IMPROVEMENTS IN PREDICTING CHILDREN'S OVERALL READING ABILITY BY MODELING VARIABILITY IN EVALUATORS' SUBJECTIVE JUDGMENTS

Matthew P. Black and *Shrikanth S. Narayanan*

Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, CA

http://sail.usc.edu

ABSTRACT

Automatic literacy assessment is one promising application of speech and language processing research. In our previous work, we showed we could accurately predict children's overall ability to read a list of English words aloud, an integral component of early literacy assessment. In this paper, we improve upon our results by exploiting the fact that evaluators' level of agreement significantly varies, depending on the child being judged. This source of evaluator variability is directly modeled using generalized least squares linear regression. In this framework, the children for which the evaluators were more confident in rating are weighted higher. Performance in predicting the mean evaluator's scores increases from a Pearson's correlation coefficient of 0.946 to 0.952, a relative improvement of 0.63%. This is a significantly higher correlation than the mean inter-evaluator agreement of 0.899 ($p < 0.05$). Critically, the mean and maximum absolute errors are significantly reduced.

Index Terms— Automatic literacy assessment, pronunciation evaluation, children's speech, generalized least squares regression

1. INTRODUCTION

Literacy assessment is an important aspect of children's early education. Automatic literacy assessment has the potential to help with this process by taking some of the burden off teachers, allowing them to concentrate on lesson planning and providing individualized help to their students. Automatic literacy assessment research has been applied to children of all ages and reading skill levels, through pronunciation evaluation of basic reading tasks like correctly reading letter-sounds and letter-names [1] and isolated words [2–4], to reading full sentences and stories [5–7].

While the majority of automatic literacy assessment research has focused on detecting reading errors made by the child, there is also a need to automatically estimate children's *overall* performance on a reading task. These high-level integrative assessments may be especially important for student stratification, providing a teacher with an objective way to quickly identify students that may need more assistance. Automatically quantifying overall performance may also be useful to track children's progress over time.

Our previous work in [4] demonstrated one approach to automatically predicting young children's overall performance in reading a list of words aloud. In this research, multiple evaluators first listened to the children's speech and rated their overall reading ability on a scale from 1 to 7. Next, acoustic features were extracted that were correlated with cues evaluators stated were most important: pronunciation correctness, fluency, and speaking rate. Finally, we used supervised least squares linear regression techniques that predicted the average evaluator's ratings with Pearson's correlation of 0.946; this exceeded the average inter-evaluator agreement correlation of 0.899, although not significantly at the 5% significance level ($p > 0.05$).

One weakness to our proposed approach in [4] was that we did not take into account the fact that the variability in ratings across evaluators was not constant for all children; evaluators were in complete agreement for some children and disagreed more for other children. This variable level of evaluator uncertainty could potentially be incorporated during model training. In addition, we will show that this *heteroscedasticity* in evaluators' subjective judgments (having a non-constant variance) violates an assumption of the least squares linear regression techniques proposed in [4]. We addressed this weakness in this paper by employing generalized least squares linear regression methods that account for this "variable variability" in evaluators' scores across children.

Section 2 describes the corpus and human evaluation, and Section 3 discusses the acoustic features we extracted. Section 4 describes the baseline and proposed supervised learning methods used to predict the children's overall reading ability. The results are presented and discussed in Section 5, and the conclusions and intended future work are provided in Section 6.

2. CORPUS

2.1. TBALL Project & Corpus

The Technology-Based Assessment of Language and Literacy (TBALL) Project was established to automatically assess the English literacy skills of young children in early education from multilingual backgrounds [8, 9]. Toward this goal, we designed a human-computer interface to test children in kindergarten to second grade on age-appropriate reading tasks. The recorded speech data, collected from children in actual elementary schools in California using a close-talking microphone, makes up the TBALL Corpus [10].

For our previous and current work, we analyzed a subset of children who were administered the isolated word-reading task. For this task, children read aloud a list of 55 pre-determined English words that progressively became more challenging; the list started with the word, "map," and ended with, "transportation." One word was displayed on the computer monitor at a time, and the child had up to five seconds to say the word before the next one was shown.

We noticed several interesting spoken phenomena during this isolated word-reading task. In an annotated subset of the data comprised of 2800 single-word utterances, we found that 37.1% of the target words were mispronounced. In addition, there were a variety of *disfluencies*; these included hesitations (e.g., partial word repetitions), sound-outs (where the child would say each individual phone in the target word before pronouncing the full word), elongations (unnaturally lengthening a phone or syllable of the target word) and

This research was supported by the National Science Foundation.

eval	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	mean
corr	0.83	0.84	0.87	0.89	0.89	0.91	0.92	0.92	0.93	0.94	0.95	**0.899**

Table 1. Inter-evaluator agreement between the 11 evaluators, computed as the Pearson's correlation between an individual evaluator's scores and the mean scores of the other 10 evaluators.

speaking the target word with a question intonation (perhaps conveying uncertainty). We found that 23.0% of the 2800 single-word utterances contained at least one disfluency [4]. Finally, the children spoke at different speaking rates, with some children immediately saying the words out loud when they were displayed and others needing the full five seconds to read each word.

2.2. Subjective human evaluation

To analyze how these various factors (pronunciation correctness, fluency, speaking rate) affected evaluators' perception on the children's performance, we selected 42 children that displayed a wide variety of performance levels and reading styles. Eleven English-speaking evaluators listened to the speech of the 42 children and rated each on his/her "overall reading ability." These subjective judgments were on an integer scale from 1 ("poor") to 7 ("excellent").

While none of the 11 evaluators were licensed teachers, we found in previous work that the inter-evaluator agreement between teachers and non-experts was not significantly different for a related pronunciation verification task [11]. In this study, we computed inter-evaluator agreement by calculating Pearson's correlation coefficient between the scores of an individual evaluator's scores and the mean scores of the other ten evaluators. Table 1 shows that agreement ranged from 0.83 to 0.95, with mean inter-evaluator agreement of 0.899. Since all 11 evaluators' agreement statistics were high, we chose to treat each evaluator equally in this work. Please see [4] for more information on the 11 evaluators' backgrounds.

Figure 1 is a plot of the mean and standard deviation in the overall reading ability scores assigned to each child, computed across all evaluators. We see from this figure that the mean scores ranged from 1.55 to 7, and the standard deviations ranged from 0 (all evaluators agreed for 2 of the 42 children) to 1.29. The lowest standard deviations occurred for the children with higher mean scores. This makes numerical sense for the children with mean scores greater than 6.5 because all evaluators assigned scores of 6 or 7. However, it can also be argued that these children are objectively *easier* to grade, since they spoke most of the words correctly and had few disfluencies.

On the other hand, evaluators tended to agree less for the children with more pronunciation errors and more disfluencies; these cues may have impacted the evaluators to differing degrees. Thus, it can be argued that it is more *subjective* to grade the children with the higher standard deviations. In particular, the child with the highest standard deviation (who was assigned scores that ranged from 2 to 7) pronounced almost all of the words correctly but sounded out each word beforehand; it is possible that some evaluators largely ignored these sound-out disfluencies, while others felt it was strong evidence that the child was not (yet) the most skilled reader.

While Figure 1 provides visual evidence that the evaluators' level of agreement varied across children, we also employed two statistical hypothesis tests for heteroscedasticity: Levene's test [12] and the Brown-Forsythe test [13]. For both tests, we could reject the null hypothesis of homoscedasticity in the evaluators' scores at the 5% significance level (Levene's: $p < 0.001$, Brown-Forsythe: $p < 0.05$). This validates our decision in this work to pursue generalized least squares linear regression methods, which do not assume the evaluators' overall scores have equal variance for each child.

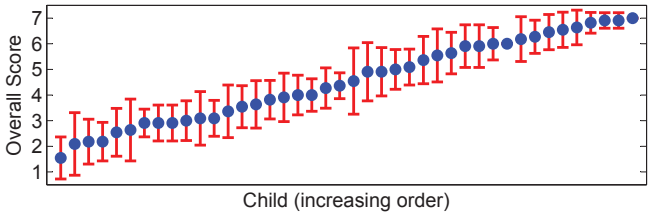

Fig. 1. The mean and standard deviation in the overall scores assigned to each child, computed across all 11 evaluators.

3. ACOUSTIC FEATURE EXTRACTION

We used the identical set of acoustic features in this paper as in our past work [4]. The feature extraction was a two stage process. In the first stage, we extracted 48 "scores" for each single-word utterance spoken by the 42 children. Each score was related to one of three cues that evaluators stated were important when judging the children's overall reading ability: pronunciation correctness, fluency, and speaking rate. In the second stage, we computed 12 functionals (e.g., mean, standard deviation) across all the words spoken by the child for each of the 48 scores. Therefore, our final feature set consisted of 576 (48×12) features per child.

In the first stage, the 48 scores were broken down as follows: 10 pronunciation correctness, 12 fluency, and 26 speaking rate. The pronunciation correctness scores were based on two common pronunciation verification methods: 1) forced alignment with a dictionary of acceptable and foreseeable unacceptable pronunciations of the target word, and 2) Goodness of Pronunciation (GOP) scoring [14]. The fluency scores were based on constrained automatic speech recognition using disfluency-specialized grammars, which were designed to detect partial word instantiations of the target word. Finally, the speaking rate scores were based on forced alignment and captured relevant timing information, such as the speech start time (relative to when the word was first displayed on the monitor) and the average speaking rate in units of syllables/s and phones/s. Further details on the feature extraction process can be found in [4].

4. LEARNING METHODS

Since we are treating all evaluators equally (Section 2.2), our goal in this paper was to predict the overall reading ability scores from the *mean* evaluator (Figure 1). We explain the baseline system in Section 4.1 and our proposed methods in Section 4.2. For all methods, we used leave-one-out cross-validation to separate training data (41 children) from the test child. We optimized all regression parameters (e.g., selected features, smoothing/tuning parameters) using another stage of leave-one-out cross-validation on each train set separately.

4.1. Least squares linear regression

The baseline learning method, least squares (LS) linear regression, was based on our previous work [4]. The problem is defined as:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector comprised of the mean evaluator scores for each child, X is the noiseless $n \times m$ feature matrix (with a $n \times 1$ ones vector appended to account for the intercept/offset term), $\boldsymbol{\beta}$ is the $m \times 1$ linear weight vector, and the $n \times 1$ residual vector $\boldsymbol{\epsilon}$ is *assumed* to be homoscedastic. The optimal linear weights $\hat{\boldsymbol{\beta}}$ that

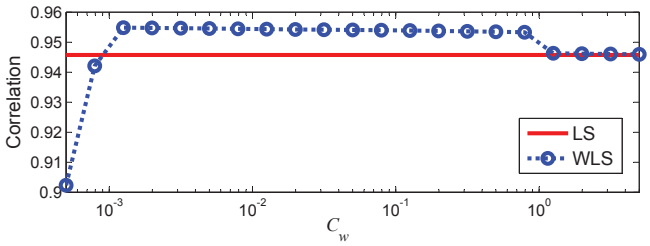

Fig. 2. Performance, in terms of Pearson's correlation, in predicting the children's overall reading ability using the weighted least squares (WLS) method, as a function of the tuning parameter C_w.

minimize the sum of the squared residual, $\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2$, are:

$$\hat{\boldsymbol{\beta}}_{ls} = \left(X^\mathsf{T}X\right)^{-1}X^\mathsf{T}\boldsymbol{y} \tag{2}$$

Due to dimensionality issues and multicollinearity effects, we did not use all 576 features in X. Instead, we used sequential forward feature selection to iteratively select features and construct X that maximized Pearson's correlation between $\boldsymbol{y}$ and $X\hat{\boldsymbol{\beta}}$ on the train set. Two or three features were selected, depending on the cross-validation fold ($m = \{2, 3\}$). Therefore, $n > m$, and we never had the problem of an under-determined system.

4.2. Generalized least squares linear regression

The least squares solution shown in Equation 2 is only optimal when the assumption of homoscedasticity in $\boldsymbol{\epsilon}$ holds. However, since we showed in Section 2.2 that $\boldsymbol{y}$ is heteroscedastic, we see in Equation 1 that $\boldsymbol{\epsilon}$ too will be heteroscedastic. This led us to employing generalized least squares linear regression methods [15]. In this formulation, the optimal linear weights, in the least squares sense, are:

$$\hat{\boldsymbol{\beta}} = \left(X^\mathsf{T}\Omega X\right)^{-1}X^\mathsf{T}\Omega\boldsymbol{y}, \tag{3}$$

where Ω is a diagonal matrix, with diagonal elements $\Omega_{jj} = 1/\sigma_j^2$, where σ_j is the "true" standard deviation in the overall reading ability of child j; see [15] for a derivation. In this paper, we estimated Ω in two ways: 1) by using the scores provided by the 11 evaluators, and 2) by iteratively estimating Ω from the prediction residuals. We refer to the former method as weighted least squares (WLS) and the latter method as feasible generalized least squares (FGLS)[1].

Equation 4 shows how we computed the WLS estimate of Ω, where $\tilde{\sigma}_j$ is the estimated standard deviation in the overall reading ability of child j, computed from the evaluators' scores (Figure 1), and C_w is a positive smoothing parameter:

$$\Omega_{wls} = \text{diag}\left(\frac{1}{\tilde{\sigma}_1^2 + C_w}, \cdots, \frac{1}{\tilde{\sigma}_n^2 + C_w}\right), \quad C_w > 0 \tag{4}$$

The WLS method has the benefit of requiring only one additional parameter, C_w, which is needed to avoid numerical problems for the case when all evaluators agree ($\tilde{\sigma}_j = 0$). C_w can also be viewed as a tuning parameter; as C_w is increased, the solution to $\hat{\boldsymbol{\beta}}_{wls}$ (Equation 3) tends to $\hat{\boldsymbol{\beta}}_{ls}$ (Equation 2). Figure 2 demonstrates the effectiveness of the WLS method in predicting the mean evaluator's overall reading ability scores for a large range of C_w values.

For the FGLS method, we iteratively estimated Ω. See Algorithm 1 for pseudocode of our implementation, which was based on [15]. We first found the WLS solution on the training data and computed the prediction residual vector, which were used to initialize the FGLS iteration process. At each FGLS iteration i, the residual vector was used to construct a new FGLS diagonal matrix Ω_i (step 8). The form of Ω_i is very similar to Ω_{wls} (Equation 4), except Ω_i is determined analytically from the trained model, while Ω_{wls} is computed from the evaluators' scores. The FGLS smoothing parameter, C_f, in step 8 of the algorithm is analogous to the C_w term in Equation 4. We selected C_f using a grid search, choosing the value that maximized the Pearson's correlation between the diagonal entries of Ω_i and Ω_{wls}; this tuning method was used to avoid overtraining and numerical issues. In steps 9 and 10 of Algorithm 1, new estimates for the FGLS linear weights $\hat{\boldsymbol{\beta}}_i$ were computed and a new residual vector was calculated. This iterative process was repeated until the sum of the squared residuals no longer decreased on the training data. After convergence, the trained model was then applied to the test data. We found the FGLS algorithm converged in 3 to 12 iterations, depending on the cross-validation fold.

Algorithm 1 Feasible generalized least squares (FGLS)

Require: Training data (feature matrix: X, dependent variable: $\boldsymbol{y}$)

1: Compute weighted least square (WLS) solution: $\hat{\boldsymbol{\beta}}_{wls}$
2: Compute WLS residual column vector: $\boldsymbol{\epsilon}_{wls} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{wls}$
3: Compute sum of squared residual: $E_{wls} = \boldsymbol{\epsilon}_{wls}^\mathsf{T}\boldsymbol{\epsilon}_{wls}$
4: Initialize FGLS: $\hat{\boldsymbol{\beta}}_0 \leftarrow \hat{\boldsymbol{\beta}}_{wls}$, $\boldsymbol{\epsilon}_0 \leftarrow \boldsymbol{\epsilon}_{wls}$, $E_0 \leftarrow E_{wls}$
5: Initialize FGLS iteration counter: $i \leftarrow 0$
6: **repeat**
7: Increment FGLS iteration counter: $i \leftarrow i + 1$
8: Compute diagonal FGLS matrix:
$\Omega_i = \text{diag}\left(\frac{1}{\epsilon_{i-1,1}^2 + C_f}, \cdots, \frac{1}{\epsilon_{i-1,n}^2 + C_f}\right), \quad C_f > 0$
9: Compute FGLS coefficients: $\hat{\boldsymbol{\beta}}_i = (X^\mathsf{T}\Omega_i X)^{-1}X^\mathsf{T}\Omega_i\boldsymbol{y}$
10: Compute FGLS residual column vector: $\boldsymbol{\epsilon}_i = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}_i$
11: Compute FGLS sum of squared residual: $E_i = \boldsymbol{\epsilon}_i^\mathsf{T}\boldsymbol{\epsilon}_i$
12: **until** $E_i \geq E_{i-1}$

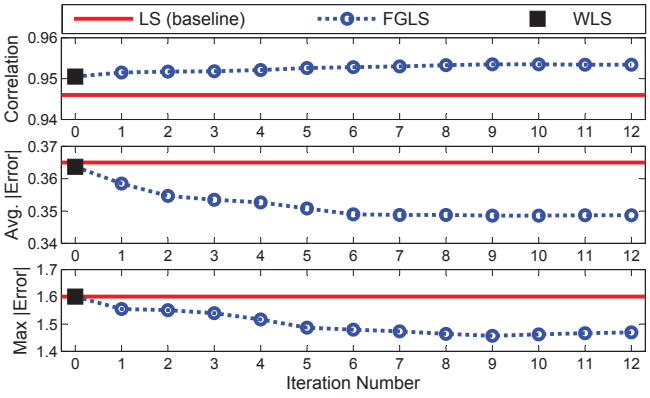

Fig. 3. Performance, in terms of the 3 metrics, of the 3 proposed systems: baseline least squares (LS), weighted least squares (WLS), and 12 iterations of feasible generalized least squares (FGLS).

For illustrative purposes, Figure 3 shows the performance of the FGLS method in predicting the mean evaluator's overall reading ability scores, as a function of the FGLS iteration. While we only attained a small gain in performance over WLS with respect to Pearson's correlation, we do get a larger relative boost in performance for the two secondary metrics used in our previous work [4]: the *mean* absolute error in predictions and the *maximum* absolute

[1]FGLS is also commonly known as iteratively reweighted least squares.

System	*Performance metric*		
	Corr	$\|E\|_{avg}$	$\|E\|_{max}$
Least Squares (LS) – Baseline	0.946	0.365	1.601
Weighted Least Squares (WLS)	0.951	0.364	1.601
Feasible Generalized LS (FGLS)	**0.952**	**0.356**	**1.579**

Table 2. Performance, in terms of the 3 metrics, of the 3 proposed systems: baseline least squares (LS), weighted least squares (WLS), and feasible generalized least squares (FGLS).

error in predictions (out of the 42 children). This suggests that the FGLS method helps improve the robustness in estimating the linear weight coefficients $\hat{\boldsymbol{\beta}}$ by starting from the WLS solution and iteratively incorporating uncertainty in the trained model.

5. RESULTS & DISCUSSION

Comparable results for the three learning methods, attained by selecting features and optimizing all learning parameters using cross-validation, are shown in Table 2. We see that both proposed methods (WLS and FGLS) equaled or outperformed the baseline LS method for all three performance metrics. While there were no significant differences in the correlation coefficients of the three methods, the incremental improvements achieved with the WLS and FGLS methods made their correlations significantly higher than the mean inter-evaluator agreement of 0.899 (Table 1), with both $p < 0.05$.

The WLS method, which directly modeled evaluators' variability across children, achieved a Pearson's correlation coefficient of 0.951 between the predicted scores and the mean evaluator's scores, a relative improvement of 0.53% over baseline LS linear regression. The best overall system for all three performance metrics was FGLS linear regression, with relative improvements over baseline LS linear regression of 0.63%, 2.5%, and 1.4% for the correlation, average absolute error, and maximum absolute error performance metrics, respectively. The FGLS method has the benefit of being initialized with the WLS solution and making further changes based on the heteroscedasticity of the residual from the trained model.

6. CONCLUSIONS & FUTURE WORK

In this work, we showed we could improve the predictive power of a high-level automatic literacy assessment system by incorporating variability in evaluators' uncertainty across children. We used two generalized least squares linear regression techniques that accurately predicted children's overall ability to read a list of words aloud, significantly outperforming average inter-evaluator agreement. These methods exploit the fact that there are variable levels of subjectivity in assessing children's reading ability, depending on the behaviors exhibited by the children. We hope the techniques proposed in this work can be applied to other learning problems that involve modeling the perceptions of multiple evaluators.

One area of future work is to take into account evaluator *reliability*, as opposed to treating each evaluator equally; this has been shown to be advantageous in the context of emotion classification [16]. The inter-evaluator agreement statistics listed in Table 1 vary for the 11 evaluators, so it is possible that some evaluators are more reliable than others. We may be able to predict the evaluators' scores better if we weighted the scores of the more reliable evaluators higher. Unfortunately, initial experiments that used evaluator reliability-weighted linear combinations of the scores (using the agreement statistics in Table 1 as a measure of reliability) did not increase automatic prediction performance. Future research will experiment with other reliability metrics to find more robust ways of combining multiple evaluators' perspectives (e.g., by using data-dependent evaluator modeling as in [17]).

Finally, we also hope to extend this high-level literacy assessment system to other important reading tasks. Our ultimate goal is to deploy this type of system in an actual elementary classroom, where it could be trained to mimic the grading trends of the teacher or a bank of teachers and provide feedback in near real-time.

7. REFERENCES

[1] M. P. Black, A. Kazemzadeh, J. Tepperman, and S. S. Narayanan, "Automatically assessing the ABCs: Verification of children's spoken letter-names and letter-sounds," *ACM Transactions on Speech and Language Processing*, vol. 7, no. 4, article 15, Aug. 2011.

[2] J. Duchateau, L. Cleuren, H. Van hamme, and P. Ghesquière, "Automatic assessment of children's reading level," in *Proc. of Interspeech*, 2007.

[3] J. Tepperman, S. Lee, A. Alwan, and S. Narayanan, "A generative student model for scoring word reading skills," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 2, pp. 348–360, 2011.

[4] M. P. Black, J. Tepperman, and S. S. Narayanan, "Automatic prediction of children's reading ability for high-level literacy assessment," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 1015–1028, Aug. 2011.

[5] J. Mostow, S. F. Roth, E. G. Hauptmann, and M. Kane, "A prototype reading coach that listens," in *Proc. of AAAI*, 1994.

[6] P. Cosi and B. Pellom, "Italian children's speech recognition for advanced interactive literacy tutors," in *Proc. of Interspeech*, 2005.

[7] A. Hagen, B. Pellom, and R. Cole, "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Communication*, vol. 49, no. 12, pp. 861–873, 2007.

[8] A. Alwan, Y. Bai, M. P. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, "A system for technology based assessment of language and literacy in young children: The role of multiple information sources," in *Proc. of MMSP*, 2007.

[9] P. Price, J. Tepperman, M. Iseli, T. Duong, M. P. Black, S. Wang, C. K. Boscardin, M. Heritage, P. David Pearson, S. Narayanan, and A. Alwan, "Assessment of emerging reading skills in young native speakers and language learners," *Speech Communication*, vol. 51, no. 10, pp. 968–984, 2009.

[10] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, "TBALL data collection: The making of a young children's speech corpus," in *Proc. of Interspeech*, 2005.

[11] J. Tepperman, J. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, and S. Narayanan, "Pronunciation verification of children's speech for automatic literacy assessment," in *Proc. of Interspeech*, 2006.

[12] H. Levene, "Robust tests for equality of variances," in *Contributions to Probability and Statistics*, I. Olkin, Ed., pp. 278–292. Stanford University Press, Palo Alto, CA, 1960.

[13] M. B. Brown and A. B. Forsythe, "Robust tests for the equality of variances," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 364–367, 1974.

[14] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[15] R. J. Carroll and D. Ruppert, *Transformation and Weighting in Regression*, chapter 2: Generalized least squares and the analysis of heteroscedasticity, Chapman & Hall, New York, NY, 1988.

[16] K. Audhkhasi and S. S. Narayanan, "Emotion classification from speech using evaluator reliability-weighted combination of ranked lists," in *Proc. of ICASSP*, 2011.

[17] K. Audhkhasi and S. S. Narayanan, "Data-dependent evaluator modeling and its application to emotional valence classification from speech," in *Proc. of Interspeech*, 2010.