8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)
20-21 September 2019, Graz, Austria

# Corpora Design and Score Calibration for Text Dependent Pronunciation Proficiency Recognition

Fred Richardson, John Steinberg, Gordon Vidaver, Steve Feinstein, Ray Budd, Jennifer Melot, Paul Gatewood and Douglas Jones

MIT Lincoln Laboratory

{frichard,John.Steinberg,Gordon.Vidaver,srf,Raymond.Budd,
Jennifer.Melot,Paul.Gatewood,daj}@ll.mit.edu

## Abstract

This work investigates methods for improving a pronunciation proficiency recognition system, both in terms of phonetic level posterior probability calibration, and in ordinal utterance level classification, for Modern Standard Arabic (MSA), Spanish and Russian. To support this work, utterance level labels were obtained by crowd-sourcing the annotation of language learners' recordings. Phonetic posterior probability estimates extracted using automatic speech recognition systems trained in each language were estimated using a beta calibration approach [1] and language proficiency level was estimated using an ordinal regression [2]. Fusion with language recognition (LR) scores from an i-vector system [3] trained on 23 languages is also explored. Initial results were promising for all three languages and it was demonstrated that the calibrated posteriors were effective for predicting pronunciation proficiency. Significant relative gains of 16% mean absolute error for the ordinal regression and 17% normalized cross entropy for the binary beta regression were achieved on MSA through fusion with LR scores.

## 1. Introduction

Automatic speech recognition (ASR) systems have been used for pronunciation assessment for some time [4]. However, some have claimed that the confidence scores derived from ASR are difficult for humans to interpret, which can limit their effectiveness in language learning applications [5]. This is likely due in part to the nature of ASR systems which traditionally attempt to solve a different problem, i.e. recognize words or phones despite variations in speaker, channel, accent, etc. Since these systems are typically trained exclusively with native speech and are not exposed to data from language learners, scores derived from ASR systems are not guaranteed to correlate well with pronunciation proficiency. This paper focuses on improving the utility of ASR scores for language learning by leveraging utterance level pronunciation labels from native speakers and language learners to (1) improve interpretability of phone level scores through a beta regression calibration and (2) predict a meaningful and ranked overall pronunciation level using an ordinal regression trained on ASR and language recognition (LR) features.

Goodness of pronunciation (GOP), a very commonly used feature for pronunciation proficiency recognition (PPR) [6, 7, 8], is essentially a phoneme segment level posterior computed using frame-level likelihoods or posteriors from an ASR system. While much prior work has used language learner data labeled at the phonetic level [5, 7, 8, 9, 10, 11], that annotation is extremely difficult to obtain as it requires expert knowledge, often suffers from significant disagreement between annotators, and takes a considerable amount of time and effort to complete. Instead, we obtain utterance level labels for language learners' recordings through a crowd-sourced annotation framework. Calibrated phoneme posteriors for these samples are inferred using a beta calibration technique [1] and utterance-level ranked ordinal PPR levels are predicted using a threshold-based ordinal regression [2]. In order to compensate for the lack of non-native speech used to train the ASR system, the GOP features are fused with scores from an LR system [3] trained on 23 languages. Experiments in modern standard Arabic (MSA), Spanish, and Russian demonstrate the utility of phoneme level beta calibration and LR fusion where average calibrated phone posteriors are shown to predict pronunciation proficiency almost as well as utterance level systems and fusion with a LR system is shown to significantly improve overall PPR performance.

## 2. Data Sets

The data used in this work consists of known read text prompts from multiple target languages where the speakers are either native speakers of the language or language learning students. It is expected that most of the students' first language (L1) is English though that information was not provided. There are up to three types of data used for a given language: native speaker data collected from native speakers of the language, data produced by language teachers (also presumed to be native speakers) and data produced by language learners whose proficiency in the target language is initially unknown. A crowd-sourced annotation framework is used to label the student data as having been spoken by a beginner, advanced, or native speaker.

### 2.1. Native Speaker Data

The native speaker data used in this work covers 22 languages including MSA, Spanish, and Russian. Most languages have roughly 20 hours of data and 60 to 70 speakers more or less divided equally between males and females. Each speaker produced between 200 and 300 recordings of 2-3 second duration each for up to two different channel conditions: (1) a clean "laptop" channel in which recordings were made with a high quality microphone in a quiet environment and (2) a slightly noisier recording made on an iOS device (iPhone or iPad). MSA, Spanish, and Russian were used to train the Kaldi ASR systems used

This work is sponsored by the Defense Language Institute Foreign Language Center under Air Force Contract FA8702-15-D-0001. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.



10.21437/SLaTE.2019-12

in this work.

### 2.2. Annotated Language Learner Data

A crowd-sourcing platform was used to annotate language learner recordings for pronunciation proficiency. The annotators, who were native speakers of the target language, were asked to rank recordings as "beginner", "advanced" or "native" (ordinal labels 1, 2 or 3 respectively). Each recording was labeled independently by 5 annotators and any data with less than 80% agreement was discarded. Additional teacher data was added to increase the number of samples for the "native" class by sampling roughly 10 recordings for each teacher. A binary task, used for the beta regression described in 5.1, was also created by assigning the "$\mathrm{nat}$" class label to "native" data and the "$\overline{\mathrm{nat}}$" class label to "beginner" data while discarding the "advanced" data. The resulting data sets have between 2,500 and 3,500 samples across more than 200 speakers with a majority of the data falling in the "advanced" class. A summary of the language learner and teacher data is shown in Table 1 for Spanish, MSA, and Russian.

## 3. System Architecture

The following is an overview of the system architecture used for pronunciation proficiency recognition in this work. The input to the system is an acoustic waveform and a transcript of $N$ words $\mathbf{W} = [w_1, ..., w_N]$ where each word token $w_i \in \{1, ..., L_w\}$. Mel frequency cepstral coefficients (MFCCs) serve as feature vectors $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_T]$ and are extracted from the waveform. MFCCs are used for two purposes:

1. To align the transcript $\mathbf{W}$ to produce a sequence of phonemes $\mathbf{Q} = [q_1, ..., q_K]$ where $q_i \in \{1, ..., L_q\}$ and their corresponding start and end times $\mathbf{S} = [\mathbf{s}_1, ..., \mathbf{s}_K]$ where $\mathbf{s}_i = [t_{\mathrm{start}}, t_{\mathrm{end}}]$ and
2. To compute phoneme posteriors $p(q|t, \mathbf{X})$ for all phonemes $q$ at each time index $t$.

From the forced alignment and frame posteriors, average phone posteriors $p(q_i|s_i, \mathbf{X})$ (GOP features) are estimated for each phone $q_i$ and corresponding segment $\mathbf{s}_i$. The average phone segment posteriors are then used in the calibration stage either to estimate the binary posterior probability that the utterance was produced by a native speaker $p(\mathrm{nat}|\mathbf{W}, \mathbf{X})$ or to estimate a speaker's pronunciation proficiency level for the utterance using a multi-class ordinal decision function $f(\mathbf{W}, \mathbf{X})$ which returns the predicted ordinal class label 1, 2 or 3 for $\mathbf{W}$ and $\mathbf{X}$. Both $p(\mathrm{nat}|\mathbf{W}, \mathbf{X})$ and $f(\mathbf{W}, \mathbf{X})$ will be described in more detail in Section 5.

### 3.1. Speech Recognition System

The feature extraction, frame posterior estimation and forced alignment stages are all components from the `Kaldi`[1] speech recognition system [12]. An existing Kaldi Switchboard multi-condition time delay neural network (TDNN) chain model recipe [13] was modified to use four hidden TDNN layers and to use wide band waveform data. The data augmentation stage of the recipe was adjusted to use higher signal to noise ratios, to apply point source noises more frequently and to draw from both simulated and real room impulse responses and background noises. The input features to the TDNN chain model are LDA transformed MFCCs and i-vectors.

[1] https://kaldi-asr.org/

### 3.2. Language Recognition System

An i-vector based LR system [3] was trained using data from 23 languages including all 22 foreign languages and WSJ English [14]. The LR system uses PLDA scoring [15] at the utterance level to estimate the log likelihood ratio $\ln p(\mathbf{X}|\mathrm{lang})/p(\mathbf{X}|\overline{\mathrm{lang}})$ where $\mathrm{lang}$ is one of the 23 target languages. The PLDA scores are fused to the ASR posterior features for both the beta and ordinal regressions. For details, refer to the baseline i-vector system described in [16].

## 4. Tasks and Performance Metrics

The annotated language learner and teacher data set described in Section 2.2 is used to define two PPR tasks - a binary task and an ordinal multi-class task. Each task has different performance metrics. In the binary case, the error rate ($\mathrm{ERR}$) can be expressed in terms of the false alarm rate ($\mathrm{FA}(t)$) and the miss rate ($\mathrm{MR}(t)$) for a decision threshold of $0.5$:

$$\mathrm{ERR} = \frac{1}{2}\left(\mathrm{MR}(0.5) + \mathrm{FA}(0.5)\right)$$

$$\mathrm{MR}(t) = \frac{1}{N_{\mathrm{nat}}} \sum_{\mathbf{W},\mathbf{X}\in\mathrm{nat}} H\left(t - p(\mathrm{nat}|\mathbf{W}, \mathbf{X})\right)$$

$$\mathrm{FA}(t) = \frac{1}{N_{\overline{\mathrm{nat}}}} \sum_{\mathbf{W},\mathbf{X}\in\overline{\mathrm{nat}}} H\left(p(\mathrm{nat}|\mathbf{W}, \mathbf{X}) - t\right)$$

where $N_{\mathrm{nat}}$ is the number of samples labeled as "native", $N_{\overline{\mathrm{nat}}}$ is the number of "not native" samples, and $H(x)$ is the Heaviside step function (which is 1 if $x$ is a positive value and 0 otherwise). The binary normalized cross entropy ($\mathrm{NCE}$) is used to measure calibration performance and can be expressed using the utterance level binary posterior $p(\mathrm{nat}|\mathbf{W}, \mathbf{X})$ as:

$$\mathrm{NCE} = -\frac{1}{2}\left(\frac{1}{N_{\mathrm{nat}}} \sum_{\mathbf{W},\mathbf{X}\in\mathrm{nat}} \log_2 p(\mathrm{nat}|\mathbf{W}, \mathbf{X}) + \frac{1}{N_{\overline{\mathrm{nat}}}} \sum_{\mathbf{W},\mathbf{X}\in\overline{\mathrm{nat}}} \log_2\left(1 - p(\mathrm{nat}|\mathbf{W}, \mathbf{X})\right)\right)$$

Both the $\mathrm{ERR}$ and $\mathrm{NCE}$ performance metrics have minimum values of $0$ where lower values correspond to better performance. A non-informative system, one where $p(\mathrm{nat}|\mathbf{W}, \mathbf{X})$ is $0.5$ for all samples, will yield an $\mathrm{NCE}$ of $1.0$ though higher values are possible ($\mathrm{ERR}$ has a maximum value of $1.0$).

For the multi-class ordinal task, a common metric is the mean absolute error ($\mathrm{MAE}$) between the predicted class label $f(\mathbf{W}, \mathbf{X})$ and the ground truth label for the sample:

$$\mathrm{MAE} = \frac{1}{M} \sum_{c\in\{1,...,M\}} \frac{1}{N_c} \sum_{\mathbf{X}\in c} |f(\mathbf{W}, \mathbf{X}) - c|$$

Here $N_c$ is the number of samples for each of the $M = 3$ classes. Similar to the binary task metrics $\mathrm{ERR}$ and $\mathrm{NCE}$, the ordinal $\mathrm{MAE}$ metric has a minimum value of $0.0$ and lower values correspond to better performance. Note also that all three metrics are balanced in that they effectively equalize the number of samples observed for each class.

## 5. Pronunciation Proficiency Calibration

As mentioned previously, we investigate two types of tasks for PPR: a binary task in which we try to estimate the posterior

| Ordinal Label | Binary Label | Source / Annotation | Spanish | | MSA | | Russian | |
|---|---|---|---|---|---|---|---|---|
| | | | Recordings | Speakers | Recordings | Speakers | Recordings | Speakers |
| 1 | $\overline{\text{nat}}$ | Learner / Beginner | 457 | 88 | 166 | 66 | 392 | 108 |
| 2 | N/A | Learner / Advanced | 1702 | 168 | 3110 | 177 | 2316 | 201 |
| 3 | nat | Learner / Native | 253 | 77 | 18 | 14 | 232 | 74 |
| 3 | nat | Teacher / Native | 201 | 21 | 241 | 27 | 190 | 19 |
| Total | | | 2613 | 211 | 3535 | 207 | 3130 | 235 |

Table 1: *Statistics for binary and ordinal classes on Spanish, MSA and Russian.*

probability that an utterance comes from a native speaker and a multi-class task in which we try to estimate the ordinal rank of a speaker's pronunciation proficiency for an utterance. A "beta" logistic regression is used for the binary task to estimate calibrated posterior probabilities while an ordinal regression is used for the multi-class task to predict pronunciation proficiency level. Both regressions are trained using the labeled data as described in Section 2. The following sections describe these models in more detail.

### 5.1. Beta Calibration

A method for calibrating posterior probabilities using a binary logistic regression is presented in [1]. The approach starts with the assumption that the observed samples are drawn from an unknown beta distribution and are bounded between zero and one. The log likelihood ratio of an observation $x$ for two classes $c_0$ and $c_1$ assuming data drawn from each class has a different beta distribution has the form:

$$ln\frac{p(x|c_1)}{p(x|c_0)} = a\ln x - b\ln(1-x) + c$$

Where $a$, $b$ and $c$ are the hyper parameters derived from the ratio of the two beta distributions and $a$ and $b$ must be non-negative to ensure that the ratio is monotonic in $x$. Note that this is equivalent to the log posterior odds ratio when the log prior odds ratio $\ln p(c_1)/p(c_0)$ is added as an offset.

Beta calibration uses a logistic regression to estimate the parameters $a$,$b$ and $c$ that minimize the normalized cross entropy (NCE) between the data labels ($c_0$ or $c_1$) and the predicted posterior probability:

$$\hat{p}(c_1|x) = \sigma\left(a\ln x - b\ln(1-x) + c\right)$$

where $\sigma(q) = \frac{1}{1+e^{-q}}$ is the sigmoid function which transforms the log posterior odds ratio into a posterior probability.

### 5.2. Inferred Phoneme Posterior Calibration

One of the goals of this work is to infer a calibrated posterior probability that a given phoneme was uttered by a native speaker using data that is labeled at the utterance level. The proposed statistical model assumes that there is only one possible segmentation $\mathbf{S}$ even if the acoustic data $\mathbf{X}$ consists of incorrect or incorrectly pronounced phonemes. The posterior probability that an utterance was produced by a native speaker can then be expressed as:

$$p(\text{nat}|\mathbf{W},\mathbf{X}) = \frac{p(\text{nat},\mathbf{W}|\mathbf{X})}{p(\mathbf{W})} \approx \frac{p(\text{nat},\mathbf{Q}|\mathbf{S},\mathbf{X})}{p(\mathbf{W})}$$

The naive Bayes assumption is used to express the joint posterior probability of the phone sequence $\mathbf{Q}$ given the data $\mathbf{X}$ and segmentation $\mathbf{S}$ as a simple product of the marginal phone posterior probabilities. Here the variable $q_i$ is used to represent the correct phoneme uttered by a native speaker:

$$p(\text{nat},\mathbf{Q}|\mathbf{S},\mathbf{X}) \approx \prod_{i=1}^{K} p(q_i|s_i,\mathbf{X}) \tag{1}$$

The probability that an utterance comes from a non-native speaker is restricted by the assumption that the segmentation $\mathbf{S}$ is fixed in that all non-native phones must occur within the same segments:

$$p(\overline{\text{nat}},\mathbf{Q}|\mathbf{S},\mathbf{X}) \approx \prod_{i=1}^{K} (1 - p(q_i|s_i,\mathbf{X})) \tag{2}$$

Note that the sum of the approximations in Equations 1 and 2 will not generally equal 1 (or even be close to 1) when $K > 1$. By combining Equations 1 and 2, an approximation to the log posterior odds ratio can now be expressed as:

$$\begin{aligned}\ln\frac{p(\text{nat},\mathbf{Q}|\mathbf{S},\mathbf{X})}{p(\overline{\text{nat}},\mathbf{Q}|\mathbf{S},\mathbf{X})} \approx &\sum_{i=1}^{K}\ln p(q_i|s_i,\mathbf{X}) \\ &- \sum_{i=1}^{K}\ln\left(1 - p(q_i|s_i,\mathbf{X})\right)\end{aligned} \tag{3}$$

In this work a logistic regression is used to estimate an utterance level posterior probability given an embedded beta calibration at the phoneme level. Using the approximation to the log posterior ratio in Equation 3 and the beta calibration approach described above in Section 5.1:

$$\begin{aligned}\hat{l}(\text{nat}|\mathbf{W},\mathbf{X}) = &\sum_{i=1}^{K} a_{q_i}\ln p(q_i|s_i,\mathbf{X}) \\ &- \sum_{i=1}^{K} b_{q_i}\ln\left(1 - p(q_i|s_i,\mathbf{X})\right) + c\end{aligned}$$

where $\hat{l}$ is the estimated log posterior odds ratio. Note that with this model, each of the unique phonemes $q$ has two weights $a_q$ and $b_q$. For this work, in order to reduce the overall number of parameters in the model the parameters $b_q$ are tied to a single parameter $b$.

From a machine learning perspective, the log beta phoneme posteriors (LBPPs) are sparse consisting of up to $L_q$ accumulated log phoneme posterior probabilities $\sum_{q_i \equiv q}\ln p(q_i|s_i,\mathbf{X})$ for each observed unique phoneme $q$ and the accumulated log phoneme posteriors $\sum_{i=1}^{K}\ln\left(1 - p(p_i|s_i,\mathbf{X})\right)$ for all other phonemes. The total number of parameters for the regression is then $L_q + 2$ which includes the intercept $c$. The software package used for estimating the regression is `glmnet_py`[2] which supports non-negative constraints on the parameters $a_q$ and $b$ and provides a means of combining $L_1$ and $L_2$ regularization [17].

[2] `https://web.stanford.edu/~hastie/glmnet_python/`

The LBPP features are normalized so that the same regression weights are applicable for both long and short utterances. While a very long utterance consisting of many words could include all $L_q$ phonemes (including repeat occurrences), a short utterance may include only a few unique phonemes occurring one or more times. To compensate for this disparity, utterance dependent weights are applied to the observed log posteriors where the per phoneme weight is $w_q = \frac{1}{c_t c_q}$ using $c_t$ as the total number of unique phonemes and $c_q$ as the number of times each unique phoneme $q$ occurs. With this approach, the calibrated posterior for each observed phoneme is given by:

$$\hat{p}(q|s, \mathbf{X}) = \sigma\left(a_q \ln p(q|s, \mathbf{X}) - b \ln\left(1 - p(q|s, \mathbf{X})\right) + c\right)$$

### 5.3. Ordinal Regression

For some regression problems, such as classifying recordings as either "beginner", "advanced", or "native", the classes are ordinal values and the the cost of an error is the difference between the class values. Several different ordinal regression methods have been developed over the years to address these types of problems. A detailed overview and comparison of these techniques can be found in [18]. In this work we have chosen to use the "all threshold" [2, 19] ordinal regression as implemented in the `mord`[3] Python package. This model is essentially a binary linear regression with a set of thresholds that define decision boundaries for the ordinal classes. Direct minimization of MAE to estimate the parameters of the ordinal regression is generally not feasible and a surrogate loss function (such as the logistic loss) is used instead [2].

Using the normalized LBPP features described above, the decision function for the ordinal regression is given by:

$$f(\mathbf{W}, \mathbf{X}) = 1 + \sum_{i=1}^{M-1} H\left(\sum_{i=1}^{K} a_{q_i} w_{q_i} \ln p(q_i|s_i, \mathbf{X}) - b \sum_{i=1}^{K} w_{q_i} \ln\left(1 - p(q_i|s_i, \mathbf{X})\right) + e_i\right)$$

where $\{e_1, ..., e_{M-1}\}$ are the decision thresholds for the $M = 3$ ordinal classes and $H(x)$ is the Heaviside step function. This approach is motivated by the idea that the approximation of the log posterior odds ratio in Equation 3 can be partitioned to predict ranked ordinal classes using the ordinal regression. The `mord` package was modified to support non-negative constraints on parameters $a_q$ and $b$ to satisfy the beta calibration constraints. The total number of parameters for the ordinal regression is $L_q + M$ (where $M = 3$ in this work).

### 5.4. Feature Fusion

Feature level fusion is straight forward for both the binary beta calibration and the ordinal regression. Features from another system are simply appending to the LBPP features. In this work LBPP features were fused with 23 scores from a LR system.

## 6. Experiments

Experiments were performed on MSA, Spanish and Russian where the Kaldi components of the system described in Section 3 were trained with the native speaker data described in Section 2.1 resulting in 30 to 50 unique phonemes ($L_q$) for each language. Data from all 22 foreign languages and WSJ English were used to train the i-vector LR system described in Section 3.2. Tenfold cross validation where each fold holds out unique speakers was used to train the beta calibration and ordinal regression using the language learner and teacher data described in Section 2.2.

[3] `https://github.com/fabianp/mord`

| Language | Features | Binary beta regression | |
|---|---|---|---|
| | | % ERR | NCE |
| MSA | GOP | 24.1 | 0.715 |
| | LBPPs | 20.6 | 0.675 |
| | + LR scores | 18.5 | 0.559 |
| Spanish | GOP | 36.1 | 0.861 |
| | LBPPs | 16.0 | 0.557 |
| | + LR scores | 14.1 | 0.524 |
| Russian | GOP | 12.6 | 0.497 |
| | LBPPs | 12.0 | 0.423 |
| | + LR scores | 10.0 | 0.400 |

Table 2: *Binary beta regression performance.*

| Language | Features | MAE |
|---|---|---|
| MSA | LBPPs | 0.524 |
| | + LR scores | 0.440 |
| Spanish | LBPPs | 0.412 |
| | + LR scores | 0.403 |
| Russian | LBPPs | 0.388 |
| | + LR scores | 0.362 |

Table 3: *Ordinal regression performance.*

Table 2 shows performance on the binary task described in Section 4 for MSA, Spanish, and Russian when using (1) uncalibrated average GOP features (2) calibrated LBPP features and (3) a fusion of calibrated LBPP and LR features. Calibration led to improvements over the baseline GOP features ranging from a 56% relative reduction in ERR for Spanish down to a 5% reduction in ERR for Russian. NCE relative improvements vary from 35% for Spanish down to 6% for MSA. Fusion with the i-vector based LR system yields a significant improvement for MSA (10% improvement in ERR and 17% improvement in NCE), Spanish (12% improvement in ERR and 6% improvement in NCE) and Russian (16% improvement in ERR and 5% improvement in NCE). Further investigation is required to understand the disparities in improvements across the different languages.

Performance on the ordinal task described in Section 4 for MSA, Spanish, and Russian are shown in Table 3. Again it is clear that performance on Spanish and Russian are significantly better than on MSA, but LR score fusion yields the highest gain for MSA (16% relative improvement in MAE compared to 2% and 7% for Spanish and Russian respectively).

## 7. Conclusions

This work demonstrates the utility of a beta calibration technique for estimating phonetic posteriors using utterance level transcriptions and using a threshold-based ordinal regression for predicting pronunciation proficiency level. Fusion of the LBPP features with LR scores yields significant gains in performance suggesting that further gains may be possible with more recent technologies such as such as the x-vector system [20]. It also appears that the beta calibrated approach is effective as indicated by the reasonably good performance of the Hybrid system.

# 8. References


[1] M. Kull, T. de Menezes e Silva Filho, and P. Flach, "Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers," in *AISTATS*, 2017.

[2] F. Pedregosa, F. Bach, and A. Gramfort, "On the consistency of ordinal regression methods," *Journal of Machine Learning Research*, 2017.

[3] N. Dehak, P. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via ivectors and dimensionality reduction," in *Interspeech*, 2011.

[4] H. Franco, L. Neumeyel, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *ICASSP*, 1997.

[5] H. Ryu and M. Chung, "Mispronunciation diagnosis of l2 english at articulatory level using articulatory goodness-of-pronunciation features," in *ISCA Workshop on Speech and Language Technology in Education*, 2017, pronunciation scoring.

[6] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communications*, 2000, pronunciation.

[7] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study," in *ISCA Workshop on Speech and Language Technology in Education*, 2009, pronunciation scoring.

[8] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communications*, 2015, pronunciation scoring.

[9] N. F. Chen, R. Tong, D. Wee, P. Lee, B. Ma, and H. Li, "iCALL corpus: Mandarin chinese spoken by non-native speakers of european descent," in *Interspeech*, 2015, pronunciation scoring.

[10] R. Duan, T. Kawahara, M. Dantsuji, and H. Nanjo, "Transfer learning based non-native acoustic modeling for pronunciation error detection," in *ISCA Workshop on Speech and Language Technology in Education*, 2017, pronunciation scoring.

[11] H. Strik, K. Truong, F. de Wet, and C. Cucchiari, "Comparing different approaches for automatic pronunciation error detection," *Speech Communications*, 2009, pronunciation scoring.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesel, "The kaldi speech recognition toolkit," in *Proc. of IEEE ASRU*, 2011.

[13] T. Ko, V. Peddinti, D. Povey, M. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*, 2017.

[14] D. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Workshop and Speech and Natural Language*, 1992.

[15] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech*, 2011, pp. 249–252.

[16] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," in *IEEE Signal Processing Letters*, 2015.

[17] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, 2010.

[18] P. Gutierrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervas-Martinez, "Ordinal regression methods: Survey and experimental study," *IEEE Transactions on Knowledge and Data Engineering*, 2016.

[19] J. Rennie and N. Srebro, "Loss functions for preference levels: Regression with discrete ordered labels," in *IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, 2005.

[20] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors," in *IEEE Odyssey*, 2018.