NN-BASED ORDINAL REGRESSION FOR ASSESSING FLUENCY OF ESL SPEECH

Shaoguang Mao[1,#] *, Zhiyong Wu*[1]*, Jingshuai Jiang*[2]*, Peiyun Liu*[2]*, Frank K. Soong*[2, *]

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

{msg16@mails,zywu@sz}.tsinghua.edu.cn,{jjsbanana,abbyliu8023}@gmail.com,frankkps@microsoft.com

ABSTRACT

Automatic assessment of a language learner's speech fluency is highly desirable for language education, e.g. for English as a Second Language (ESL) learning. In this paper, we formulate the fluency assessment as a problem of Ordinal Regression with Anchored Reference Samples (ORARS), where the fluency of a speech utterance is predicted by an ordinal regression neural network (NN) trained with anchored reference samples. The ORARS is trained and tested by: picking human expert labeled samples in each mean opinion score (MOS) bucket as the anchored reference samples and pairing them with input speech samples as training couplets; training an NN-based binary classifier to determine which sample in a pair is better in fluency; predicting the rank (MOS) of a test sample based upon the posteriors of all binary comparisons between the test sample and all anchored reference samples. Experimentally, our proposed approach outperforms the traditional NN-based methods and reaches a performance of "human parity", i.e. as comparable as human experts, in its fluency assessment of collected ESL speech. To the best of our knowledge, this is the first attempt to assess speech fluency with an ordinal regression framework where a test input is paired with bucketed and anchored reference samples.

Index Terms— Computer Assisted Language Learning (CALL), speech fluency assessment, ordinal regression, anchored reference sample, mean opinion score (MOS)

1. INTRODUCTION

Fluency, an important attribute of human speech, carries a speaker's intention, naturalness and emotions of speech in oral communication. In Computer Assisted Language Learning (CALL), a capable automatic speech fluency assessment is both necessary and highly desirable [1]-[6].

Fluency can be defined as "the degree to which speech flows easily without pauses and other disfluency markers" [4] and various subjective rating schemes have been proposed to measure speech fluency levels [1], such as the commonly used 5-point mean opinion score (MOS), 1 for bad and 5 for excellent. For automatic assessment, a quantitative and objectively measurable scoring method is needed. Different scoring models have been proposed, such as multi-class classifier [1][2], Gaussian model [3], etc., where fluency levels are assumed to be independent and the intrinsic ordinal (rank ordering) property is ignored. Also, regression approaches were utilized where the bracketed levels are regarded as numerical values [2][5][6]. But the difference between rated levels, e.g. the difference between 4 and 5 and that between 2 and 3, are not equidistant. For example, low levels scores, say 1 and 2, may be more related to unnatural breaks, while high level scores, say 4 and 5, may be more relevant to advanced skills, such as stress and intonation [1][4]. Training non-stationary kernels for regression is challenging and tends to over fit the data in training.

Since speech fluency scoring is a non-stationary process, one important information we can exploit more is the natural order in comparing a pair of speech samples, i.e., better or not. Inspired by research results reported in solving similar problems, e.g. age estimation [7]-[9], credit rating [10]-[11], facial beauty assessment [12] and more [13]-[15], we adopt the ordinal regression approach to speech fluency assessment.

Ordinal regression aims at classifying or predicting numerical values from labelled patterns where the labels of the target variables exhibit a natural ordering [15]. Spoken fluency tests involve rating based on an ordinal scale, i.e. 5-point MOS of [bad, poor, fair, good, excellent], which can be used to build better models.

In collecting subjectively assessed fluency data, we found that for human labelers, it is much easier if they only need to compare two samples and judge which one is better, i.e. a binary preference test of two samples is easier, quicker and more accurate. Based upon this observation, we propose to solve ordinal regression by utilizing anchored reference samples. The rank of a test sample is predicted by a series of "preference selections" between it and the anchored reference samples selected from each rank bucket. Specifically, in the training stage, expert assessed speech samples of each rank are pre-stored in each bracketed bucket. The samples are picked as reference anchors and paired with the remaining samples to form training couplets. A binary classifier is then trained to determine which sample in each paired couplet is better. In inference (testing), the rank of an unseen test sample is then predicted based upon posterior probabilities of all binary comparison results between the test sample and all reference anchors.

The contributions of our work are as follows: (1) formulating the automatic fluency assessment as an ordinal

[#]: Work performed as an intern in Microsoft.

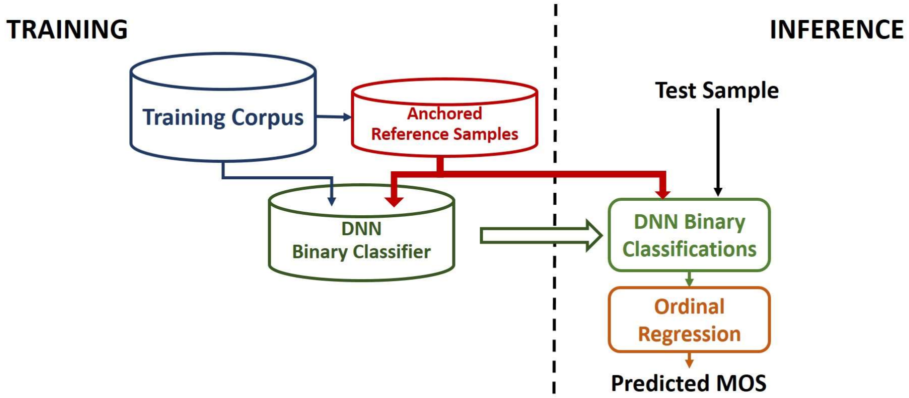

Fig.1. Diagram of the proposed *Ordinal Regression with Anchored Reference Samples (ORARS)* framework

regression problem; (2) proposing to use anchored reference samples to improve ordinal regression performance; (3) demonstrating the new approach achieves a performance close to "human parity" in speech fluency assessment.

2. ORDINAL REGRESSION

2.1 Problem Formulation

Assume that the i-th sample is represented in an input space, $\boldsymbol{x_i} \in X$, and in the output space, $y_i \in Y = \{r_1, r_2, \dots, r_K\}$, with ordered rank, $r_K \succ r_{K-1} \succ \cdots \succ r_1$, where the symbol $\succ$ indicates the ordering relation among different ranks. Given a training set $D = \{\boldsymbol{x_i}, y_i\}_{i=1}^{N_D}$ with N_D samples, ordinal regression is to find a mapping from inputs to ranks, $h(\cdot): X \rightarrow Y$, to minimize the loss function, $L(h)$, with a predefined cost function $C: X \times Y \rightarrow L$.

The *cost matrix* C [17] is adopted to measure the cost between the predicted ranks and the ground-truth ranks. In particular, $C_{y,r}$ is the cost of predicting a sample $(\boldsymbol{x}, y)$ with a rank r. Generally, it is assumed that $C_{y,y} = 0$ and $C_{y,r} > 0$ for $r \neq y$. The absolute cost, defined as $C_{y,r} = |y - r|$, is a common choice for ordinal regression.

2.2 Algorithm Survey

Ordinal regression algorithms can be grouped into two categories [9][15]:

1) Naive approaches with the well-known machine learning algorithms: These algorithms treat ordinal regression problems with specific assumptions. For example, casting different labels into real values and applying standard regression techniques [18]; methods like cost-sensitive classifications [19]; a new support vector machine (SVM) to handle multiple thresholds [20].

2) Ordinal binary decompositions: In this category, ordinal regression problems are decomposed into a series of simpler binary classification sub-problems [9][16][17], which can be solved directly with well-studied classification algorithms. For instance, reduction applied to SVM (REDSVM) was proposed in [17] to reduce an ordinal regression as a set of binary classification problems where several SVMs are used to solve the resultant sub-problems. Recently, multi-task ordinal regression (MTOR) was proposed [9], where an end-to-end convolutional neural network-based, multi-task framework was implemented to train sub-classifications together from input photos for estimating the age of a person.

The ordinal binary decomposition can convert the ordinal regression problem to a set of simpler sub-problems that take ordering into account, it shows a better performance. In this paper, we adopt REDSVM and MTOR as baselines of the ordinal regression for performance comparison.

3. ORDINAL REGRESSION WITH ANCHORED REFERENCE SAMPLES

In this study, we introduce a new ordinal regression framework to assess speech fluency by pairing a given speech sample with anchored reference samples to convert the conventional ordinal regression into a series of simpler "binary preference" tests.

3.1 Anchored Reference Samples

As mentioned earlier, it is easier subjectively by human expert or objectively by machine, to compare two samples and decide which one is better than to give an absolute MOS score of a test sample. We follow the binary preference comparison idea with anchored reference samples.

The subjectively MOS assessed samples are put into the corresponding MOS-labeled buckets. The anchored reference samples can then be picked from the buckets and compared with any given new sample to train a binary classifier. Since the anchored reference samples form a representative sampling of the target sample distribution, by comparing a test sample with them, we can infer the relative position of the test by a sequence of binary classification sub-tests. Hence, we pair samples with anchor samples and to predict results according to contrasts between sample pairs.

3.2 Ordinal Regression with Anchored Reference Samples (ORARS)

In Ordinal Regression with Anchored Reference Samples (ORARS) framework as shown in Fig.1, we focus on determining the relative preference of the two samples in a pair. The model input is a sample pair and the model concentrates on judging which sample is better in fluency. For an unseen sample, its rank is predicted through comparisons between it and all the picked anchored reference samples. Specifically, the proposed ORARS framework consists of **training** and **inference**:

Training:

1. Anchored reference samples selection: given a dataset $X = \{\boldsymbol{x_i}, y_i\}_{i=1}^{N_X}$ with N_X samples, for each rank r_k $(k = 1,..,K)$, we randomly select N_A samples $(\boldsymbol{x}, y)$ with $y = r_k$ to form the anchored reference sample set A. The training set D includes all samples that belong to X but not belong to A, i.e. $D \cup A = X, D \cap A = \emptyset$.

2. Training pairs generation: given the training set $D = \{\boldsymbol{x_j}, y_j\}_{j=1}^{N_D}$ with N_D samples and the anchored reference sample set $A = \{\boldsymbol{x_a^k}, r_k\}_{a=1,k=1}^{N_A,\ K}$, where N_A is the number of picked anchored reference samples of each rank, K is the total number of ranks, a training pair set $P = \{(\boldsymbol{x_j}, \boldsymbol{x_a^k}), z_j^k\}_{j=1,a=1,k=1}^{N_D,\ N_A,\ K}$ is generated based on *Cartesian Product* between D and A, where the first sample $(\boldsymbol{x_j}, y_j) \in D$, the second sample $(\boldsymbol{x_a^k}, r_k) \in A$ and z_j^k is the label for model training representing whether y_j is higher than r_k

$$z_j^k = \begin{cases} 1 & \text{if}(y_j \geq r_k), \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

3. Binary classifier training: a binary classifier is trained with P to determine which sample in a pair is better in fluency. A Deep Neural Network (DNN) binary classifier is trained with the whole training pair set.

Inference:

For a test sample $\boldsymbol{x}'$, to predict its rank y', sample pairs are similarly formed by retrieving anchored reference samples from the set A. After contrasts between $\boldsymbol{x}'$ and all the samples in A, the predicted rank is computed as follows,

$$\mathrm{h}(x') = \frac{\sum_{k=1}^{K}\sum_{a=1}^{N_A} F(\boldsymbol{x}', \boldsymbol{x_a^k})}{N_A} \qquad (2)$$

$F(\boldsymbol{x}', \boldsymbol{x_a^k})$ is the binary classification output by comparing the test sample $\boldsymbol{x}'$ with the anchored reference sample $(\boldsymbol{x_a^k}, r_k)$ via the trained model. $F(\boldsymbol{x}', \boldsymbol{x_a^k})$ can be implemented with a zero or one hard binary decision, or with a soft decision, e.g. posteriors. The cumulative quantization errors of hard binary decisions may result in a final performance degradation, which can be alleviated by a soft decision decoding [21][22] with the posteriors estimated by the DNN binary classifier. The confidence of whether a test sample is better than an anchored reference sample in fluency is estimated naturally as a posterior by the *softmax* function in the last layer of DNN. Thus, the $F(\boldsymbol{x}', \boldsymbol{x_a^k})$ is defined as:

$$F(\boldsymbol{x}', \boldsymbol{x_a^k}) = \mathrm{P}(y' \geq r_k | \boldsymbol{x}', \boldsymbol{x_a^k}) \quad (3)$$

3.3 Approach Analysis

Compared with the traditional machine learning and ordinal regression approach, the ORARS judges whether the rank of a sample is higher than rank k by comparing the sample with anchored reference samples in rank k.

The advantages of ORARS are two-fold: (1) it simplifies the traditional multi-class classification or regression to a simple binary classifier for "preference selection" that is an easier task to train the model reliability; on the other hand, the combinations of data to formulate training pairs greatly enlarge the training data size and reduce data imbalance; (2) it introduces anchored reference samples to improve the performance. When the anchored reference samples, which form a sampling of the rank space, are compared with the test, a more accurate ordinal regression process is established.

4. CORPUS

The Chinese Learners' English Prosody Corpus (CLEPC) are utilized for this study. The corpus is collected by Microsoft Research Asia mTutor team for evaluating fluency of ESL speech of Chinese learners.

All speech utterances are spoken by Chinese ESL learners in a "read after me" practice, where example sentences spoken by native speakers are available for listening before repeating them. Labelers with acoustic-phonetic background are recruited to assess the recorded sentences in 5-point MOS scores. The fluency of each speech utterance was rated with a score in [1, 2, 3, 4, 5] for [Bad, Poor, Fair, Good, Excellent], respectively. Each recorded sentence was manually labeled by two labelers. If the scores of the same sentence by the two labelers are larger than 1, an additional 3rd score is obtained and the closest 2 among the 3 scores are used.

Altogether, speech samples recorded from 182 speakers, and 8000 samples are collected. The Correlation coefficient (Corr), Mean absolute error (MAE), Cumulative score (CS), Fine error, Coarse error are used to measure the labeling consistency between the two labelers:

$$Corr = \frac{Cov(S_1, S_2)}{\sqrt{Var[S_1]Var[S_2]}} \qquad (4)$$

$$\mathrm{MAE} = \frac{\sum_{n=1}^{N}|s_{1n} - s_{2n}|}{N} \qquad (5)$$

$$CS(s) = \frac{N_s}{N} \times 100\% \qquad (6)$$

$$Fine\_error = CS(0.5) \qquad (7)$$

$$Coarse\_error = 1 - CS(1.5) \qquad (8)$$

where S_1, S_2 are scoring sequences from two labelers, $s_{in} \in S_i$ is the score of n-th speech sample from i-th labeler, N is the set size, $Cov()$ is covariance of two variables, $Var[]$ is varance of a set, and N_s is the number of speech samples whose absolute error between two rating scales is less than s.

The labeling quality of human labelers is shown in the Table 1. Integrating two labelers' opinions, we use the average score from two labelers as the final label for each

Table 1. Labeling quality of human labelers

Metrics	Corr	MAE	Fine err	Coarse err
Human labeler	0.612	0.602	48.3%	7.25%

Table 2. Detailed information about designed Fluency Feature Vector

Category	Feature	Correlation	Extraction
Break	Break similarity with reference including position and duration	0.257	Ref.[23]
	Break position similarity with reference	0.256	Ref.[23]
	The percentage of break duration	-0.523	Break duration / whole duration
Rate	Syllable per second (including break time)	0.387	Syllable number / whole duration
	Syllable per second (ignoring break time)	0.543	Syllable number / non-silence duration
Pronunciation	Average pronunciation quality of all phonemes	0.146	mTutor pronunciation scoring API [24]

sample. Hence, there are nine ranks [1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5] and ▁▂▇▅▂ is the histogram of different ranks.

5. EXPERIMENTS

5.1. Experimental Setup

500 speech samples from CLEPC corpus are randomly picked as the test set D'. For each experiment, N_A speech samples with consistent labeling (i.e. $S_{1n} = S_{2n}$) in each MOS score (1 to 5) bucket are randomly selected as the anchored reference sample set A, and all remaining samples are employed as the training set D.

Six systems are tested in our experiments: 1) DNN classifier; 2) SVM classifier; 3) linear SVR; 4) reduction applied to SVM (REDSVM) [17]; 5) multi-task ordinal regression (MTOR) [9]; 6) ordinal regression with anchored reference samples (ORARS). Experiments on systems 1 to 5 are conducted first to verify whether ordinal regression algorithms will be a better choice to solve the fluency assessment problem. Then, systems 4 to 6 are evaluated to confirm if the proposed ORARS framework can improve the performance further.

After preliminary experiments on trying different NN configurations, all NN models are trained with six hidden layers with 512 units per layer with *tanh* as the activation function, and a *softmax* output layer.

To verify ORARS is robust to the number and selection of the anchor points, N_A value in systems 6 varies from 10 to 50 in an increment of 5, and for each N_A setting, six random selection experiments are conducted.

Corr, MAE, Fine error, Coarse error between predicted scores and ground truth labels are adopted as the experimental quality metrics (Eqs. 4, 5, 7, and 8). To compare our model with human assessment capability, the labelers' performance on the test set is also measured.

5.2 Feature Selection

We take the subjective factors that human use to evaluate the speech fluency into our feature selection to form a six-dimension feature vector "Fluency Feature Vector (FFV)" in Break, Speech Rate and Pronunciation quality [2][23]. Detailed information about FFV is listed in Table 2. The preliminary experimental results indicate that FFV shows better than the conventional raw features (MFCC or Mel-spectrogram) and they are employed in all experiments.

Table 3. Experimental results on test set

Method	Corr	MAE	Fine Err (%)	Coarse Err (%)
Human labelers	0.652	0.554	47.8	3.2
DNN Classifier	0.540	0.550	31.8	11.2
SVM Classifier	0.503	0.583	28.2	12.4
SVR	0.614	0.545	52.4	6.0
REDSVM	0.583	0.539	30.4	9.4
MTOR	0.590	**0.524**	31.6	9.2
ORARS	**0.640** **(±0.002)**	0.533 (±0.005)	**56.0** **(±0.6)**	**3.5** **(±0.3)**

5.3 Experimental Results

By conducting a series of random experiments for ORARS, we obtain the means and standard deviations of random anchor selection experiments as listed in Table 3, which shows the performance of ORARS predicted MOS is highly stable and distributes in a small range. The results show that the proposed methods are robust with respect to random selections of the anchors. ORARS slightly outperforms human labeling performance with a lower MAE and a smaller fine error and are close to human labeling in other measures.

Hence, we can draw the following conclusions: treating speech fluency assessment as an ordinal regression problem is both viable and desirable; ORARS framework outperforms traditional ordinal regression in assessing speech fluency of ESL utterances and is robust to random anchor selections.

6. CONCLUSION

We propose an ordinal regression approach to automatic assessment of ESL speech fluency. The MOS scoring process is decomposed into a sequence of NN-based binary classifications by pairing speech test sample with human MOS-scored anchored references. The paired samples are compared to make a series of “better or not” decisions. The final rank (MOS score) of a test sample is predicted by the output posteriors of all paired binary comparisons. The new approach achieves a performance better than other ordinal regression algorithms. It reaches a performance of “human parity” level by outperforming slightly human labeling performance with a lower MAE and a smaller fine error measure. The correlation with human labeling and a measure of coarse error, are comparable or close, respectively. The excellent performance of the new approach indicates its high potential for other subjective scoring applications.

7. REFERENCES

[1] Chung, H., Lee, Y. K., & Park, J. G., Ground truth estimation of spoken English fluency score using decorrelation penalized low-rank matrix factorization. [in] *Proc. ASRU*, pp. 445-449, 2017.

[2] Zechner, K., Higgins, D., Xi, X., &Williamson, D. M., Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, vol. 51, no. 10, pp. 883-895, 2009.

[3] Knill, K. M., & Gales, M. J., Automatically grading learners' English using a Gaussian process. [in] *Proc. ISCA*, 2015.

[4] Derwing, T. M., & Munro, M. J., *Pronunciation fundamentals: evidence-based perspective for L2 teaching and research*. vol. 42. John Benjamins Publishing Company, 2015.

[5] Fontan, L., Le Coz, M., & Detey, S., Automatically measuring L2 speech fluency without the need of ASR: a proof-of-concept study with Japanese learners of French. [in] *Proc. Interspeech 2018*, 2018.

[6] Hassanali, K. N., Yoon, S. Y., & Chen, L., Automatic scoring of non-native children's spoken language proficiency. [in] *Proc. SLaTE*, pp. 13-18, 2015.

[7] Chang, K. Y., Chen, C. S., & Hung, Y. P., Ordinal hyperplanes ranker with cost sensitivities for age estimation. [in] *Proc. CVPR*, pp. 585-592, 2011.

[8] Cao, D., Lei, Z., Zhang, Z., Feng, J., & Li, S. Z., Human age estimation using ranking svm. [in] *Proc. Chinese Conference on Biometric Recognition*, pp. 324-331, 2012.

[9] Niu, Z., Zhou, M., Wang, L., Gao, X., & Hua, G., Ordinal regression with multiple output cnn for age estimation. [in] *Proc. CVPR*, pp. 4920-4928, 2016.

[10] Dikkers, H., & Rothkrantz, L., Support vector machines in ordinal classification: An application to corporate credit scoring. *Neural Network World*, vol.15, no.6, 491, 2005

[11] Kim, K. J., & Ahn, H., 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.

[12] Yan, H., Cost-sensitive ordinal regression for fully automatic facial beauty assessment. *Neurocomputing*, vol. 129, pp. 334-342, 2014.

[13] Bender, R., & Grouven, U., Ordinal logistic regression in medical research. *Journal of the Royal College of physicians of London*, vol. 31, no. 5, pp. 546-551. 1997.

[14] Rudovic, O., Pavlovic, V., & Pantic, M., Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. [in] *Proc. CVPR*, pp.2634-2641, 2012.

[15] Gutiérrez, P. A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F., & Hervas-Martinez, C., Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, vol.28, no.1, pp.127-146, 2016.

[16] Frank, E., & Hall, M., A simple approach to ordinal classification. [in] *Proc. European Conference on Machine Learning*, pp. 145-156, 2001.

[17] Li, L., & Lin, H. T., Ordinal regression by extended binary classification. [in] *Proc. Advances in neural information processing systems*, pp. 865-872, 2007.

[18] Torra, V., Domingo-Ferrer, J., Mateo-Sanz, J. M., & Ng, M., Regression for ordinal variables without underlying continuous variables. *Information Sciences*, vol.176, no.4, pp.465-474, 2006.

[19] Tu, H. H., & Lin, H. T., One-sided Support Vector Regression for Multiclass Cost-sensitive Classification. [in] *Proc. ICML*, pp. 1095-1102, 2010.

[20] Shashua, A., & Levin, A., Ranking with large margin principle: Two approaches. [in] *Proc. Advances in neural information processing systems*, pp. 961-968, 2003.

[21] Hartmann, C. R., *Soft Decision Decoding. In Algebraic Coding Theory and Applications*, Springer, Berlin, Heidelberg. pp. 333-365, 1979.

[22] Hagenauer, J., & Hoeher, P., A Viterbi algorithm with soft-decision outputs and its applications. [in] *Proc. GLOBECOM*, pp. 1680-1686, 1989.

[23] Xiao, Y., & Soong, F. K., Proficiency Assessment of ESL Learner's Sentence Prosody with TTS Synthesized Voice as Reference. [in] *Proc. Interspeech 2017*, pp. 1755-1759, 2017.

[24] Hu, W., Qian, Y., Soong, F. K., & Wang, Y., 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.