A UNIVERSAL ORDINAL REGRESSION FOR ASSESSING PHONEME-LEVEL PRONUNCIATION

Shaoguang Mao, Frank Soong, Yan Xia, Jonathan Tien

Microsoft Research Asia
{shamao, frankkps, yanxia, jtien}@microsoft.com

ABSTRACT

The efficacy and robustness of Ordinal Regression (OR) in assessing speech pronunciation for language learning at phrase level has been shown before. However, for assessing phoneme pronunciation, we need to: 1. collect human scoring annotations for phoneme tokens of a short duration (60-70 ms); 2. train an ordinal regression model for each phoneme with the corresponding training and inference costs. In this paper, we propose to train a Universal Ordinal Regression (UOR) model instead of multiple, separate models for different phonemes, and evaluate its performance accordingly. A single universal binary classifier in UOR is trained to make a binary preference decision (better or worse) between a pair of two tokens with the same phoneme ID. In inference, labeled anchored tokens of specific phoneme ID in the training data are paired with test phoneme token to make binary preference decisions. By evaluating the new UOR on Speechocean762, a public speech database for pronunciation evaluation, we show the advantages of the proposed new approach. Improvements of Pearson Correlation Coefficient by 16.7% and Mean Square Error by 25.0%, all relatively, are obtained against the state-of-the-art systems.

Index Terms— Computer Assisted Language Learning, Universal Ordinal Regression, Pronunciation Assessment, Mispronunciation Detection

1. INTRODUCTION

Pronunciation evaluation is an important task in computer assisted language learning (CALL) [1-4]. It brings helpful assessment to language learners for improving their oral practice, particularly, many on mispronunciation detection and diagnosis (MD&D) [1-21].

Nowadays, two main MD&D solutions are scoring-based approaches [5-9] and phoneme recognition-based approaches [10-18]. The former one is to compute a score by evaluating a test token against trained models. One classical method is Goodness of Pronunciation (GOP) [6-9], which is a pronunciation score to measure the acoustic model outputs. Commonly, scores for each phoneme are derived first and then word-level scores or sentence-level scores are obtained with average operations [7].

The phoneme recognition-based methods treat MD&D as a phoneme recognition task [10-18]. Extended Recognition Network (ERN) incorporates common mispronunciation patterns in a recognition network [14, 15]. ERN can detect designed mispronunciations well but not effectively for errors unseen in the training data. Acoustic-Phonemic Model (APM) is to evaluate a test token by utilizing the input of acoustic features and the reference canonical phoneme sequences [13, 16-18]. Different structures are implemented to build APM, including multi-distribution neural network [13], convolution neural network, recurrent neural network plus attention mechanism [18], etc. Even though phoneme recognition-based methods can predict more diagnosis information in substitution, deletion and insertion errors, the performance is still not good enough for large-scale applications [18]. Meanwhile, proficiency evaluation is still very much needed, say for placement purposes.

Recently, some Ordinal Regression (OR) pronunciation assessment models have been proposed for evaluating sentence-level pronunciation and fluency [9, 21]. Different from GOP, OR models take the relative rank ordering information between samples into account to achieve state-of-the-art performance in the corresponding tasks. However, current ordinal regression framework usually provides only an overall score at the sentence level but not for fine-grained, like syllable or phoneme, scores. Hence, sentence level scores are not helpful enough for learners to pinpoint their pronunciation deficiency in a more focused manner.

To take full advantage of ordinal regression for a more precise and focused assessment at the phoneme level, a simple approach is to repeat the ordinal regression training for each individual phoneme. However, some challenges still exist. A large-scale high-quality fine-grained annotations of phoneme pronunciations are necessary. The human effort for labeling large amounts of such annotations is high and the consistency across different human labelers, or even the same labeler, is difficult to guarantee. Besides, if we train the ordinal regression models for all phonemes individually, tens of such models need to be trained. In inference, tens of the trained models need to be hosted in practice.

In this paper, we propose a universal ordinal regression (UOR) approach to phoneme-level assessment. Different from the traditional ordinal regression approach, UOR trains only one universal binary classifier across different

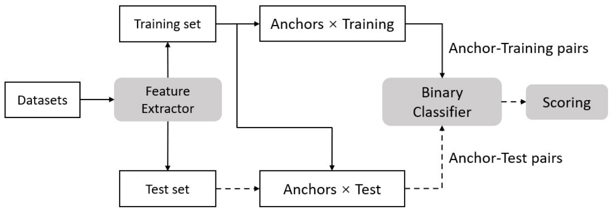

Fig.1. Framework of Ordinal Regression with Anchored Reference Samples.

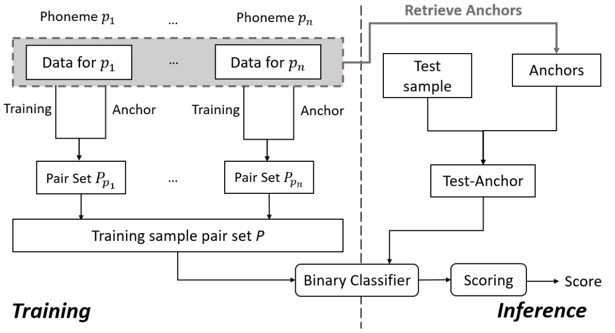

Fig.2. Framework of Universal Ordinal Regression (UOR) for Phoneme Pronunciation Scoring.

phonemes to assess the relative ranking, i.e., better or worse, between any two paired samples from the same phoneme ID. In inference, corresponding anchored reference samples of a specific phoneme are retrieved to compare with the test sample. Compared with GOP or OR-based method, the proposed UOR has the following advantages: 1) increasing the available training data for training the universal model; 2) exploiting the similarity across different phonemes, e.g., the similarity between phonemes “p” and “b”, or other same place-of-articulation phonemes, is more effectively utilized for training a more robust model; 3) reducing the repetitive model training process for more efficient memory storage and streamlined hosting process.

2. RELEATED WORK

2.1. Ordinal Regression Problems

Ordinal Regression (OR) is a regression approach to predicting a test sample’s rank in the dataset where the rank ordering information of all data is given [9, 21, 22]. It has been widely used in practical problems where the natural ranking information among samples can be assessed, like subjective scoring, credit ranking, age estimation, etc.

Specifically, given a dataset D, assume that $(\boldsymbol{x_i}, r_i)$ is the i-th samples in D, where $\boldsymbol{x_i} \in \boldsymbol{X}$ is the input feature and $r_i \in R$ is the corresponding label. The objective of OR is to find a function $r : \boldsymbol{X} \rightarrow R$, where the $r_i \in R$ contains rank ordering information.

Pronunciation scoring can be defined as an OR problem since the different proficiency levels form a natural rank order. For instance, in the 5-point Mean Opinion Score (MOS) scenario, where [1, 2, 3, 4, 5] represent [Bad, Poor, Fair, Good, Excellent], a higher rank of a sample indicates it is better than samples with lower ranks. In previous works, some OR algorithms were proposed to solve prosody or pronunciation scoring.

2.2. Ordinal Regression with Anchored Reference Samples

The Ordinal Regression with Anchored Reference Samples (ORARS) is adopted to address the prosody or pronunciation scoring and achieve state-of-the-art performance. ORARS constructs sample pairs between test sample with all anchored samples and predicts a score based on all comparison results. The rank ordering information is explicitly modeled with a binary classifier which focuses on the relative preference, i.e., better or worse, between samples.

ORARS is constructed with three sub-modules: feature extractor, binary classifier, scoring module [9]. First, a feature vector is extracted for each sample. The extraction process varies with the specific problem. Then, a binary classifier is trained for comparisons between sample pairs. Finally, a scoring module is responsible for predicting a score based on the comparison results between the test sample and anchored reference samples. For example, a scoring method could be predicting the relative order of the test sample in the training set first and then using the score of the corresponding sample as the prediction.

3. UNIVERSAL ORDINAL REGRESSION FOR PHONEME-LEVEL EVALUATION

3.1. Universal Ordinal Regression (UOR)

Different from ORARS, where the binary classifier for predicting preferences between samples are independently trained for each specific task, the Universal Ordinal Regression (UOR) is to train **one** universal model for all binary preference checking tasks. UOR is suitable for the scenario where a series of related tasks are involved, and the phoneme pronunciation evaluation is a typical case.

Specifically, the UOR includes three steps: 1) Task abstraction; 2) Data preparation; 3) Universal model training.

Task Abstraction: To train the evaluator for phoneme pronunciation assessment, $E: (X \rightarrow R)$, traditional ORARS method trains an individual model M_{p_i} for each phoneme p_i, $i \in [1, \mathrm{N}]$,where p_i indicates i-th phoneme. However, the data limitation of labeling data for each phoneme and repeatable training processes makes it hard to implement. Targeting to better utilize the commonality across phonemes, UOR abstracts all sub-tasks into one universal comparison model M.

Data Preparation: To train the universal comparison model M, the training sample pair set P is constructed as the union of all P_{p_i}. $P_{p_i}$is the sample pair set to train the M_{p_i}.

$$P = \bigcup_{i=1}^{N} P_{p_i} \quad (1)$$

Assume that $D_{p_i}: (\boldsymbol{x}^{\boldsymbol{p_i}}, y^{p_i})$ is the training set for task M_{p_i}, P_{p_i} is constructed as $(\boldsymbol{x}_{\boldsymbol{a}}^{\boldsymbol{p_i}}, \boldsymbol{x}_{\boldsymbol{b}}^{\boldsymbol{p_i}}, l_{ab}^{p_i})$, where $\boldsymbol{x}$ is the input feature, y is its score and $(\boldsymbol{x}_{\boldsymbol{a}}^{\boldsymbol{p_i}}, y_a^{p_i})$, $(\boldsymbol{x}_{\boldsymbol{b}}^{\boldsymbol{p_i}}, y_b^{p_i})$ belong to D_{p_i}. The label $l_{ab}^{p_i}$ would be assigned as Eq.2:

$$l_{ab}^{p_i} = \begin{cases} 1, y_a^{p_i} > y_b^{p_i} \\ 0, y_a^{p_i} \leq y_b^{p_i} \end{cases} \quad (2)$$

Universal Model Training: With the training pair sample set *P,* a universal binary model *M* is trained. The output of *M* is $[\, p(y_a^{p_i} > y_b^{p_i}), p(y_a^{p_i} \leq y_b^{p_i}) \,]$ with the input $(\boldsymbol{x}_{\boldsymbol{a}}^{\boldsymbol{p_i}}, \boldsymbol{x}_{\boldsymbol{b}}^{\boldsymbol{p_i}})$, which is activated by a SoftMax function. The loss function for model training is as Eq.3:

$$loss_{ab}^{p_i} = w_{ab}^{p_i} * [-\left(1 - l_{ab}^{p_i}\right) * logp\left(y_a^{p_i} \leq y_b^{p_i}\right) - l_{ab}^{p_i} p\left(y_a^{p_i} > y_b^{p_i}\right)] \quad (3)$$

A weight function $w_{ab}^{p_i}$ is introduced for mitigating the negative influences from human's subjective drafting [9]. $w_{ab}^{p_i}$ will be decreased when two samples' labels are close.

$$w_{ab}^{p_i} = \min\left(\left|y_a^{p_i} - y_b^{p_i}\right|, 1\right) \quad (4)$$

For a test sample $\boldsymbol{x}_{\boldsymbol{test}}^{\boldsymbol{p_i}}$ for p_i, the $D_{p_i}: (\boldsymbol{x}^{\boldsymbol{p_i}}, y^{p_i})$ will be retrieved as the anchored reference sample set, and sample pairs are constructed with the combination of $\boldsymbol{x}_{\boldsymbol{test}}^{\boldsymbol{p_i}}$ and all samples from D_{p_i}. With the predicted posterior probabilities of $p(y_{test}^{p_i} > y_{anchor}^{p_i})$, the relative rank k_{test} of $\boldsymbol{x}_{\boldsymbol{test}}^{\boldsymbol{p_i}}$ in D_{p_i} could be inferred with Eq.5.

$$k_{test} = \left\lfloor \sum_{(\boldsymbol{x}_{\boldsymbol{a}}^{\boldsymbol{p_i}}, y_a^{p_i}) \in D_{p_i}} p\left(y_{test}^{p_i} > y_a^{p_i}\right) \right\rfloor \quad (5)$$

$\lfloor\ \rfloor$ represents a round-down operation. The score of the k_{test}-th sample in $D_{p_i}: (\boldsymbol{x}^{\boldsymbol{p_i}}, y^{p_i})$ will be outputted as the predicted score.

3.2. Discussions

UOR is an evolution of ORARS, which is suitable for the scenarios where a sort of sub-task coexists. It keeps the key idea of ORARS, transferring predicting the absolute score to predicting its relative position in an ordinal space.

A universal model shares the following benefits. On the one hand, the common patterns across tasks may be learned for a more robust model. For instance, to compare the relative pronunciation proficiency for a vowel, the knowledge from the comparisons within other vowels may help. On the other hand, co-training reduces the dependence of the data amount of each phoneme. Imagining the data imbalance fact, that some phonemes are relatively rare, reducing the data requirements for some phonemes is helpful.

Beyond the phoneme-level pronunciation evaluation, there are also plenty of other similar applicable scenarios. For example, in an essay quality evaluation case, a universal binary classifier can be trained across different grades, and in the inference stage, the test sample can be assessed by being paired with samples from the corresponding grade. The wide applicable scenarios highlight the proposed method's values.

Table 1. Pearson correlation coefficient between experts' phoneme-level scoring.

Expert	1	2	3	4	5	Avg.
1		0.54	0.62	0.61	0.60	0.59
2	0.54		0.58	0.52	0.53	0.54
3	0.62	0.58		0.61	0.63	0.61
4	0.61	0.52	0.61		0.61	0.58
5	0.60	0.53	0.63	0.61		0.59
						0.58

From an industry perspective, the proposed UOR saves the cost of repeatable model training and model hosting in the inference stage while achieving better performance.

4. EXPERIMENTS

4.1. Datasets

For the pronunciation evaluation task, two datasets are required. One is a native speaker (L1) corpus for acoustic model (AM) training, and the other one is a non-native (L2) speaker corpus for scoring model training. We reference the data setups from kaldi gop_speechocaen762 [25]. The L1 corpus is *LibriSpeech* [23] and the L2 corpus is *Speechocean762* [24]. All training and test set splitting are the same as preceding works, and the test set is strictly separated during model training stages.

Speechocean762 is a free public dataset for the pronunciation scoring task and includes 5,000 English sentences [24]. All speakers' mother tongue is Mandarin. The scores were annotated by five experts and each expert scores independently under the same metric. There are phoneme-level, word-level, and sentence-level scoring information but only phoneme-level scores are utilized in this work.

For the phoneme-level scoring, there are three ranks: 0 denotes the pronunciation is incorrect or missed, 1 indicates the pronunciation is right but has a heavy accent, and 2 indicates the pronunciation is correct. To better compare model's performance with human subjects, Table 1 lists the Pearson Correlation Coefficient of phoneme pronunciation scoring between human subjects in the test set.

4.2. Feature Selection

Phone-level features are as the input for scoring, which consists of Log Phone Posterior (*LPP*), Log Posterior Ratio (*LPR*) [7, 24].

The Log Phone Posterior (*LPP*) is defined as:

$$LPP(p) = \log \mathrm{p}(p|o; t_s, t_e) \quad (6)$$

The Log Posterior Ratio (*LPR*) between phoneme p_j and p_i is defined as:

$$LPR\left(p_j|p_i\right) = \log \mathrm{p}\left(p_j|o; t_s, t_e\right) - \log \mathrm{p}(p_i|o; t_s, t_e) \quad (7)$$

where o is the observation, i.e., acoustic features, t_s, t_e are start time and end time of p. The $\mathrm{p}(p|o; t_s, t_e)$ is derived from an AM trained with L1, and t_s, t_e are derived from the forced alignment between audio and phoneme sequence.

Table 2. Performance comparison between different models for phoneme-level scoring.

Methods	Mean Square Error	Pearson Correlation	F1-score	
			Macro Avg	Weighted Avg
(I) Goodness of Pronunciation (GOP)	0.84	0.36	/	/
(II) Linear Regression, Separate	0.13	0.38	0.37	0.76
(III) Supported Vector Regression, Separate	0.16	0.45	0.50	0.90
(IV) Neural Net Regression, Separate	0.35	0.21	0.31	0.67
(V) Neural Net Regression, Universal	0.16	0.09	0.34	0.92
(VI) Ordinal Regression, Separate	**0.10**	0.46	0.45	**0.91**
(VII) Our Proposal, UOR	0.12	**0.52**	**0.51**	**0.91**

Additionally, considering the universal model needs to identify which phoneme is assessed now, a one-hot vector to indicate the phoneme ID will be appended to the phone-level features for universal models.

4.3. Metrics

Mean Square Error (MSE), Pearson Correlation Coefficient (PCC), Macro Average of F1-score (macro F1) and Weighted Average of F1-score (weighted F1) are employed as evaluation metrics [24].

Let M be the total number of samples, f_i is the prediction of model for i-th sample and y_i is the corresponding actual value. MSE is computed as Eq. 8.

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(f_i - y_i)^2 \tag{8}$$

PCC is a statistical measure of linear correlation between two data sequences. A higher PCC indicates higher relevance between data sets.

$$PCC(f,y) = \frac{\sum_{i=1}^{M}(f_i-\bar{f})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{M}(f_i-\bar{f})^2}\sqrt{\sum_{i=1}^{M}(y_i-\bar{y})^2}} \tag{9}$$

Macro F1 is the mean of F1 score of all classes and weighted F1 is the weighted sum of F1 score. The previous one measures the average performance across all tasks and the latter favors the performance of major tasks. Specifically, each phoneme evaluation is treated as a task and F1-score is calculated with a ternary classification, i.e., 0, 1, 2. Assuming N is the size of the phoneme set and M_{p_i} is the sample number of p_i, macro F1 and weighted F1 computed as follows.

$$Macro\_F1 = \frac{1}{N}\sum_{i=1}^{N}F1_{p_i} \tag{10}$$

$$Weighted\_F1 = \sum_{i=1}^{N}\frac{M_{p_i}}{M}F1_{p_i} \tag{11}$$

4.4. Experiments settings

Seven models are involved for comparisons: (I) Goodnees of Pronunciation (GOP) [7]; (II) Separate Linear Regression for each phoneme; (III) Separate Supported Vector Regression for each phoneme; (IV) Separate Neural Net Regression for each phoneme; (V) Universal Neural Net Regression for all phonemes; (VI) Separate Ordinal Regression for each phoneme; (VII) Our model, UOR for all phonemes.

The acoustic model used to compute the GOP and extract phone-level features are trained with Kaldi LibriSpeech default receipt [25]. (II) and (III) take GOP and phone-level features as input respectively and train a regressor for each phoneme, it can be reproduced with Kaldi gop_speechocaen 762 receipt. All neural networks in (IV), (V), (VI), (VII) include three hidden layers with 128 hidden units. The loss functions for (IV), (V) are MSE and the loss functions for (VI), (VII) are Eq.3. For (IV) and (VI), the input features are phone-level features (LPP+LPR) while for (V) and (VII), a one-hot vector to specify the phoneme ID is appended.

Model (I) – (V) are set as baselines. Model (VI), (VII) are to verify whether ordinal regression methods can enhance phoneme scoring performance or not. Besides, when contrasting (VI) with (VII), the gains and losses from universal training can be observed.

4.5. Results

The experimental results are listed on Table.2. Neural Net regression performs poorly compared to other approaches. The reason may be that training a neural network based on 2,500 samples is difficult. But the ordinal regression genres are effectively trained thanks to the sample pair construction.

Compared to the SVR, the separate ordinal regression models slightly increase the PCC and MSE but decreased the macro F1. This indicates that OR improves the performance of some majority phonemes while decreasing the performance of some relatively rare phonemes. The lack of data may cause deteriorated performance of some phonemes. However, the UOR increases PCC 16.7% and decreases MSE 25.0% relatively, while keeping F1 score with SVR. The experimental results illustrate the advantages of UOR.

5. CONCLUSION

We proposed a novel Universal Ordinal Regression (UOR) model for objective pronunciation assessment at phoneme level. Different from previous works, the UOR model is trained to assess any phoneme token with one universal binary classifier. The trained universal binary classifier exploits the training data of all phonemes and compensates for relatively inadequate data of some phonemes. Experimental results show that the new UOR, compared with prior state-of-the-art performance, can improve the PCC by 16.7%, and decrease MSE by 25.0%, all relatively. Additionally, the UOR, as a streamlined model with the advantage of more efficient storage and compact processing, is highly desirable for online phoneme pronunciation assessment services.

6. REFERENCES

[1] Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, Patti Price, "Automatic text-independent pronunciation scoring of foreign language student speech.", 1996. ICSLP 96. Proceedings., Fourth International Conference on Vol. 3, pp. 1457-1460, 1996

[2] Horacio Franco, Leonardo Neumeyer, Yoon Kim, Orith Ronen, "Automatic pronunciation scoring for language instruction.", ICASSP 1997., Vol. 2, pp. 1471-1474,

[3] Chul-Ho Jo, Tatsuya Kawahara, Shuji Doshita, and Masatake Dantsuji, "Automatic pronunciation error detection and guidance for foreign language learning.", Fifth International Conference on Spoken Language Processing. 1998.

[4] Shaoguang Mao, Xu Li, Kun Li, Zhiyong Wu, Xunying Liu, and Helen Meng, "Unsupervised discovery of an extended phoneme set in l2 English speech for mispronunciation detection and diagnosis." I in ICASSP, 2018, pp. 6244-6248.

[5] Keelan Evanini and Xinhao Wang, "Automated speech scoring for non-native middle school students with multiple task types.", in INTERSPEECH, 2013, pp. 2435– 2439.

[6] Su-Youn Yoon and Klaus Zechner, "Combining human and automated scores for the improved assessment of non-native speech," Speech Communication, vol. 93, pp. 43–52, 2017.

[7] Wenping Hu, Yao Qian, Frank K Soong, and Yong 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.

[8] Jiatong Shi, Nan Huo, and Qin Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training.", in INTERSPEECH, 2020, pp. 3057– 3061.

[9] Bin Su, Shaoguang Mao, Frank Soong, Yan Xia, Jonathan Tien and Zhiyong Wu, "Improving pronunciation assessment via ordinal regression with anchored reference samples", in ICASSP, 2021, pp. 7748-7752.

[10] K. P. Truong, Ambra Neri, Catia Cucchiarini, and Helmer Strik, "Automatic pronunciation error detection: an acoustic-phonetic approach.", InSTIL/ICALL Symposium 2004, 2004.

[11] Ann Lee, and James R. Glass, "Context-dependent pronunciation error pattern discovery with limited annotations.", INTERSPEECH, 2014

[12] Xiaojun Qian, Helen Meng, and Frank Soong, "A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training.", IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 24.6 (2016), pp.1020-1028, 2016

[13] Kun Li, Xiaojun Qian, and Helen Meng, "Mispronunciation detection and diagnosis in l2 English speech using multi-distribution deep neural networks.", IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.1 (2017), pp.193-207, 2017

[14] Alissa M. Harrison, Wai-Kit Lo, Xiao-jun Qian, and Helen Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training.", SLaTE, 2009.

[15] Wai-Kit Lo, Shuang Zhang, and Helen Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system.", INTERSPEECH, 2010

[16] Shaoguang Mao, Zhiyong Wu, Runnan Li, Xu Li, Helen Meng, and Lianhong Cai, "Applying multitask learning to acoustic-phonemic model for mispronunciation detection and diagnosis in l2 english speech.", ICASSP 2018, pp. 6254-6258.

[17] Yiqing Feng, Guanyu Fu, Qingcai Chen, and Kai Chen, "SED-MDD: Towards sentence dependent end-to-end mispronunciation detection and diagnosis.", ICASSP 2020, pp. 3492-3496. IEEE, 2020.

[18] Kaiqi Fu, Jones Lin, Dengfeng Ke, Yanlu Xie, Jinsong Zhang, and Binghuai Lin, "A Full Text-Dependent End to End Mispronunciation Detection and Diagnosis with Easy Data Augmentation Techniques." arXiv preprint arXiv:2104.08428 (2021).

[19] Bi-Cheng Yan, Shao-Wei Fan Jiang, Fu-An Chao, and Berlin Chen. "Maximum f1-score training for end-to-end mispronunciation detection and diagnosis of L2 English speech." arXiv preprint arXiv:2108.13816 (2021).

[20] Korzekwa, Daniel, Jaime Lorenzo-Trueba, Szymon Zaporowski, Shira Calamaro, Thomas Drugman, and Bozena Kostek. "Mispronunciation Detection in Non-Native (L2) English with Uncertainty Modeling." ICASSP 2021, pp. 7738-7742.

[21] Shaoguang Mao, Zhiyong Wu, Jingshuai Jiang, Peiyun Liu, and Frank K. Soong, "NN-based ordinal regression for assessing fluency of ESL speech." In ICASSP 2019, pp. 7420-7424.

[22] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua, "Ordinal regression with multiple output CNN for age estimation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4920-4928. 2016.

[23] Vassil Panayotov, Guoguo Chen, Daniel Povey, and San jeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in ICASSP, 2015, pp. 5206–5210.

[24] Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. "speechocean762: An Open-Source Non-native English Speech Corpus For Pronunciation Assessment." arXiv preprint arXiv:2104.01378 (2021).

[25] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann et al. "The Kaldi speech recognition toolkit." In IEEE 2011 workshop on automatic speech recognition and understanding.