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

Weakly-supervised word-level pronunciation error detection in non-native English speech

Daniel Korzekwa[1,2], *Jaime Lorenzo-Trueba*[3], *Thomas Drugman*[3], *Shira Calamaro*[3], *Bozena Kostek*[2]

[1]Amazon, Poland
[2]Gdansk University of Technology, Faculty of ETI, Poland
[3]Amazon, UK

korzekwa@amazon.com

Abstract

We propose a weakly-supervised model for word-level mispronunciation detection in non-native (L2) English speech. To train this model, phonetically transcribed L2 speech is not required and we only need to mark mispronounced words. The lack of phonetic transcriptions for L2 speech means that the model has to learn only from a weak signal of word-level mispronunciations. Because of that and due to the limited amount of mispronounced L2 speech, the model is more likely to overfit. To limit this risk, we train it in a multi-task setup. In the first task, we estimate the probabilities of word-level mispronunciation. For the second task, we use a phoneme recognizer trained on phonetically transcribed L1 speech that is easily accessible and can be automatically annotated. Compared to state-of-the-art approaches, we improve the accuracy of detecting word-level pronunciation errors in AUC metric by 30% on the GUT Isle Corpus of L2 Polish speakers, and by 21.5% on the Isle Corpus of L2 German and Italian speakers.

Index Terms: automated pronunciation assessment, speech processing, second-language learning, deep learning

1. Introduction

It has been shown that Computer-Assisted Pronunciation Training (CAPT) helps people practice and improve pronunciation skills [1, 2]. Despite significant progress over the last two decades, standard methods are still unable to detect mispronunciations with high accuracy. These methods can detect phoneme-level mispronunciations at about 60% precision and 40%-80% recall [3, 4, 5]. By further raising precision we can lower the risk of providing incorrect feedback, whereas with higher recall, we can detect more mispronunciation errors.

Standard methods aim at recognizing the phonemes pronounced by a speaker and compare them with expected (canonical) pronunciation of correctly pronounced speech. Any mismatch between recognized and canonical phonemes yields a pronunciation error at the phoneme level. Phoneme recognition-based approaches rely on phonetically transcribed speech labeled by human listeners. Human-based transcription is a laborious task, especially, in the case of L2 speech where listeners have to identify mispronunciations. Sometimes, it might be even impossible to transcribe L2 speech because different languages have different phoneme sets and it is unclear which phonemes were pronounced by the speaker.

Phoneme recognition-based approaches generally fall into two categories. The first category uses forced-alignment techniques [6, 7, 8, 9] based on the work by Franco et al. [10] and the Goodness of Pronunciation (GOP) method [11]. The GOP uses Bayesian inference to find the most likely alignment between canonical phonemes and the corresponding audio signal (forced alignment). Then, the GOP uses the likelihoods of the aligned audio signal as an indicator for mispronounced phonemes. In the second category there are methods that recognize phonemes pronounced by a speaker purely from a speech signal, and only then align them with canonical phonemes [12, 13, 14, 15, 16]. Techniques falling into both categories can be complemented with the use of a reference signal obtained either from a database of speech [17, 18, 19] or generated from phonetic representation [4, 20].

There are two challenges for the phoneme recognition approaches. First, phonemes pronounced by a speaker have to be recognized accurately, which has been shown to be difficult [5, 21, 22, 23]. Second, standard approaches expect only a single canonical pronunciation of a given text, but this assumption does not always hold true due to phonetic variability of speech. In [4], we addressed these problems by incorporating a pronunciation model of L1 speech, but this approach still relies on phonetically transcribed L2 speech.

In this paper, we introduce a novel model (noted as WEAKLY-S) for the detection of word-level pronunciation errors that does not require phonetically transcribed L2 speech. The model produces the probabilities of mispronunciation for all words, conditioned on a spoken sentence and canonical phonemes. Mispronunciation error types include any of phoneme replacement, addition, deletion or unknown speech sound. During training, the model is weakly supervised, in the sense that we only mark mispronounced words in L2 speech and the data do not have to be phonetically transcribed. Due to the limited availability of L2 speech and the fact it is not phonetically transcribed, the model is more likely to overfit. To solve this problem, we train the model in a multi-task setup. In addition to a primary task of word-level mispronunciation detection, we use a phoneme recognizer trained on automatically transcribed L1 speech for the secondary task. Both tasks share common parts of the model, which makes the primary task less likely to overfit. Additionally, we address the overfitting problem with synthetically generated pronunciation errors that are derived from L1 speech.

Leung et al. [3] used a phoneme recognizer based on Connectionist Temporal Classification (CTC) for pronunciation error detection. Instead, we use an attention-based phoneme recognizer following Chorowski et al. [22] so that we can regularize the model by both tasks sharing a common component (attention). With a CTC-based phoneme recognizer it would not be possible because this technique does not use attention that could be shared between both tasks. Zhang et al. [5] employed a multi-task model for pronunciation assessment, but with two important differences. First, they use a Needleman-Wunsch al-

http://dx.doi.org/10.21437/Interspeech.2021-38

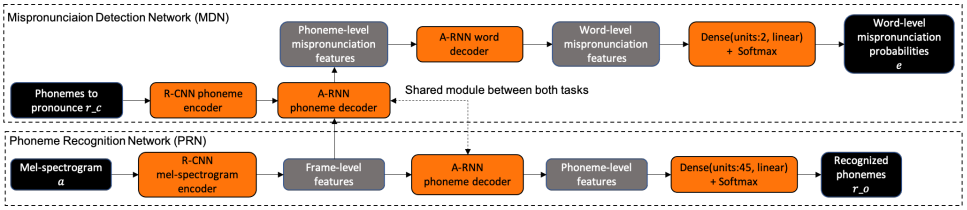

Figure 1: *Neural network architecture of the WEAKLY-S model for word-level pronunciation error detection.*

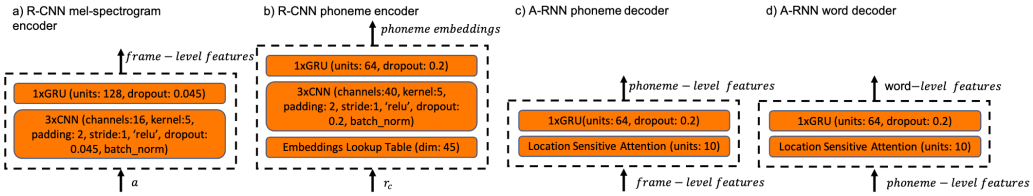

Figure 2: *Details of the neural network architecture of the WEAKLY-S model for word-level pronunciation error detection.*

gorithm [24] for aligning canonical and recognized sequences of phonemes, but this algorithm cannot be tuned towards sequences of phonemes. We use an attention mechanism that automatically maps the speech signal to the sequence of word-level pronunciation errors. Second, Zhang et al. detect pronunciation errors at the phoneme level and they expect L2 speech to be phonetically transcribed. This differs from our method of recognizing pronunciation errors at the word level with no need for phonetic transcriptions of L2 speech. To the best of our knowledge, this is the first approach to train word-level pronunciation error detection model that does not require phonetically transcribed L2 speech and can be optimized directly towards word-level mispronunciation detection.

2. Proposed Model

2.1. Model Definition

The model is made of two sub-networks: *i)* a word-level Mispronunciations Detection Network (MDN) detects word-level pronunciation errors $\mathbf{e}$ from the audio signal $\mathbf{a}$ and canonical phonemes $\mathbf{r_c}$, *ii)* a Phoneme Recognition Network (PRN) recognizes phonemes $\mathbf{r_o}$ pronounced by a speaker from the audio signal $\mathbf{a}$ (Fig. 1).

More formally, let us define the following variables: $\mathbf{a}$ - speech signal represented by a mel-spectrogram, $\mathbf{r_c}$ - canonical phonemes that the speaker was expected to pronounce, $\mathbf{r_o}$ - phonemes pronounced, and $\mathbf{e}$ - the probabilities of mispronouncing words in the spoken sentence. The model outputs the probabilities of word-level mispronunciation, denoted as $\mathbf{e} \sim p(\mathbf{e}|\mathbf{a}, \mathbf{r_c}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represent parameters of the model.

We train the WEAKLY-S model in a multi-task setup. In addition to the primary task $\mathbf{e}$, we use a phoneme recognizer denoted as $\mathbf{r_o} \sim p(\mathbf{r_o}|\mathbf{a}, \boldsymbol{\theta})$ for the secondary task. The parameters $\boldsymbol{\theta}$ are shared between both tasks, which makes the MDN less likely to overfit. We define the loss function as the sum of two losses: a word-level mispronunciation loss and a phoneme recognition loss. Its formulation for the *ith* training example is presented in Eq. 1. We train the model using two types of training data: phonetically transcribed L1 speech (both losses are used) and untranscribed L2 speech (only the mispronunciation loss is used). Having a separate loss for word-level mispronunciation lets us train the model from speech data that are not phonetically transcribed.

$$\mathcal{L}(\boldsymbol{\theta}) = log(p(\mathbf{e}|\mathbf{a}, \mathbf{r_c}, \boldsymbol{\theta})) + log(p(\mathbf{r_o}|\mathbf{a}, \boldsymbol{\theta})) \quad (1)$$

2.2. Neural Network Details

Following Sutskever et al. [25], the MDN network encodes the mel-spectrogram $\mathbf{a}$ and the canonical phonemes $\mathbf{r_c}$ with Recurrent Convolutional Neural Network (RCNN) encoders (Fig. 2a and Fig. 2b). These encoded representations are passed into an attention-based [26] Recurrent Neural Network (A-RNN) decoder (Fig. 2c) that generates phoneme-level mispronunciation features. Phoneme-level features are transformed into word-level features (Fig. 2d) based on an attention mechanism and these finally are used for computing word-level mispronunciation probabilities $\mathbf{e}$.

The PRN recognizes phonemes $\mathbf{r_o}$ pronounced by the speaker. It is similar to the attention-based phoneme recognizer by Chorowski et al. [22]. To generate phoneme-level features, it uses the same RCNN mel-spectrogram encoder and A-RNN decoder as the MDN. The only difference is that the A-RNN decoder is not conditioned on canonical phonemes. Phoneme-level features are transformed to the probabilities of pronounced phonemes. We added a phoneme recognition task due to the limited amount of L2 speech annotated with word-level mispronunciations. Without it, the MDN would be prone to overfitting if it was trained only on its own. By sharing common parts between both models, the PRN acts as a backbone for the MDN and makes it more robust.

The model was implemented in MxNet framework [27] and tuned for hyper-parameters with AutoGluon Bayesian optimization framework [28]. The model was first pretrained on L1 and L2 speech corpora and then the MDN part was fine-tuned only on L2 speech data. We used the Adam optimizer with learning rate 0.001 and gradient clipping 5. Training data were segmented into buckets with batch size 32, using GluonCV [29]. The A-RNN phoneme and word decoders are based on Location Sensitive Attention by Chorowski et al. [22].

3. Experiments

We present three experiments. We start with comparing our model against state-of-the-art approaches in the task of word-

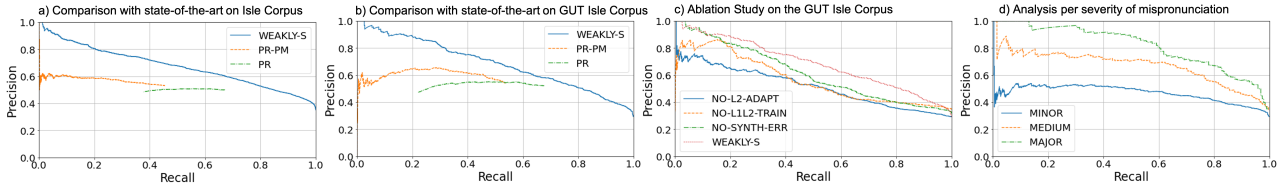

Figure 3: *Precision-recall curves for the WEAKLY-S and baseline models, PR-PM and PR, (a) tested on German and Italian speakers and (b) Polish speakers. (c) Ablation study on the GUT Isle corpus. (d) Analysis of mispronunciation severity levels.*

level mispronunciation detection. In an ablation study we analyze which elements of the model contribute the most to its performance. Finally, we analyze how the severity of pronunciation error affects the accuracy of the model.

3.1. Speech Corpora and Metrics

In our experiments, we use a combination of L1 and L2 English speech. L1 speech is obtained from TIMIT [30] and LibriTTS [31] corpora. L2 data come from the Isle [32] corpus (German and Italian speakers) and the GUT Isle [33] corpus (Polish speakers). In total, we collected 102,812 utterances, summarized in Table 1. We split the data into training and test sets, holding out 28 L2 speakers (11 German, 11 Italian, and 6 Polish) only for testing the performance of the model.

The L2 corpus of Polish speakers was annotated for word-level pronunciation errors by 5 native English speakers. Annotators marked mispronounced words and indicated their severity levels using one of the three possible values: 1 - MINOR, 2 - MEDIUM, 3 - MAJOR. The Isle corpus of German and Italian speakers comes with phoneme level mispronunciations. Words with at least one mispronounced phoneme were automatically marked as mispronounced. The Isle corpus is not mapped to severity levels of mispronunciations. In total, there are 35,555 L2 words, including 8035 mispronounced words. All data were re-sampled to 16 kHz.

We extended the train set with 292,242 utterances of L1 speech with synthetically generated pronunciation errors. We use a simple approach of perturbing phonetic transcription for the corresponding speech audio. First, we sample these utterances with replacement from L1 corpora of human speech. Then, for each utterance, we replace phonemes with random phonemes with a probability of 0.2. In [34] we found that generating incorrectly stressed speech using Text-To-Speech (TTS) improves the accuracy of detecting lexical stress errors in L2 speech. Although, as opposed to using TTS, we create pronunciation errors by perturbing the text, we expect this simpler approach should still help recognizing word-level pronunciation errors.

Table 1: *Summary of speech corpora used in experiments. * - audiobooks read by volunteers from all over the world [31]*

Native Language	Hours	Speakers
English	90.47	640
Unknown*	19.91	285
German and Italian	13.41	46
Polish	1.49	12

To evaluate our model, we use three standard metrics: Area Under Curve (AUC), precision and recall. The AUC metric provides an overall performance of the model accounting for all possible trade offs between precision and recall. Precision-recall plots illustrate relations between both metrics. Complementary, to analyze precision, in all our experiments we consistently fix recall at the value of 0.4 to be comparable with two baseline models that do not cover the whole range of recall values (see Section 3.2).

3.2. Comparison with State-of-the-Art

We compare our proposed WEAKLY-S model against two state-of-the-art baselines. The phoneme recognizer (PR) model by Leung et al. [3] is our first baseline. The PR is based on CTC loss [35] and it outperforms multiple alternative approaches for pronunciation assessment. The original CTC-based model uses a hard likelihood threshold applied to recognized phonemes. To compare it with two other models, following our work in [4], we replaced hard likelihood threshold with a soft threshold. The second baseline is the PR extended by a pronunciation model (PR-PM model [4]). The pronunciation model accounts for phonetic variability of speech produced by native speakers, which results in higher precision of detecting pronunciation errors.

The results are presented in Fig. 3a, Fig. 3b and Table 2. The WEAKLY-S model turns out to outperform the second best model in AUC by 30% from 52.8 to 68.63 and in precision by 23% from 61.21 to 75.25 on the GUT Isle Corpus of Polish speakers. We observe similar improvements on the Isle Corpus of German and Italian speakers.

Table 2: *Accuracy metrics of detecting word-level pronunciation errors. WEAKLY-S vs baseline models.*

Model	AUC [%]	Precision [%,95%CI]	Recall [%,95%CI]
Isle corpus (German and Italian)			
PR	55.52	49.39 (47.59-51.19)	40.20 (38.62-41.81)
PR-PM	48.00	54.20 (52.32-56.08)	40.20 (38.62-41.81)
WEAKLY-S	**67.47**	71.94 (69.96, 73.87)	40.14 (38.56, 41.75)
GUT Isle corpus (Polish)			
PR	52.8	54.91 (50.53-59.24)	40.29 (36.66-44.02)
PR-PM	50.50	61.21 (56.63-65.65)	40.15 (36.51-43.87)
WEAKLY-S	**68.63**	75.25 (71.67-78.59)	40.38 (37.52-43.29)

One difference between our model and the two baselines is that they both use the Needleman-Wunsch algorithm [24] for aligning canonical and recognized sequences of phonemes. This is a dynamic programming-based algorithm for comparing biological sequences and cannot be optimized for mispronunciation errors. Our model automatically finds the mapping between regions in the speech signal and the corresponding canonical phonemes, and then identifies word-level mispronunciation errors. In this way, we eliminate the Needleman-Wunsch algorithm as a possible source of error.

The second difference is the use of phonetic transcriptions for L2 speech. Both baselines use automatic transcriptions provided by an Amazon-proprietary grapheme-to-phoneme model.

In [4] we found that for the PR and PR-PM models it is better to use automatically transcribed L2 speech for training a phoneme recognizer than not use L2 speech at all. Note that these automatic transcriptions will include phoneme mistakes for mispronounced speech. Our model does not use transcriptions of L2 speech, and instead it is guided by the word-level pronunciation errors of L2 speech in a weakly-supervised fashion.

3.3. Ablation Study

We now investigate which elements of our new model contribute the most to its performance. Along with the WEAKLY-S model, we trained three additional variants, each with a certain feature removed. The NO-L2-ADAPT variant does not fine-tune the model on L2 speech, though it is still exposed to L2 speech while it is trained on a combined corpus of L1 and L2 speech. The NO-L1L2-TRAIN model is not trained on L1/L2 speech, and fine-tuning on L2 speech starts from scratch. It means that the model will not use a large amount of phonetically transcribed L1 speech data and ultimately the secondary task of the phoneme recognizer will not be used. In the NO-SYNTH-ERR model, we exclude synthetic samples of mispronounced L1 speech. It significantly reduces the amount of incorrectly pronounced words used during training from 1,129,839 to only 5,273 L2 words.

L2 Fine-tuning (NO-L2-ADAPT) is the most important factor that contributes to the performance of the model (Fig. 3c and Table 3), with an AUC of 51.72% compared to 68.63% for the full model. Training the model on both L2 and L1 speech together is not sufficient. We think it is because L2 speech accounts for less than 1% of the training data and the model naturally leans towards L1 speech. The second most important feature is training the model on a combined set of L1 and L2 speech (NO-L1L2-TRAIN), with AUC of 56.46%. L1 speech accounts for more than 99% of the training data. These data are also phonetically transcribed, and therefore can be used for the phoneme recognition task. The phoneme recognition task acts as a 'backbone' and reduces the effect of overfitting in the main task of detecting word pronunciation errors. Finally, excluding synthetically generated pronunciation errors (NO-SYNTH-ERR) reduces the AUC from 68.63% to 61.54%.

Table 3: *Ablation study for the GUT Isle corpus.*

Model	AUC [%]	Precision [%]	Recall [%]
NO-L2-ADAPT	51.72	57.89	40.11
NO-L1L2-TRAIN	56.46	59.73	40.20
NO-SYNTH-ERR	61.54	67.22	40.38
WEAKLY-S	**68.63**	75.25	40.38

3.4. Severity of Mispronunciation

When providing feedback to the L2 speaker about mispronounced words, we want to reflect the severity of mispronunciation, in order to focus on more severe errors and not report them all at once. We segment pronunciation errors into three categories: LOW, MEDIUM and HIGH, based on an inter-tester agreement of annotating sentences for word-level mispronunciations. Mispronounced words with less than 40% inter-tester agreement belong to the LOW category, between 40% and 80% to MIDDLE, and over 80% to HIGH. We validated that the proposed inter-tester agreement bands are well correlated with explicit listener opinions on the severity of mispronunciation, as shown in Table 4. This result shows that data on mispronunciation severity can be derived automatically, without the need to collect it.

Table 4: *Severity of mispronunciation by inter-tester agreement for the GUT Isle Corpus. 1 - MINOR, 2 - MEDIUM, 3 - MAJOR.*

Inter-tester agreement	Severity [mean and 95% CI]
LOW (Less than 40%)	1.32 (1.28-1.35)
MEDIUM (Between 40% and 80%)	1.58 (1.54-1.62)
HIGH(Higher than 80%)	2.08 (2.03-2.13)

We aim at detecting the words of HIGH inter-tester agreement with higher precision to provide more relevant feedback to L2 speakers. To make AUC, precision, and recall metrics comparable between different levels of inter-tester agreement, we enforce the ratio of mispronounced words across all categories to the same level of 29.2% by randomly down-sampling correctly pronounced words. This value is the proportion of mispronounced words across all inter-tester agreement levels in the GUT Isle Corpus. We observe that we can detect pronunciation errors of HIGH inter-tester agreement with 91.67% precision at 40.38% recall (Fig. 3d and Table 5). By segmenting pronunciation errors into three difference bands, we can report to a language learner only the errors of HIGH inter-tester agreement, and improve their learning experience.

Table 5: *Accuracy metrics for different severity levels of mispronunciation for the GUT Isle Corpus.*

Inter-test agreement	AUC [%]	Precision [%]	Recall [%]
LOW	46.99	51.84	40.48
MEDIUM	66.90	71.89	40.80
HIGH	81.48	91.67	40.31

4. Conclusions and Future Work

We proposed a model for detecting pronunciation errors in English that can be trained from L2 speech labeled only for word-level mispronunciations. The data do not have to be phonetically transcribed. The model outperforms state-of-the-art models in AUC metric on the GUT Isle Corpus of Polish speakers and the Isle Corpus of German and Italian speakers. The limited amount of L2 speech and the lack of phonetically transcribed speech makes this model prone to overfitting. We overcame this issue by proposing a multi-task training with two tasks: a word-level pronunciation error detector trained on L1 and L2 speech, and a phoneme recognizer trained on L1 speech. The most important factors that contribute to the model accuracy are: *i)* fine-tuning on L2 speech, *ii)* pre-training on a joined corpus of L1 and L2 speech, and *iii)* use of synthetically generated pronunciation errors.

The level of inter-tester agreement in annotating pronunciation errors correlates with explicit human opinions about the severity of mispronunciation. By detecting pronunciation errors only for high inter-tester agreement, we may significantly lower the number of false positives reported to a language learner.

In the future, we want to experiment with discrete phoneme representations such as Vector-Quantized Variational-Auto-Encoder (VQ-VAE) [36, 37], which should fit better to discrete nature of phonemes. Second, we plan to generate synthetic mispronounced speech, which is motivated by our recent work on using speech synthesis for generating lexical stress speech errors [34].

5. References

[1] A. Neri, O. Mich, M. Gerosa, and D. Giuliani, "The effectiveness of computer assisted pronunciation training for foreign language learning by children," *Computer Assisted Language Learning*, vol. 21, no. 5, pp. 393–408, 2008.

[2] C. Tejedor-García, D. Escudero, E. Cámara-Arenas, C. González-Ferreras, and V. Cardeñoso-Payo, "Assessing pronunciation improvement in students of english using a controlled computer-assisted pronunciation tool," *IEEE Transactions on Learning Technologies*, 2020.

[3] 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.

[4] D. Korzekwa, J. Lorenzo-Trueba, S. Zaporowski, S. Calamaro, T. Drugman, and B. Kostek, "Mispronunciation detection in non-native (l2) english with uncertainty modeling," *arXiv preprint arXiv:2101.06396, accepted to ICASSP 2021*, 2021.

[5] Z. Zhang, Y. Wang, and J. Yang, "Text-conditioned transformer for automatic pronunciation error detection," *arXiv preprint arXiv:2008.12424*, 2020.

[6] H. Li, S. Huang, S. Wang, and B. Xu, "Context-dependent duration modeling with backoff strategy and look-up tables for pronunciation assessment and mispronunciation detection," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*. ISCA, 2011, pp. 1133–1136.

[7] 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.

[8] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities." in *INTERSPEECH*, 2019, pp. 954–958.

[9] S. Cheng, Z. Liu, L. Li, Z. Tang, D. Wang, and T. F. Zheng, "Asr-free pronunciation assessment," *arXiv preprint arXiv:2005.11902*, 2020.

[10] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *1997 IEEE international conference on acoustics, speech, and signal processing*, vol. 2. IEEE, 1997, pp. 1471–1474.

[11] 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.

[12] N. Minematsu, "Pronunciation assessment based upon the phonological distortions observed in language learners' utterances," in *INTERSPEECH*, 2004.

[13] 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 *Intl. Workshop on Speech and Language Technology in Education*, 2009.

[14] A. Lee and J. R. Glass, "Pronunciation assessment via a comparison-based system," in *SLaTE*, 2013.

[15] P. Plantinga and E. Fosler-Lussier, "Towards real-time mispronunciation detection in kids' speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 690–696.

[16] S. Sudhakara, M. K. Ramanathi, C. Yarra, A. Das, and P. Ghosh, "Noise robust goodness of pronunciation measures using teacher's utterance," in *SLaTE*, 2019.

[17] Y. Xiao, F. K. Soong, and W. Hu, "Paired phone-posteriors approach to esl pronunciation quality assessment," in *bdl*, 2018, vol. 1, no. 782d, p. 3.

[18] M. Nicolao, A. V. Beeston, and T. Hain, "Automatic assessment of english learner pronunciation using discriminative classifiers," in *2015 IEEE Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5351–5355.

[19] J. Wang, Y. Qin, Z. Peng, and T. Lee, "Child speech disorder detection with siamese recurrent network using speech attribute features." in *INTERSPEECH*, 2019, pp. 3885–3889.

[20] X. Qian, H. Meng, and F. Soong, "Capturing l2 segmental mispronunciations with joint-sequence models in computer-aided pronunciation training (capt)," in *2010 7th Intl. Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 84–88.

[21] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," *arXiv preprint arXiv:1412.1602*, 2014.

[22] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[23] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[24] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

[25] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[27] T. e. a. Chen, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv preprint arXiv:1512.01274*, 2015.

[28] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, and A. Smola, "Autogluon-tabular: Robust and accurate automl for structured data," *arXiv preprint arXiv:2003.06505*, 2020.

[29] J. Guo *et al.*, "Gluoncv and gluonnlp: Deep learning in computer vision and natural language processing." *Journal of Machine Learning Research*, vol. 21, no. 23, pp. 1–7, 2020.

[30] 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," *STIN*, vol. 93, p. 27403, 1993.

[31] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Proc. Interspeech 2019*, 2019, pp. 1526–1530.

[32] E. Atwell, P. Howarth, and D. Souter, "The isle corpus: Italian and german spoken learner's english," *ICAME Journal: Intl. Computer Archive of Modern and Medieval English Journal*, vol. 27, pp. 5–18, 2003.

[33] D. Weber, S. Zaporowski, and D. Korzekwa, "Constructing a dataset of speech recordings with lombard effect," in *24th IEEE SPA*, 2020.

[34] D. Korzekwa, B. Kostek *et al.*, "Detection of lexical stress errors in non-native (l2) english with data augmentation and attention," *arXiv preprint arXiv:2012.14788*, 2020.

[35] A. Graves, "Connectionist temporal classification," in *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, 2012, pp. 61–93.

[36] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[37] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6306–6315, 2017.