Roughly 4% and 3% of the train and test set phone annotations are mispronunciations. Out of the mispronunciations, <unk> takes up to 26% and 25%, respectively. The phone sets of TIMIT and L2-arctic were mapped into CMUDict to be combined with SpeechOcean762 phone set and were used for both phone recognition fine-tuning and multi-task learning.

3.2. Evaluation metrics

The APA performance is measured using Pearson Correlation Coefficient (PCC) between model prediction scores and human annotated scores. For MDD performance, Precision, Recall, and F1 scores are calculated according to the metrics used in [31, 32] and are reported for both correct pronunciations (CP) and mispronunciations (MP) following [18]. True Acceptance (TA) refers to predicting CP as CP, False Acceptance (FA) refers to predicting MP as CP, True Rejection (TR) refers to predicting MP as MP, and False Rejection (FR) refers to predicting CP as MP. As metrics are reported for both CP and MP, classes used for metrics contradict each other. For example, True Positive (TP) corresponds to TA for CP and TR for MP. This applies to False Positive (FP), True Negative (TN), and False Negative (FN) as well, where $Precision = TP/(TP + FP)$, $Recall = TP/(TP + FN)$, $F1 = 2 \times Precision \times Recall/(Precision + Recall)$.

3.3. Implementation details and baseline

For all procedures, models had the feature encoder frozen and were trained with 8 batch sizes, an AdamW optimizer, a training epoch of 100, and a linear scheduler with a learning rate of 1e-4 and a warm-up ratio of 0.1. Pre-trained SSL models were implemented using HuggingFace [33]. For multi-task learning, all the experiments were repeated for 3 trials with different random seeds and are reported with the mean value. The BLSTM layer was fixed to 128 hidden dimensions.

For the main results, we utilize Wav2Vec2-robust as the backbone model, TIMIT (**L1**) as the phone recognition fine-tuning dataset, and α=0.25 for the joint loss weight. For the joint model without additional fine-tuning, α was set to 0.0 for the first 50 epochs, as CTC loss is much bigger than cross-entropy loss without fine-tuning. Section 3.4 shows the main results of the proposed method. Section 4.1 tries on different loss weights, backbone SSL model (**XLS-R**, **HuBERT**, **WavLM**), and dataset (L2-arctic (**L2**), a sum of TIMIT/L2-arctic (**MIX**)) to compare their influence on model performance.

3.4. Experiment results

Table 1 demonstrates the experiment results with regard to transfer learning, and multi-task learning of APA and MDD. The proposed joint model (**Joint-CAPT** [$\alpha = 0.25, \beta = 1.0$]) is compared to respective STL models (**APA** [$\alpha = 1.0, \beta = 0.0$], **MDD** [$\alpha = 0.0, \beta = 1.0$]), for both raw self-supervised model (**SSL**) and fine-tuned model (**L1**). First, multi-task learning greatly improves APA performance, with Joint-CAPT-SSL and Joint-CAPT-L1 both having higher PCC for all scores than the respective APA-SSL and APA-L1, with an average of 0.108 and 0.057 increase. The average PCC of Joint-CAPT-SSL is even higher than APA-L1 which leverages the extra knowledge by a mean of 0.049. Second, although more subtle than APA performance increase, multi-task learning also improves MDD performance. For both correct pronunciations and mispronunciations, Joint-CAPT-SSL (0.962, 0.415) and Joint-CAPT-L1 (0.962, 0.414) have higher F1 scores than the respective MDD-SSL and MDD-L1. The performance gain was achieved from higher recall for CP, and higher precision for MP. Altogether, this proves the effectiveness of jointly training APA and MDD.

Table 3: *Model performances compared w.r.t backbone SSL model and dataset used for fine-tuning. Robust, L1 is the baseline.*

Transferred Model	Pronunciation Scores (PCC)				PER	Correct Pronunciations			Mispronunciations		
	Accuracy	Fluency	Prosodic	Total		Precision	Recall	F1	Precision	Recall	F1
Robust, L1	**0.719**	**0.775**	**0.773**	**0.743**	9.93%	0.997	0.928	0.962	0.267	0.914	0.414
XLS-R, L1	0.685	0.759	0.756	0.711	13.49%	**0.998**	0.895	0.943	0.203	**0.936**	0.334
HuBERT, L1	0.694	0.763	0.760	0.730	10.12%	0.997	0.927	0.961	0.261	0.901	0.405
WavLM, L1	0.704	0.766	0.757	0.728	**9.42%**	0.997	**0.932**	**0.963**	**0.273**	0.886	**0.418**
Robust, L2	0.710	0.773	0.767	0.734	10.06%	0.997	0.927	0.961	0.264	0.914	0.409
Robust, MIX	0.707	0.769	0.765	0.729	9.95%	0.997	0.928	0.961	0.265	0.908	0.410

Lastly, similar to multi-task learning, auxiliary phone recognition fine-tuning significantly improves APA performance, for both APA-L1 and Joint-CAPT-L1. However, it slightly reduces MDD performance as F1 scores were reduced for both MDD-L1 and Joint-CAPT-L1, caused by lower precision of MD. The mismatch between the transferred L1 and the target L2 data may be the cause, as dialectal differences of TIMIT may be misled to mispronunciations.

4. Discussion

4.1. Analysis on loss weight and transferred model

We additionally explore the influence of **(1) loss weight**, and **(2) transferred model** on model performances. First for α, Table 2 shows that smaller weights on L_{APA} (0.1, 0.25) help the model achieve better results on fluency and prosodic correlation. Increasing the weights (0.25, 0.5) results in better accuracy and total scores, with better MDD performances as well. Yet, the performance decreases as the weight gets too big. (1.0) Overall, $\alpha = 0.25$ achieved the most decent performance for joint APA/MDD.

For the SSL model, XLS-R consistently showed the weakest performance in both APA and MDD as in Table 3. As XLS-R had been pre-trained with multi-lingual data, the model may not have been fit for the scope. WavLM surpassed the baseline model performance on F1 scores. As HuBERT which had a similar amount of pre-training data showed conflicting results, the MDD performance improvement of WavLM may lie in its learning scheme. For the dataset, using L1 for transfer learning outperformed both L2 and MIX. This is an interesting finding given that L2 has more similar acoustic characteristics to the target Speechocean762 data, and the mix of the two has a larger amount of data.

4.2. Correlation analysis of human and model assessments

To explore how the model leverages the correlation between proficiency scores and phonetic errors, correlation analysis was conducted using PCC for both human evaluators and predictions of the proposed Joint-CAPT-L1. The test set of Speechocean762 was used for analysis. The results are plotted using linear regression in Fig. 2. For both plots, accuracy, fluency, prosodic, and total score all showed a correlation with mispronunciations and were statistically significant (p<.001).

Specifically, for human evaluators, the total score had the highest negative correlation with mispronunciations (r=-0.656), followed by accuracy (r=-0.624), fluency (r=-0.606), and prosodic (r=-0.593). This suggests that the human assessors of Speechocean762 were influenced by phone errors when grading the scores for all aspects, which complies with the findings of previous studies. Interestingly, model predictions also showed a similar pattern where accuracy (r=-0.541) and total score (r=-0.534) had the highest negative correlation with mispronunciation, followed by prosodic (r=-0.476) and fluency (r=-0.461). This also corresponds to the performance results of Joint-CAPT-L1, where accuracy and total score gained the most performance increase compared to APA-L1. In other words, the statistical analysis provides evidence that the integrated model leveraged the correlation between APA and MDD tasks to gain performance improvement.

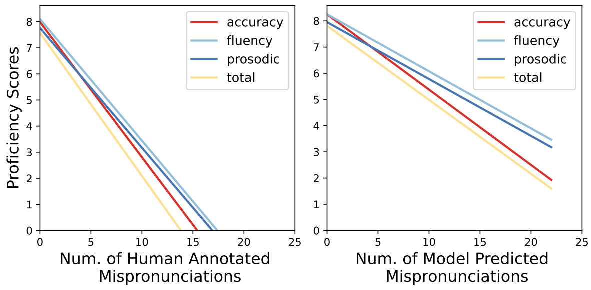

Figure 2: *Correlation between proficiency scores of four aspects and the number of mispronunciations are plotted for both human evaluations and model predictions.*

5. Conclusions

This study presents a novel architecture that jointly trains automatic pronunciation assessment and mispronunciation detection and diagnosis with a multi-task learning perspective, motivated by the high linguistic correlation between proficiency scores and phonetic errors. The significant performance improvement of the proposed joint model over single-task APA and MDD on Speechocean762 proves that the correlation between two tasks can benefit each other, which is further supported by correlation analysis. The proposed model not only conforms to the linguistic mechanism of non-native speech assessment, but shows its usefulness in practical assessment scenarios where learners are graded in various aspects with a single utterance.

6. Acknowledgements

This paper is supported by the Ministry of Science and ICT and the National Information Society Agency under the project titled ”Voice Data of the Korean Language Construction Project for Language Education - Spoken by Western and Asian Language Users” (2022-Data-WI57)

7. References

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

[2] D. Litman, H. Strik, and G. S. Lim, "Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities," *Language Assessment Quarterly*, vol. 15, no. 3, pp. 294–309, 2018.

[3] P. M. Rogerson-Revell, "Computer-assisted pronunciation training (capt): Current issues and future directions," *RELC Journal*, vol. 52, no. 1, pp. 189–205, 2021.

[4] B. Lin, L. Wang, X. Feng, and J. Zhang, "Automatic scoring at multi-granularity for l2 pronunciation." in *Interspeech*, 2020, pp. 3022–3026.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] B. Lin and L. Wang, "Exploiting information from native data for non-native automatic pronunciation assessment," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 708–714.

[10] M. G. O'Brien, "L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative german speech," *Language Learning*, vol. 64, no. 4, pp. 715–748, 2014.

[11] 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.

[12] W. Hu, Y. Qian, and F. K. Soong, "An improved dnn-based approach to mispronunciation detection and diagnosis of l2 learners' speech." in *SLaTE*, 2015, pp. 71–76.

[13] Y. Feng, G. Fu, Q. Chen, and K. Chen, "Sed-mdd: Towards sentence dependent end-to-end mispronunciation detection and diagnosis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3492–3496.

[14] K. Fu, J. Lin, D. Ke, Y. Xie, J. Zhang, and B. Lin, "A full text-dependent end to end mispronunciation detection and diagnosis with easy data augmentation techniques," *arXiv preprint arXiv:2104.08428*, 2021.

[15] M. Yang, K. Hirschi, S. D. Looney, O. Kang, and J. H. Hansen, "Improving Mispronunciation Detection with Wav2vec2-based Momentum Pseudo-Labeling for Accentedness and Intelligibility Assessment," in *Proc. Interspeech 2022*, 2022, pp. 4481–4485.

[16] D. Zhang, A. Ganesan, S. Campbell, and D. Korzekwa, "L2-GEN: A Neural Phoneme Paraphrasing Approach to L2 Speech Synthesis for Mispronunciation Diagnosis," in *Proc. Interspeech 2022*, 2022, pp. 4317–4321.

[17] N. Zheng, L. Deng, W. Huang, Y. T. Yeung, B. Xu, Y. Guo, Y. Wang, X. Chen, X. Jiang, and Q. Liu, "CoCA-MDD: A Coupled Cross-Attention based Framework for Streaming Mispronunciation Detection and Diagnosis," in *Proc. Interspeech 2022*, 2022, pp. 4352–4356.

[18] M. A. H. Wadud, M. Alatiyyah, and M. F. Mridha, "Non-autoregressive end-to-end neural modeling for automatic pronunciation error detection," *Applied Sciences*, vol. 13, no. 1, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/1/109

[19] S.-H. Yang and M. Chung, "Linguistic factors affecting evaluation of l2 korean speech proficiency." in *SLaTE*, 2017, pp. 53–58.

[20] 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.

[21] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language learning*, vol. 45, no. 1, pp. 73–97, 1995.

[22] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-arctic: A non-native english speech corpus." in *Interspeech*, 2018, pp. 2783–2787.

[23] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Proc. Interspeech 2021*, 2021, pp. 721–725.

[24] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[25] 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.

[26] 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.

[27] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning." in *Interspeech*, vol. 2021, 2021, pp. 4508–4512.

[28] E. J. Yeo, K. Choi, S. Kim, and M. Chung, "Automatic severity classification of dysarthric speech by using self-supervised model with multi-task learning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[29] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[30] R. Weide *et al.*, "The carnegie mellon pronouncing dictionary," *release 0.6, www. cs. cmu. edu*, 1998.

[31] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2016.

[32] W.-K. Leung, X. Liu, and H. Meng, "Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8132–8136.

[33] T. Wolf *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*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6