INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

A study on fine-tuning wav2vec2.0 Model for the task of Mispronunciation Detection and Diagnosis

Linkai Peng[1], *Kaiqi Fu*[1], *Binghuai Lin*[2], *Dengfeng Ke*[1], *Jinsong Zhang*[1]

[1]Beijing Language and Culture University, China
[2]Smart Platform Product Department,Tencent Technology Co., Ltd, China

penglinkai96@gmail.com,
kaiq.fu@gmail.com,binghuailin@tencent.com,dengfeng.ke@blcu.edu.cn,jinsong.zhang@blcu.edu.cn

Abstract

Mispronunciation detection and diagnosis (MDD) technology is a key component of computer-assisted pronunciation training system (CAPT). The mainstream method is based on deep neural network automatic speech recognition. Unfortunately, the technique requires massive human-annotated speech recordings for training. Due to the huge variations in mother tongue, age, and proficiency level among second language learners, it is difficult to gather a large amount of matching data for acoustic model training, which greatly limits the model performance. In this paper, we explore the use of Self-Supervised Pretraining (SSP) model wav2vec2.0 for MDD tasks. SSP utilizes a large unlabelled dataset to learn general representation and can be applied in downstream tasks. We conduct experiments using two publicly available datasets (TIMIT, L2-arctic) and our best system achieves 60.44% f1-score. Moreover, our method is able to achieve 55.52% f1-score with 3 times less data, which demonstrates the effectiveness of SSP on MDD[1].

Index Terms: self-supervised, mispronunciation detection and diagnosis (MDD), computer-aided pronunciation training (CAPT), wav2vec 2.0, pre-training

1. Introduction

Computer-Assisted Pronunciation Training (CAPT) system can meet people's needs for language learning in fragmented time with flexible devices. Mispronunciation detection and diagnosis system (MDD) is an indispensable component of the CAPT system. Similar to the role of teachers in oral practice lessons, MDD can provide instant feedback about pronunciation problems for users to improve their speaking skills. Considering the rapidly increasing number of language learners, a high-performance MDD is needed to provide precise diagnoses of pronunciation errors at the phonetic and prosodic levels. Here, we focus on phonetic mispronunciation in second-language learning.

The pronunciation error detection framework can be roughly divided into two categories. The first method is based on confidence measures mainly obtained from automatic speech recognition (ASR). Whether the pronunciation is correct or not is judged by calculating the confidence score of the frame/phoneme level with the help of forced alignment[1, 2, 3]. The second is based on extended search lattice and one of the most popular approaches is extended recognition network (ERN) [4]. ERN incorporates a finite number of phonetic error patterns into the decoding network to provide detailed diagnostic information related to the possible error patterns. However, some limitations exist in these methods: 1) Confidence measures based approaches lack the ability to provide specific diagnostic information; 2) ERN cannot guarantee that all mispronunciations are covered. Unseen mispronunciations will lead to bad performance; 3) Multistage systems usually have complex structures whose building process is laborious. Inspired by deep neural network (DNN) based ASR framework, the Connectionist Temporal Classification (CTC) [5] based End-to-End (E2E) method was applied to MDD and achieved promising performance [6, 7]. The end-to-end modeling approach avoids the complicated modeling process and hand-designed dictionaries. Though effective, this kind of method requires massive supervised corpora to make all the weights in the network fully trained. Note that MDD is a data-scarce task. Large-scale non-native speech data is difficult to collect and the corresponding annotations rely on the support of experienced experts, which is very costly and time-consuming.

A developing research area of deep learning, transfer learning has the potential to alleviate the problem of data scarcity. In particular, reusing a pre-trained model or applying general feature representation learned from related tasks to downstream tasks has been proved successful in several domains [8, 9, 10]. A related sub-filed, self-supervised pretraining (SSP) attempts to learn powerful context representation from unlabeled data with the target computed from the input signal itself or modifications of the input [11, 12]. In speech processing, several SSP models have been introduced: APC [13], Mockingjay [14], TERA [15], wav2vec [16] and its variation (vq-wav2vec [17]; wav2vec2.0 [12]). Most of them are becoming the building block in many applications, such as phoneme recognition, speech translation [18] and ultra-low resource ASR [19]. In this paper, we focus on wav2vec2.0 which has consistently achieved state-of-the-art (SOTA) results in many tasks and explore its application in MDD tasks. Wav2vec2.0 uses convolutional neural networks (CNN) to encode the raw waveform input and the Transformer structure to contextualize the latent representation. Contrastive Predict Coding (CPC) [20] criterion and vector quantization (VQ) [21] are adopted to make the pre-training process more reasonable and effective. In [22], CPC is applied to pre-train model with 150 hours of unlabeled data. Experiments performed on private datasets preliminary show the effectiveness of self-supervised pre-training on MDD.

Based on the above situation, the two main purposes of this paper are: 1) to introduce a public pre-training model to the E2E MDD task. The availability of pre-trained models trained on large-scale data (over 50k hours) can save researchers large amounts of time and computational cost; 2) to compare the performance of the proposed method under different configurations. Even if we focused our experiments on the model wav2vec2.0, our exploration can provide insight to any other

[1] Configuration files and code are available at https://github.com/vocaliodmiku/wav2vec2mdd

http://dx.doi.org/10.21437/Interspeech.2021-1344

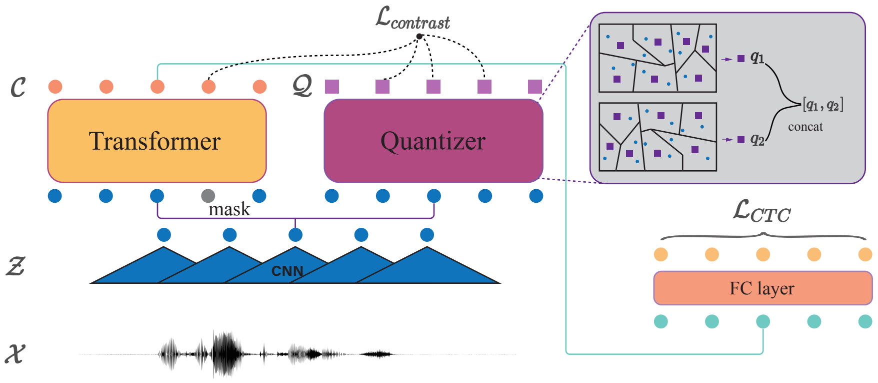

Figure 1: *Left: The wav2vec2.0 framework and corresponding criterion in pre-training stage. Right: Visualisation of the vector quantization module and the diagram of applying wav2vec2.0 to MDD task with CTC loss.*

model pre-trained with SSP. The rest of this paper is organized as follows: Section 2 presents the proposed MDD framework. The experimental setup and the results are presented in Section 3. Finally, conclusions and future work are presented in Section 4.

2. Method

We first briefly introduce the model structure of wav2vec2.0 and the self-supervised pre-training method used. Then we describe how to apply the pre-trained model to MDD task. The pre-trained model we use comes from fairseq toolkit [23].

2.1. Pre-trained Wav2vec2.0 model

The wav2vec2.0 model architecture is shown in Figure 1. It consists of a CNN-based encoder network, a transformer-based context network and a vector quantization module. The encoder network $f : \mathcal{X} \mapsto \mathcal{Z}$ encodes the raw audio sample $x_i \in X$ into latent speech representation $(z_1, z_2, ..., z_T)$ using seven blocks of temporal convolution layer. Followed by layer normalization and GELU activation layer, the convolutions in each block have 512 channels with strides (5,2,2,2,2,2,2) and kernel sizes (10,3,3,3,3,2,2), which compress about 25ms of 16kHZ audio every 20ms. Then the representation is fed into the context network $g : \mathcal{Z} \mapsto \mathcal{C}$ to build context representations $c_i = g(z_i, ..., z_T)$ over the entire latent speech representations. The context network consists of 24 blocks with model dimension 1024, inner dimension 4096 and 16 attention heads.

During pre-training, a certain proportion of consecutive time steps are masked in the latent representation Z and then CPC is performed. Given the contextual representation conditioned on the masked Z, the objective of CPC is to distinguish the latent representation from a series of distractors sampled from other masked time steps. Different from the APC which attempts to reconstruct the high-dimensional signal itself, CPC is an easier learning objective. Before the contrast training, VQ $g : \mathcal{Z} \mapsto \mathcal{Q}$ is used to discretize the unmasked Z in continuous space to a finite number of entries. Considering the discrete nature of phonetic units, VQ is applied to learning meaningful speech representation that correspond to the underlying speech units without supervision, and the authors report that the discrete representations learned are related to phonemes. The quantization module has $G = 2$ codebooks and $V = 320$ entries each. All the discrete entries selected are concatenated and fed into the contrast training task.

We make an assumption that these general representations learned could be useful for MDD.

2.2. Fine-tuning

Once the pre-training is completed, wav2vec2.0 can be treated as a front-end feature extractor or encoder. It is worth mentioning that in [12], a fully connected layer stacked on wav2vec2.0 encoder achieves the SOTA performance of phone recognition on TIMIT. Similarly, we choose to simply add a fully connected layer here to show the effectiveness of the SSP model with discretized speech representations on the MDD task. We handle the MDD task in phone recognition style: the model takes the original speech as input and outputs the phoneme sequence. The training criterion is CTC. We refer to wav2vec2.0 as w2v2.0 for simplicity.

3. Experiments

3.1. Dataset

We use the publicly available data sets TIMIT [24] and L2-arctic [25] to conduct our experiments. TIMIT is a native (L1) English corpus containing 6,300 utterances from 630 speakers. We use its original training subset. The L2-arctic corpus is a non-native English speech corpus that is intended for research in voice conversion, accent conversion, and mispronunciation detection. It contains utterances with mispronunciation of 24 (12 males and 12 females) non-native speakers whose L1 languages include Hindi, Korean, Spanish, Arabic, Vietnamese and Chinese. In order to merge these two datasets for training, we use the open-source tool SoX [26] to reduce the sampling rate of L2-arctic data to 16000HZ with default parameters. For the phone set, we map the TIMIT 61-phone to 39-phone according the mapping table from [27] and combine it into L2-arctic phone set.

3.2. Implementation Details

In our experiments, we used publicly available pre-trained wav2vec2.0 models: wav2vec2.0-LARGE, wav2vec2.0-LV60 and wav2vec2.0-XLSR. These three models use the same architecture but different amounts of data for pre-training. The

LARGE model uses 960h hours of Librispeech [28] for pre-training and the LV60 model uses 53,200 hours of LibriVox. For the XLSR model, consisting of 53 languages, 56000 hours of speech data is used for pre-training. According to the detailed statistic of each language, the total duration of languages that we are interested in is less than 53000 hours. To better understand the influence of the amount and type of data during fine-tuning, we conducted experiments with various training data configurations described below and Table 1 summarizes related duration statistics.

- **Default** In the default configuration, we solely use L2-arctic data as the training data. The division of the training and test set is consistent with prior work [29], i.e. six speakers (NJS, TLV, TNI, TXHC, YKWK, ZHAA) are selected as test set while the rest are merged to build training set. We further generate a subset from training set as development set.
- **-33%** Considering the success of the pre-trained model on low-resource tasks, we try to conduct experiments with less L2 data to explore the feasibility of ultra-low resource MDD. Specifically, the number of speakers for each language on the training set dropped from three to two, i.e. six speakers are removed from the default training set.
- **-66%** Same as above, the number of speakers for each language in the training set is further reduced from two to one, i.e. another six speakers are removed from the above training set.
- **+TIMIT** Due to the scarcity of labeled L2 data, most of the previous works utilized L1 data to ensure that the ASR-based model has the ability to identify the correct part of pronounced utterances accurately. We use this data configuration to compare with previous works.

Table 1: *The total duration in hours at different data configurations*

	Train	Dev	Test
Default	2.50	0.28	0.88
-33%	1.49	0.37	0.88
-66%	0.73	0.19	0.88
+TIMIT	6.07	0.28	0.88

In the case of **-33%** and **-66%**, we generate all possible permutations of these speakers and randomly sample six distinct combinations to conduct experiments. We report the average results in Table 3.

The target sequence used is the manually labeled annotation sequence. We trained our models on one Tesla K80 GPU using Adam optimizer, and all models are trained with the same number of steps approximately (400 epochs for default data configuration). Most of the training configurations are consistent with the open-source community. The fully connected layer we added is randomly initialized, and the encoder part is frozen in the first 10,000 steps.

3.3. Performance of phone recognition

We report the phone error rate (PER) on L2-arctic test set to evaluate the performance of phone recognition. As shown in Table 3, when only considering the default data configuration (marked by -), models based on LV60 (53k-hrs) and XLSR (56k-hrs) obtain better PER than the LARGE (960-hrs) model. The result reveals that models using a bigger amount of data to pre-train can well utilize unsupervised data and generate stronger representation to deal with the L2 data. Note that the performance of XLSR is worse than when using the additional TIMIT data. The mismatch between these two datasets may be an underlying cause. We focus on the task of Mispronunciation Detection and Diagnosis in the following sections.

3.4. Evaluation

We followed the evaluation metrics of previous study [30]. The detection of pronunciation errors can be achieved by comparing the prediction sequence and the reference text sequence. When the recognized phoneme is inconsistent with the reference text, the pronunciation error is detected. For all canonical phones, true accept (TA) and false rejection (FR) indicate the model's ability to distinguish the correct pronunciation. For all mispronounced phones, false accept (FA) and true rejection (TR) indicate the model's ability to distinguish the mispronounced phones. Further, true rejection can be divided into correct diagnosis and diagnosis error. Other metrics like recall (TR/(FA + TR)), precision (TR/(FR + TR)) and the F-1 score (2*((precision*recall)/(precision+recall))) can be calculated based on the accumulated statistics.

3.5. Performance comparison with previous method

Results in Table 2 show that the model based on XLSR surpasses the well-known GOP algorithm and previous work. In CTC-ATT [7], the authors use TIMIT, L2-arctic, and a small portion of the Librispeech corpus to build a three-stage model. Compared with CTC-ATT, XLSR achieves a 4.44% improvement in F1 score (60.44% v.s 56.02%). Even without the use of the TIMIT data, XLSR can still achieve a promising performance (59.37%). This means that the representation from a model pre-trained with a large amount of data is effective and has great potential in pronunciation error detection tasks.

Table 2: *Performance of mispronunciation detection and diagnosis with different approaches.*

Models	PR(%)	RE(%)	F1(%)
GOP[31]	35.42	52.88	42.42
CTC-ATT[7]	46.57	70.28	56.02
CNN-RNN-CTC+VC[32]	56.04	56.12	56.08
w2v2.0-XLSR	63.12	56.05	59.37
w2v2.0-XLSR(+TIMIT)	62.86	58.20	**60.44**

3.6. Performance comparison with different amount data for pre-training

Relative results are listed in Table 3. We can observe that the model LV60 and XLSR achieve better results than LARGE one, which is consistent with the observation in Section 3.3. We summarize these two observations as mispronunciation detection benefits from the general feature representation extracted from large amounts of unlabeled data.

Table 3: *Performance of mispronunciation detection and diagnosis with different pre-trained models and training data sets. The results of phone recognition are included.*

Models	Data	Canonicals		Mispronunciations			F1	PER
		True Accept	False Rejection	False Accept	True Rejection			
					Corroct Diag.	Diag. Error		
w2v2.0-LARGE	-	94.12% (24226)	5.88% (1514)	49.53% (2113)	65.86% (1418)	34.14% (735)	54.28%	16.97%
w2v2.0-LV60	-	94.01% (24198)	5.99% (1542)	43.37% (1850)	68.08% (1645)	31.91% (771)	58.75%	16.01%
w2v2.0-XLSR	-	94.57% (24343)	5.43% (1397)	43.95% (1875)	65.75% (1572)	34.25% (819)	**59.37%**	**15.43%**
w2v2.0-XLSR	-33%	94.11% (24156)	5.89% (1584)	41.23% (1802)	69.13% (1712)	30.87% (752)	59.27%	-
w2v2.0-XLSR	-66%	93.35% (23048)	6.65% (2692)	46.06% (1592)	64.67% (1870)	35.33% (804)	55.52%	-
w2v2.0-XLSR	+TIMIT	94.30% (24273)	5.70% (1467)	41.80% (1783)	70.72% (1756)	29.28% (727)	60.44%	16.20%

3.7. Performance comparison between multilingual pre-training and monolingual pre-training

A related theory, language transfer theory, believes that second language learners will transfer the phonetic phenomenon of their mother tongue to second language learning [33], and some works have shown that the use of cross-language training corpus can help improve the performance of pronunciation error detection systems [34]. Results in Table 3 show that compared with the monolingual pre-training model LV60, the multilingual pre-training model XLSR yields an improvement, but the improvement is minor (58.75% v.s. 59.37%). It seems that the multilingual pre-trained model can transfer cross-language information for pronunciation evaluation, but this claim would have to be proved by disentangling the impact of multilingual training, unrelated languages and simply training on more data.

3.8. Training on few data

We further explore the effectiveness of the pre-trained model on ultra-low resource MDD by decreasing the training L2 data. All the experiments are conducted on the XLSR model. Using only one speaker data for each language (**-66%**), the model gets a 55.52% f1-score, which is close to the previous work using all second language training data. Figure 2 details the pre-phone performance of models trained on **default** and **-66%**, and we can see that their results share similar patterns. The model using less annotated data can retain most of the ability to distinguish phones. This suggests that the feature representations generated from the SSP model can rapidly generalize on MDD tasks even when access to annotated data is limited.

Finally, we use TIMIT to enhance the performance of fine-tune. The results show that additional L1 data can slightly improve the performance of the model, 1% f1-score (59.37% v.s. 60.44%). We conclude that the knowledge about the correct pronunciation from the L1 data can break the bottleneck of using only L2 data.

4. Conclusions

In this work, we explore the application of the pre-training model wav2vec2.0 on the task of mispronunciation detection and diagnosis. By conducting several experiments, we have verified the self-supervised pre-training model can take advantage of unlabeled data and provide useful speech representations for the MDD task. We also explore the feasibility of ultra-low resource MDD. Even when the annotated second language training data is scarce, the model can achieve competitive performance. In future work, we will try to build a more effective system based on SSP model and explore the use of self-supervised pre-training models in children's speech assessment.

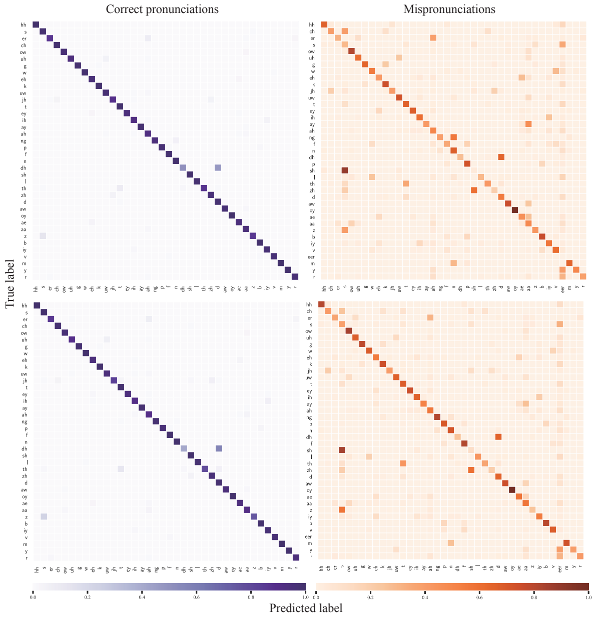

Figure 2: *Confusion matrices of XLSR models trained on* ***default*** *(up) and* ***-66%*** *(down). Note that 'True label' in the mispronunciations side means speakers are expected to pronounce these phones correctly but they failed to do so. The diagonal cells indicate True Accept for the correct pronunciations and False Accept for the mispronunciations while the rest represent False Rejection and True Rejection separately.*

5. Acknowledgements

This study was supported by Advanced Innovation Center for Language Resource and Intelligence (KYR17005), National Social Science Foundation of China (18BYY124), Wutong Innovation Platform of Beijing Language and Culture University (19PT04), the Science Foundation and Special Program for Key Basic Research fund of Beijing Language and Culture University (the Fundamental Research Funds for the Central Universities)(21YJ040004). Jinsong Zhang is the corresponding author.

6. References

[1] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[2] 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.

[3] J. Zheng, C. Huang, M. Chu, F. K. Soong, and W.-p. Ye, "Generalized segment posterior probability for automatic mandarin pronunciation evaluation," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–201.

[4] A. M. Harrison, W.-K. Lo, X.-j. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *International Workshop on Speech and Language Technology in Education*, 2009.

[5] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[6] 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.

[7] B.-C. Yan, M.-C. Wu, H.-T. Hung, and B. Chen, "An End-to-End Mispronunciation Detection System for L2 English Speech Leveraging Novel Anti-Phone Modeling," in *Proc. Interspeech 2020*, 2020, pp. 3032–3036. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1616

[8] 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.

[9] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[10] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?" in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 201–208.

[11] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2051–2060.

[12] 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, 2020.

[13] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. R. Glass, "An unsupervised autoregressive model for speech representation learning," in *INTERSPEECH*, 2019.

[14] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6419–6423.

[15] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *arXiv preprint arXiv:2007.06028*, 2020.

[16] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition." in *INTERSPEECH*, 2019.

[17] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*, 2019.

[18] H. Nguyen, F. Bougares, N. Tomashenko, Y. Estève *et al.*, "Investigating self-supervised pre-training for end-to-end speech translation," 2020.

[19] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," *arXiv preprint arXiv:2012.12121*, 2020.

[20] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[21] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *NIPS*, 2017.

[22] L. Yang, K. Fu, J. Zhang, and T. Shinozaki, "Pronunciation erroneous tendency detection with language adversarial represent learning," *Proc. Interspeech 2020*, pp. 3042–3046, 2020.

[23] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *NAACL-HLT (Demonstrations)*, 2019.

[24] 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.

[25] G. Zhao, S. Sonsaat, A. O. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-arctic: A non-native english speech corpus," *Perception Sensing Instrumentation Lab*, 2018.

[26] SoX, "audio manipulation tool," http://sox.sourceforge.net/, (accessed March 15, 2021).

[27] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[28] 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.

[29] 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.

[30] 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.

[31] B.-C. Yan and B. Chen, "End-to-end mispronunciation detection and diagnosis from raw waveforms," *arXiv preprint arXiv:2103.03023*, 2021.

[32] 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.

[33] C. B. Chang and A. Mishler, "Evidence for language transfer leading to a perceptual advantage for non-native listeners," *The Journal of the Acoustical Society of America*, vol. 132, no. 4, pp. 2700–2710, 2012.

[34] R. Duan, T. Kawahara, M. Dantsuji, and H. Nanjo, "Cross-lingual transfer learning of non-native acoustic modeling for pronunciation error detection and diagnosis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 391–401, 2019.