IMPROVING PRONUNCIATION ASSESSMENT VIA ORDINAL REGRESSION WITH ANCHORED REFERENCE SAMPLES

Bin Su[1,*]*, Shaoguang Mao*[2]*, Frank Soong* [2]*, Yan Xia*[2]*, Jonathan Tien*[2]*, Zhiyong Wu*[1]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies, and Systems,
Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Microsoft Research Asia, Beijing, China

sub18@mails.tsinghua.edu.cn, {shamao, frankkps, yanxia, jtien}@microsoft.com, zywu@se.cuhk.edu.hk

ABSTRACT

Sentence level pronunciation assessment is important for Computer Assisted Language Learning (CALL). Traditional speech pronunciation assessment, based on the Goodness of Pronunciation (GOP) algorithm, has some weakness in assessing a speech utterance: 1) Phoneme GOP scores cannot be easily translated into a sentence score with a simple average for effective assessment; 2) The rank ordering information has not been well exploited in GOP scoring for delivering a robust assessment and correlate well with a human rater's evaluations. In this paper, we propose two new statistical features, average GOP (aGOP) and confusion GOP (cGOP) and use them to train a binary classifier in Ordinal Regression with Anchored Reference Samples (ORARS). When the proposed approach is tested on Microsoft mTutor ESL Dataset, a relative improvement of Pearson correlation coefficient of 26.9% is obtained over the conventional GOP-based one. The performance is at a human-parity level or better than human raters.

Index Terms— Computer Assisted Language Learning, Ordinal Regression, Pronunciation Assessment, Goodness of Pronunciation

1. INTRODUCTION

Sentence level pronunciation assessment is an important task in Computer Assisted Language Learning (CALL), which is commonly required by oral practice and assessment [1, 2].

The Goodness of Pronunciation (GOP) [3] was proposed to assess pronunciation at the phoneme level. Then, the NN-based GOP [4] replaced the GMM-based acoustic model in GOP with a deep neural network and shows a better performance. Besides, considering the change of posterior probability entropy within a phoneme, the Center GOP and transition-aware pronunciation score (TAScore) were proposed [5] and achieved improved performance than GOP.

Most GOP-based algorithms e.g. [3, 4, 5] are designed for phoneme level pronunciation assessment. To get a pro-nunciation score at the sentence level, the simple average of all phoneme scores is adopted [5, 6]. However, there are some flaws in this method: 1) simple average of all phonemes ignores the statistical difference between different phonemes. 2) The range of GOP based method is unbounded and doesn't match well with the range of human opinion scores.

Furthermore, the rank ordering information tends to be overlooked or ignored in the GOP-based algorithms, while the rank ordering information is highly related to pronunciation assessment. For example, in a five point MOS pronunciation assessment, to give a score of three, raters may focus on the intelligibility only, and to give a score of five, raters may demand more advanced pronunciation skills. When comparing two samples, how to decide which one is better is easier. By summarizing these decisions, a more reliable sentence level decision could be obtained. The pairwise sample comparison is easy to perform and a neutral way to use the rank ordering information [7, 8].

Although the pairwise ranking algorithm is suitable for comparing two samples, the objective of pairwise ranking is to minimize the average number of inversions in ranking, which is different from the target of pronunciation assessment. To predict a sample's rank, the pointwise ranking algorithm [8] is suitable. Ordinal Regression (OR) [9, 10] is a pointwise ranking algorithm and has shown its effectiveness for many applications [11, 12, 13].

By combining the ordinal regression with pairwise ranking algorithm, the Ordinal Regression with Anchored Reference Samples (ORARS) is proposed in speech fluency assessment [14]. The ORARS utilizes the rank ordering information to predict the rank directly. By introducing the rank ordering information, the ORARS achieves a better performance than other machine-learning or OR-based [13] methods.

In this paper, we propose a new ORARS-based framework to assess the pronunciation at the sentence level. The proposed method introduces the rank ordering information and sentence level pronunciation features to improve the assessment performance. Compared with the GOP-based method, the proposed method solves two problems: 1) The phoneme

---

level GOP cannot be easily translated to sentence level score with simple average, which is commonly adopted by most researches, 2) The rank ordering information is ignored by the previous methods.

2. RELATED WORK

2.1. Ordinal Regression

Ordinal regression (OR) is a regression algorithm to predict a sample's rank in a set. The rank is marked with a non-negative number usually. Given a dataset $\mathbf{D}$, the i-th samples in $\mathbf{D}$ is $(\boldsymbol{x_i}, r_i)$ where $\boldsymbol{x_i} \in \mathcal{X}$ is the input feature and $r_i \in \mathcal{R}$ is the label of the i-th samples in $\mathbf{D}$. The objective of OR is to find a function $r : \mathcal{X} \rightarrow \mathcal{R}$, which is similar to the regression. However, the $r_i \in \mathcal{R}$ contains the rank ordering information.

In a naïve approach, the ordinal regression task could be solved by regression [15] or classification [16]. Besides, there are a lot of approaches to solve ordinal regression, like ordinal binary decomposition [17, 18]. However, these methods do not utilize the rank ordering information between samples.

2.2. Ordinal Regression with Anchored Reference Samples

The Ordinal Regression with Anchored Reference Samples (ORARS) was proposed [14] to predict the fluency score (rank). The ORARS compares each test sample with all samples in the "Anchor Set" to determine the score. It models the rank ordering information explicitly and get an improvement compared with other OR methods.

Let $\mathbf{D}$ is the training dataset, the scores of all samples in $\mathbf{D}$ are discretized to M ranks firstly. Then $\mathbf{D}$ is separated into two disjoint subsets $\mathbf{D_A}$ and $\mathbf{D_T}$. $\mathbf{D_A}$ is called "Anchor Set", which contains N samples in each rank. $\mathbf{D_T}$ is called training set.

A binary classifier is trained by comparing the samples $(\boldsymbol{x_i^a}, y_i^a) \in \mathbf{D_A}$ and $(\boldsymbol{x_j^t}, y_j^t) \in \mathbf{D_T}$ to output the probability $p_{ij} = f(\boldsymbol{x_i^a}, \boldsymbol{x_j^t})$ of $y_i^a < y_j^t$. The score of a new sample $\boldsymbol{x_t}$ is predicted as Eq.1 where $p_{it} = f(\boldsymbol{x_i^a}, \boldsymbol{x_t})$.

$$y_t = \frac{\sum_{i=1}^{NM} p_{it}}{N} \tag{1}$$

3. PROPOSED FRAMEWORK

The proposed framework is shown as Fig.1. The system includes two modules. The Feature Extraction (FE) module transfers raw acoustic features to fixed dimension feature vectors. The Ordinal Regression (OR) module compares two feature vectors got from FE module to predict their relative rank.
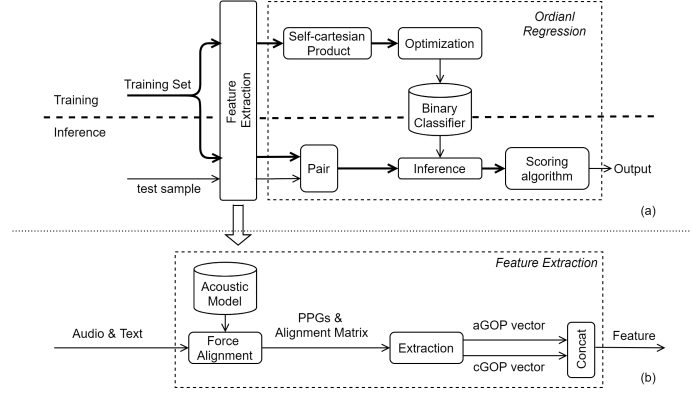

Fig. 1. (a) Framework of Ordinal Regression with Anchored Reference Samples (ORARS) for Speech Pronunciation Scoring. (b) The details of Feature Extraction module.

3.1. Feature Extractor (FE) Module

Average GOP (aGOP) vector and confusion GOP (cGOP) vector are two statistical features derived from the raw acoustic features. To obtain the aGOP and cGOP vectors, Phonetic Posteriorgrams (PPGs) and corresponding alignment matrix are required. The PPGs are computed from the input Mel-frequency Cepstral Coefficient (MFCC) sequences in a sentence. The alignment matrix is calculated from PPGs and text with a pre-trained acoustic model which is trained with transcribed audios.

The aGOP vector contains average GOPs on all phonemes. Let X be the logarithmic PPGs with shape (T, C), and Y is the corresponding alignment matrix with the same shape. T is the time dimension and C is the phoneme dimension. Each time step in Y is a one-hot vector denoted that which phoneme should be correspondingly based on the alignment. The aGOP vector V could be calculated as Eq.2 which keeps the difference between different phonemes. E_x is a matrix whose dimension is $(1, x)$ and all elements are 1. The $\epsilon = 10^{-6}$ is used to avoid the division by zero error. The $\dot{\div}$ indicates element-wise division. The X^T means the transposition of X.

$$V = \frac{E_C X^T Y}{E_T Y + \epsilon} \tag{2}$$

The cGOP vector indicates how pronunciation unit is confused with other phonemes. To calculate cGOP vector:

1. pop the posterior probability of the target phoneme in all time steps from logarithmic PPGs X, the output matrix is denoted as X_c whose shape is $(T, C-1)$;
2. sort the elements in each time step of X_c as descending order, the output is denoted as X_s;
3. calculate the mean M and standard deviation S on X_s along the time axis;
4. cGOP is $[M, S]$, whose shape is $(1, 2 * C - 2)$, where $[\,]$ means concatenate operation.
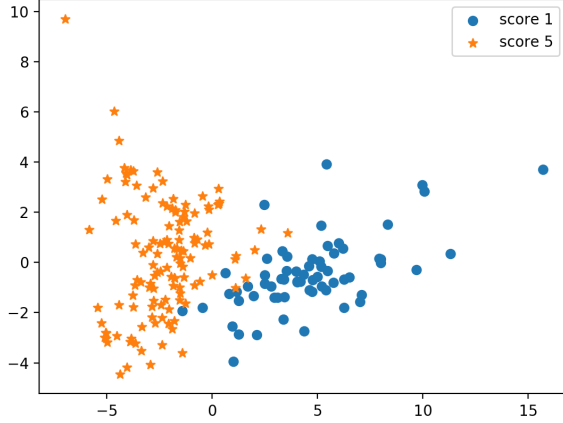
Fig. 2. The scatter plot of the test utterances in their first 2 PCA axes for "score 1" and "score 5" classes in terms of corresponding MOS scores.

Fig.2 shows the Principal Components Analysis (PCA) [19] result of cGOP on "score 1" and "score 5" samples (five-mark). The scatter plot shows that the cGOP vector is quite discriminate on distinct MOS scores, and is an appropriate feature for predicting rank.

3.2. Ordinal Regression (OR) Module

The OR module contains a binary classifier implemented with deep neural network and is used to predict the relative relationship in a given sample pair.

To train the binary classifier, two samples (x_i, y_i) and (x_j, y_j) are selected from the training dataset $\mathbf{D}$. Let f denote the FE module, the input to the binary classifier is $z_{ij} = [f(x_i), f(x_j)]$, the corresponding label l_{ij} could be assigned as Eq.3

$$l_{ij} = \begin{cases} 1 & y_i > y_j \\ 0 & y_i \le y_j \end{cases} \quad (3)$$

The output of binary classifier is $[p_{ij}^0, p_{ij}^1]$ with input z_{ij}, which is activated by softmax function. The loss function could be written as Eq.4. When the training loop is done, the p_{ij}^1 could be regraded as the probability of $y_i > y_j$.

$$loss_{ij} = -(1 - l_{ij}) * logp_{ij}^0 - l_{ij} * logp_{ij}^1 \quad (4)$$

To reduce the subjective drafts from human raters, a weight function is introduced to the loss as Eq.5. When two samples' scores are close, their loss weight will be decreased. The weight function mitigates the risk of incorrect label and makes the training process more stable.

$$w_{ij} = min(|y_i - y_j|, 1) \quad (5)$$

The scoring method proposed in [14] requires a balanced "Anchor Set" which is hard to satisfy. Hence, an alternative method is proposed. The idea is that predict the rank of a sample in a given sample distribution, and then predict its score as the sample's score in corresponding rank.

Specifically, a test sample x_t is compared with all samples $(x_i, y_i) \in \mathbf{D}$ in the training set, and get a series of posterior probabilities $\{p_{ti}^1 | i \in [1, N]\}$ as mentioned above, where N is the size of $\mathbf{D}$. The rank of x_t could be determined by Eq.6

$$k = \sum_{i=1}^{n} p_{ti}^1 + 1 \quad (6)$$

Then the score in $\{y_i | i \in [1, n]\}$ with rank $\lfloor k \rfloor$ is the predicted score, where the $\lfloor \ \rfloor$ means round down.

4. EXPERIMENT

4.1. Datasets

The proposed model has two sub-modules needed to be trained: the acoustic model (AM) in the FE module and the binary classifier in OR module. In our experiment, two pre-trained acoustic models are used to show the robustness of proposed method, and are noted as AM1 and AM2, respectively.

The AM1 is trained with all data in WSJ0 [20], LibriSpeech [21], Common Voice [22], King ASR 249, American Children Words. The AM2 is provided by Microsoft Azure Speech Service, trained on the Microsoft EN* dataset containing more than 100k hours speech. AM2 has better performance than AM1 for it is trained with more data.

The binary classifier is trained with Microsoft mTutor ESL dataset, which contains 2445 utterances, and each utterance is rated by four raters (experienced English teachers) with $\{0, 1, 2, 3, 4, 5\}$. The mean of the four scores is employed as the label. The five-fold cross-validation is used to evaluate the model performance.

4.2. Metrics

The performance of different models are evaluated by Mean Absolute Error (MAE), Pearson Correlation Coefficient (PCC) and Spearman Correlation Coefficient (SCC) [23].

Let $x = [x_1, x_2, \cdots, x_N]$ and $y = [y_1, y_2, \cdots, y_N]$ be two sequences. The MAE and PCC could be calculated as following

- MAE

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |x_i - y_i| \quad (7)$$

- PCC

$$PCC(x, y) = \frac{cov(x, y)}{std(x) * std(y)} \quad (8)$$

where $cov(\boldsymbol{x}, \boldsymbol{y})$ donates the covariance between $\boldsymbol{x}$ and $\boldsymbol{y}$, and $std(\boldsymbol{x})$ is the standard deviation of sequence $\boldsymbol{x}$. The $*$ is the product between two scalars.

The PCC and SCC describe the correlation between the predictions and the ground truths in different perspectives. The higher the PCC or SCC, the better the assessment performance. The MAE describes the error between the predictions and ground truths. A smaller MAE denotes a better performance.

4.3. Experiment settings

The AM1 is a TDNN acoustic model trained with Kaldi's [24] default hyper-parameter settings. The AM2 is provided by Microsoft Azure Speech Service without any modifications.

The binary classifier is a fully connected neural network, which contains three ReLU activated hidden layers with hidden units are 128, 256 and 128, respectively. The output layer has two output units activated by softmax function. The loss function is introduced in section 3.2. The binary classifier is trained with Adam optimizer for 30 epochs with the learning rate set as 10^{-4}, and the batch size set as 1024.

To verify the proposed feature is effective on sentence level pronunciation scoring, the GOP based methods are tested including classic GOP [3] and latest TAScore [5]. To show the ordinal regression is effective, we trained a deep neural network-based regressor (NNR) with the proposed sentence level feature as comparative models.

To train the NNR, all samples' features are extracted by FE module denoted by $\boldsymbol{f}$. Let $(\boldsymbol{x_i}, y_i)$ be the i-th samples in training dataset, and the NNR is trained on $\mathbf{D}_f = \{(\boldsymbol{f}(\boldsymbol{x_i}), y_i)\}$.

The NNR contains three ReLU activated hidden layers with hidden units are 128, 256 and 128, respectively. The output layer has single unit without activation function. The NNR is trained by Adam optimizer for 30 epochs with learning rate set as 10^{-4}, the Mean Square Error (MSE) as shown as Eq.9 is selected as the loss function. The preliminary experiments denote that training with a smaller batch size could improve training speed and performance. Hence, the batch size in the NNR model is set to 4.

$$MSE(x, y) = (x - y)^2 \tag{9}$$

All the experiments are conducted on five-fold cross-validation. In each fold, 10% of training set was used as validation set, and the validation set for all different models are the same. When training the NNR and the proposed model, the model with the smallest loss on the validation set was used in inference stage.

4.4. Results

The results are shown in Table.1. Compared with the GOP-based algorithm, the FE+NNR model obtains at least 17.3%

Table 1. Performance comparison between different assessment algorithms.

Acoustic Model	Algorithm	MAE	PCC	SCC
AM1	GOP	/	0.51	0.48
	TASocre	/	0.52	0.48
	FE+NNR	0.51	0.61	0.56
	Proposed	**0.50**	**0.66**	**0.58**
AM2	GOP	/	0.50	0.48
	TAScore	/	0.51	0.49
	FE+NNR	0.45	0.74	0.67
	Proposed	**0.42**	**0.77**	**0.70**
Human rating*		0.61	0.66	0.63

*Human rating is computer by inter-rater method.

and 16.7% relative improvement in PCC and SCC, respectively. These results indicate that the proposed sentence level feature is much more suitable for sentence level pronunciation assessment than GOP.

Comparing with the FE+NNR model, the proposed model gets at least 2.0%, 4.1% and 3.5% relative performance improvement in MAE, PCC and SCC, respectively. The proposed ORARS-based framework is effective for improving pronunciation assessment over the tested algorithms.

With the sentence level feature vector and the proposed ORARS-based framework, we have improved the performance of sentence level pronunciation assessment. The proposed model, achieves at least 26.9% and 20.8% relative improvement over the GOP-based methods. When a better acoustic model is available, more improvement can be achieved.

Compared with the inter-rater reliability, the proposed method shows even better performance and achieves human-parity.

5. CONCLUSION

In this study, we proposed a new set of feature vectors, aGOP and cGOP, for effective assessment of pronunciation proficiency of a given speech utterance. Additionally, the rank ordering information in the training set, which is pre-assessed (MOS scored) by human listeners for every sentence, is also exploited to train an ordinal regression-based binary classifier via an arrangement of anchored reference samples for more performance improvement. The new set of features and the ordinal regression based binary classification, when tested on the Microsoft mTutor ESL speech database, can improve the assessment performance measured by the Pearson correlation coefficient by 27.2%, relatively, over the traditional GOP-based approach. In comparing with human evaluation, the new assessment achieves a performance of human-parity or better and a smaller mean absolute error.

6. REFERENCES

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

[2] 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.

[3] Silke Maren Witt et al., *Use of speech recognition in computer-assisted language learning*, Ph.D. thesis, University of Cambridge Cambridge, United Kingdom, 1999.

[4] 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.

[5] Jiatong Shi, Nan Huo, and Qin Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," *arXiv preprint arXiv:2008.08647*, 2020.

[6] Kun Li, Xiaojun Qian, and Helen 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.

[7] Yunbo Cao, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, and Hsiao-Wuen Hon, "Adapting ranking svm to document retrieval," in *SIGIR*, 2006, pp. 186–193.

[8] Tie-Yan Liu, *Learning to rank for information retrieval*, Springer Science & Business Media, 2011.

[9] Pedro Antonio Gutiérrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez, "Ordinal regression methods: survey and experimental study," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, 2015.

[10] Koby Crammer and Yoram Singer, "Pranking with ranking," in *NIPS*, 2002, pp. 641–647.

[11] K. Chang, C. Chen, and Y. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *CVPR*, 2011, pp. 585–592.

[12] Kyoung-jae Kim and Hyunchul Ahn, "A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach,"

Computers & Operations Research, vol. 39, no. 8, pp. 1800–1811, 2012.

[13] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua, "Ordinal regression with multiple output cnn for age estimation," in *CVPR*, 2016, pp. 4920–4928.

[14] 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.

[15] Vicenç Torra, Josep Domingo-Ferrer, Josep M Mateo-Sanz, and Michael Ng, "Regression for ordinal variables without underlying continuous variables," *Information Sciences*, vol. 176, no. 4, pp. 465–474, 2006.

[16] Alan Agresti, *Analysis of ordinal categorical data*, vol. 656, John Wiley & Sons, 2010.

[17] Hong Wu, Hanqing Lu, and Songde Ma, "A practical svm-based algorithm for ordinal regression in image retrieval," in *Proceedings of the eleventh ACM international conference on Multimedia*, 2003, pp. 612–621.

[18] Kyoung-jae Kim and Hyunchul Ahn, "A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach," *Computers & Operations Research*, vol. 39, no. 8, pp. 1800–1811, 2012.

[19] Karl Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[20] John S. Garofolo, David Graff, Doug Paul, and David Pallett, "CSR-I (WSJ0) Complete," 2016.

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

[22] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *LREC*, May 2020, pp. 4218–4222.

[23] Charles Spearman, "General intelligenc objectively determined and measured.," 1961.

[24] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *Workshop on ASRU*, Dec. 2011.