



How Does Alignment Error Affect Automated Pronunciation Scoring in Children’s Speech?

Prad Kadambi¹, Tristan Mahr², Lucas Annear², Henry Nomeland², Julie Liss¹, Katherine Hustad², Visar Berisha¹

¹Arizona State University, USA

²University of Wisconsin-Madison, USA

pkadambi@asu.edu, tristan.mahr@wisc.edu, katie.hustad@wisc.edu, visar@asu.edu

Abstract

Automated goodness of pronunciation scores measure deviation from typical adult speech by first phonetically segmenting speech using forced alignment and then computing phoneme likelihoods. Care must be taken to distinguish between the impact of alignment error (a spurious signal) and true acoustic deviation on the automated score. Using mixed effects modeling, we predict $\Delta PLLR$, the difference between pronunciation scores computed using manual alignment ($PLLRR_m$) versus computed using automatic forced alignments ($PLLRR_a$). Pronunciation deviations and alignment error are both magnified in children’s speech and may be influenced by factors such as phoneme position and phoneme type. Our methodology shows that alignment error has a moderate effect on $\Delta PLLR$, and other variables have small to no effect. Manual PLLR closely matches automatically calculated PLLR following cross utterance averaging. Thus, practical comparisons between child speakers should be very comparable across the two methods.

Index Terms: forced alignment, goodness of pronunciation, automatic pronunciation evaluation, phoneme segmentation, alignment error

1. Introduction

Automated pronunciation scoring systems quantify a speaker’s acoustic deviation relative to typical adult speech [1]. A variety of applications rely on automated pronunciation scores such as the study of clinically disordered speech [2, 3], assessment of second language (L2) learners [4, 5, 6], disfluency detection in children [7], and clinical analysis of child speech [8]. They can also play an important role in computer-aided pronunciation training (CAPT) systems—useful tools for clinicians in treating and evaluating preschool-to-school age children with phonological disorders, which affect up to 10% of children [9].

1.1. Goodness of Pronunciation Using PLLR

A family of algorithms for automated phonetic pronunciation scoring is based on the goodness of pronunciation (GOP) metric [10, 11]. Specifically, we focus on a variant of GOP which computes a phoneme log-likelihood ratio (PLLR) between the speaker’s target phoneme and the produced phoneme. The GOP pipeline is comprised of two steps: 1. a phonetic segmentation step to identify phoneme time boundaries and 2. evaluation of corresponding phoneme log-likelihoods at each time interval using an acoustic model (Figure 1).

GOP computation begins with an utterance and its orthographic transcription. Based on the transcription, the expected phoneme sequence $Q = [q_1, \dots, q_K]$ is generated.

Phonetic Segmentation. Next, phonetic segmentation—either

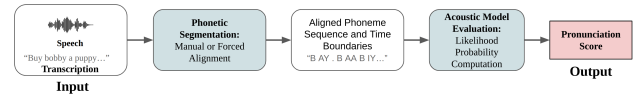


Figure 1: Pipeline for calculating pronunciation score (GOP).

via manual annotation of alignments or automatic segmentation with forced alignment (FA)—identifies time boundaries of each phoneme $q_k \in Q$. Alignment provides the frame indices \mathbf{f}_k where phoneme q_k is expected to occur. These frames are 10-ms segments in which acoustic features are calculated. We term PLLR computed via automatic forced alignment $PLLRR_a$, or automatic pronunciation score. PLLR computed via manual alignment annotation is termed $PLLRR_m$, or manual pronunciation score.

Acoustic Model Evaluation. Given these phoneme time boundaries, the GOP for the k^{th} phoneme is the phoneme log-likelihood ratio:

$$PLLRR(q_k) = \frac{1}{|\mathbf{f}_k|} \sum_{i \in \mathbf{f}_k} \log \frac{Pr(O_i|q_k)}{Pr(O_i|q^*)} \quad (1)$$

where O_i is the observed acoustic feature for frame i . The PLLR compares the likelihood of the *target* phoneme from segmentation (q_k) to the likelihood of the *most likely* phoneme (q^*) (i.e., the phoneme that best fits the acoustic features). The probabilities in (1) are typically evaluated using an acoustic model trained on adult speakers from the general population. This group serves as the pronunciation standard.

The PLLR score is affected by two main factors: true pronunciation deviation and forced alignment errors. Common FA systems [12, 13] rely on acoustic models trained for ASR on adult speakers alone and can perform poorly on speech not represented in their training data. Child speech deviates significantly from adult speech and exhibits far higher intra- and interspeaker variability. Indeed, study of ASR on child speech has found word error rates (WER) 2–5x higher than WER for adult speech [14], and evaluation of FA systems on child speech [15] has found age dependence in aligner performance and differences in alignment performance across phoneme classes. The presence of FA error obfuscates the interpretability of $PLLRR_a$ as a pronunciation score, as any variation in $PLLRR_a$ could be due to alignment error, underlying acoustic deviation in the utterance, or both.

In [8], the authors attempt to disentangle the relationship between acoustic deviation (as measured by $PLLRR_m$), alignment error, and $PLLRR_a$, by using a linear model to predict $PLLRR_a$ as a function of alignment error. They found that alignment error alone explained a moderate percentage of the variance in $PLLRR_a$. However, when estimating $PLLRR_a$ as a linear function of *both* alignment error and $PLLRR_m$, they found that

$PLL R_m$ explained almost all of the variance in $PLL R_a$ and that the explanatory contribution of alignment error was scant. Although their analysis provides useful information on the relative importance of alignment error in predicting $PLL R_a$, it does not provide information on the conditions under which the difference between the manual and automatic pronunciation scores ($\Delta PLL R$) is large. That is, their analysis did not fully explain what the downstream impact of alignment error is on automatic pronunciation scoring.

To ensure confident use of GOP-like automatic pronunciation scores in CAPT systems and similar applications, it is necessary to understand for which conditions $PLL R_m$ differs significantly from $PLL R_a$. If large discrepancies exist between the two for certain phoneme positions in an utterance, certain phoneme types, or beyond a certain alignment error, this disagreement serves as impetus for improving either of the two components of the automated pronunciation scoring pipeline (alignment and acoustic model), surfacing this information to eventual users, or disqualifying its use for those subsets of the data.

In this paper, we demonstrate a methodology for evaluating the key factors contributing to $\Delta PLL R$ —that is, the difference in pronunciation scores calculated using manual alignments ($PLL R_m$) versus those calculated using automatic forced alignments ($PLL R_a$). This methodology helps verify that for our corpus of child speech, the automated scores agree with manual scores under specific groupings of interest (by phoneme type and phoneme position). We find that the largest predictor of PLLR mismatch is alignment error, although the effect is only moderate in size. The effects of phoneme position and speaker age in estimating mismatch were statistically significant but small in magnitude. Additionally, general phoneme type (e.g., fricative, vowels, etc.) was not a significant predictor of mismatch. In fact, for all phoneme types, the standardized distributions of utterance-averaged $PLL R_a$ and $PLL R_m$ were highly similar, and correlation between the two was high.

2. Methods

2.1. Dataset

A gender matched speech corpus consisting of 42 typically developing children—native speakers of American English—was used. By age group, 10 children were 3–4-years-old, 10 children were 4–5-years-old, 12 children were 5–6-years-old, and 10 children were 6–7-years-old. A total of 3764 files (total duration 128 minutes) were collected. An average of 89 utterances were elicited per child. Single word, multi-word, and full sentence phrases were elicited in a picture-prompted repetition task based on the Test of Children’s Speech [16].

Manual Alignment. Manual alignments were first annotated from forced alignments generated using Prosodylab [12]. These alignments were then manually corrected by adjusting phoneme time boundaries in Praat [17]. Two trained manual annotators, a certified speech-language pathologist (SLP) and an SLP graduate student, provided manual annotations of phonetic boundaries (labeling 2801 and 963 files respectively).

Forced Alignment. Forced alignments were generated using a modified, speaker-adaptive version of the system in [18]. This system was trained on CommonVoice [19]. Manual alignments were used as the ground truth labels for training the alignment system such that forced alignments were generated using leave-one-speaker-out cross evaluation. That is, the following procedure was applied for each speaker: data for a given speaker was

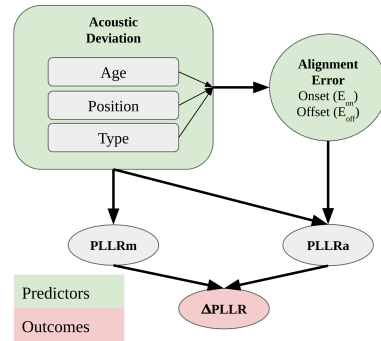


Figure 2: Acoustic deviation impacts $PLL R_m$, $PLL R_a$, and alignment error. However, alignment error also impacts $PLL R_a$. Age, phoneme position, and phoneme category (type) are key factors that determine the level of deviation.

held out, the pretrained alignment model was fine tuned on the manual alignments from the remaining 41/42 speakers, and the fine-tuned model was evaluated on the held out speaker’s data. This procedure was repeated for every speaker beginning with a pretrained model every time to ensure no data leakage. For comparison, a second set of forced alignments was also generated using the popular Montreal Forced Aligner (MFA) [13] with speaker adaptation enabled, and the statistical analyses were repeated for this second set of forced alignments.

2.2. PLLR Calculation

PLLR scores were calculated individually for each phoneme in each utterance as in Equation 1. Probabilities were evaluated in Kaldi [20] with a TDNN-F acoustic model trained on 960 hours of audio from the LibriSpeech [21] dataset (healthy adult native speakers). PLLR scores were calculated for manual alignments, $PLL R_m$, and for forced alignments, $PLL R_a$.

2.3. Statistical Modeling

Outcome Variable. We sought to identify the factors that lead to the largest discrepancy between $PLL R_m$ and $PLL R_a$. Thus, the target variable was the difference, $\Delta PLL R$:

$$\Delta PLL R = PLL R_m - PLL R_a. \quad (2)$$

Predictor Variables. $\Delta PLL R$ was modeled as a function of phoneme type (*Type*), phoneme position (*Position*), and alignment error. Both percent onset error and percent offset error (E_{on} and E_{off}) were included as predictors. These errors for an interval were calculated as the percentage onset/offset error relative to the length of the hand-aligned interval:

$$\begin{aligned} E_{on} &= 100 * (t_{on} - \hat{t}_{on}) / (t_{off} - t_{on}) \\ E_{off} &= 100 * (t_{off} - \hat{t}_{off}) / (t_{off} - t_{on}) \end{aligned} \quad (3)$$

where t_{on} and t_{off} are the hand-annotated onset/offset times and \hat{t}_{on} and \hat{t}_{off} are the onset/offset times estimated by FA.

Position had three levels: utterance-initial, utterance-final, or utterance-medial. This signified the first phoneme in the utterance, the last phoneme in the utterance, and any phoneme in between the first and last phonemes respectively. These utterance-position features were selected because they deal with a special subproblem in alignment: identifying the start/stop of an utterance. *Type* had six levels corresponding to particular classes of phonemes: vowels (in ARPAbet: IY, IH, EY, EH, AE,

Table 1: Average forced alignment error by phoneme category and utterance position. Proportion and count indicate the fraction of the dataset and the number of examples, respectively.

Category	E_{on} % (SD)	E_{off} % (SD)	Proportion (Count)
Vowels	5.9 (18.7)	3.6 (20.2)	.39 (10213)
Approximants	11.4 (46.0)	6.1 (27.2)	.065 (1719)
Nasals	9.2 (28.9)	7.0 (24.6)	.06 (1561)
Stops	16.2 (49.6)	10.0 (25.6)	.27 (7163)
Fricatives	12.0 (33.5)	7.0 (21.4)	.2 (5260)
Affricate	8.9 (22.9)	8.4 (17.0)	.013 (1719)
Utterance Initial	10.5 (28.4)	8.8 (20.7)	.125 (3287)
Utterance Medial	11.1 (37.1)	6.5 (23.5)	.75 (19715)
Utterance Final	7.1 (29.4)	4.1 (21.5)	.125 (3257)
Overall	10.5 (35.4)	6.5 (22.9)	1.0 (26259)

AH, UW, UH, OW, AO, AA, AW, AY, OY, ER), fricatives (F, V, TH, DH, S, Z, SH, HH), plosives (P, B, T, D, K, G), nasals (M, N, NG), approximants (the glides W, Y, and the liquids L, R), and affricates (CH, JH).

Linear mixed model. We fit the mixed-effects linear model in (4) (following lme4 [22] notation). Fixed effects for speaker age, percentage onset error, percentage offset error, phoneme position, and phoneme type were included. Both by-phoneme and by-speaker random intercepts were included, as well as by-phoneme random slopes for onset error and offset error.

$$\Delta PLLR \sim \text{Age} + E_{on} + E_{off} + \text{Type} + \text{Position} + (1 + E_{on} + E_{off} | \text{Phoneme}) + (1 | \text{Speaker}) \quad (4)$$

E_{on} , E_{off} and $\Delta PLLR$ were all standardized, and categorical variables were effect contrast coded. The coefficients for each category can be interpreted as effect sizes (Z scores) for a typical speaker and a “typical” phoneme.

Data Preprocessing. Individual phonemes for which $E_{on} < -100\%$ and $E_{off} > 100\%$ were removed from the dataset. This step ensured that some part of the ground truth interval was always contained within the forced aligned interval, avoiding ceiling effects in equation (1) and when predicting $\Delta PLLR$. Only 2.2% of the data was removed by this condition.

3. Results and Discussion

3.1. Alignment Error

Table 1 reports alignment error by phoneme category and by utterance position. Overall, average onset error was larger than offset error for every phoneme class and phoneme position. Alignment error for utterance-final phonemes was smaller than alignment error for utterance-initial and utterance-medial phonemes. We observed the largest alignment errors for stops (16.2% onset error), fricatives (12.0%), and approximants (11.4%). The higher alignment error for stops makes sense given the short duration of plosives. The higher-than-average alignment error for fricatives and approximants can be attributed to acoustic deviations, as children acquire phonemes in these classes at an older age [23].

3.2. Results of Statistical Analysis

The effect sizes for the terms of the linear model predicting $\Delta PLLR$ are shown in Figure 3. Onset error (E_{on}) and offset error (E_{off}) had the largest effects on $\Delta PLLR$. Phoneme position

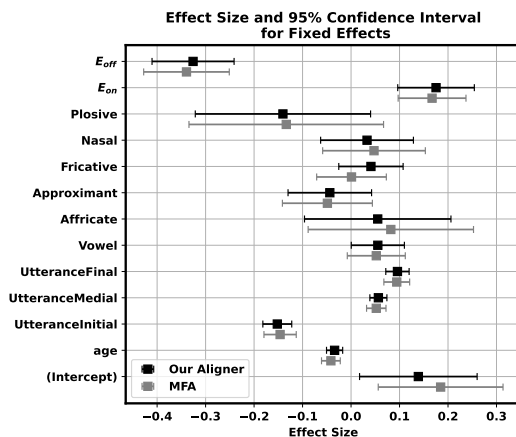


Figure 3: Effect size of fixed effects in linear model (4). $PLL R_a$ and $\Delta PLLR$ were calculated using our forced alignment system and used to fit model (4) to generate the effect sizes in black. This procedure was repeated with forced alignments from the MFA to generate the effect sizes shown in grey.

(utterance-final, utterance-medial, and utterance-initial) had a small but statistically significant effect, as did speaker age. For phoneme type, all of the 95% confidence intervals for the effect sizes included zero.

Upon recomputing alignments with the widely used Montreal Forced Aligner (MFA), $\Delta PLLR$ was recomputed and used to refit model (4). Figure 3 shows that the effect sizes are similar even when using a different alignment method.

Random Effects. Analysis of the random effects in the model are provided in Table 2. The significance of the random effects was established via a likelihood ratio test. While all of random effects were statistically significant, the comparatively small effect size of the between-speaker variance ($SD=0.045$) indicates that speaker identity has little impact on $\Delta PLLR$. Effect size was similar ($SD \approx 0.25$) for the remaining three random effects: random intercepts by phoneme, and random slopes by phoneme for E_{on} and E_{off} , suggesting moderate variability in $\Delta PLLR$ by phoneme type and for alignment error by phoneme type.

Accurately Modeling Alignment Error. We did not take the absolute value of alignment error in Equation (3). In fact, fitting (4) using the absolute values $|E_{on}|$ and $|E_{off}|$ reduces effect size magnitude more than sixfold. $\Delta PLLR$ may be affected differently whether the FA boundary is assigned before or after the manual boundary. Positive alignment error and negative alignment error are depicted in Figure 4, Case 1 and Case 2 respectively. Thus, accurate modeling of the impact of alignment error on $\Delta PLLR$ requires accounting for the sign of E_{on} and E_{off} .

Additionally, the sign of the effect size of alignment error (Figure 3) reveals that alignment error can cause an unexpected

Table 2: Statistical analysis of model (4).

Random Effect	Group	SD	χ^2	$Pr(> \chi^2)$
Intercept	Speaker	0.045	28.2	< .0001
Intercept	Phoneme	0.25	3534	< .0001
E_{on} Slope	Phoneme	0.231	1464	< .0001
E_{off} Slope	Phoneme	0.251	779	< .0001

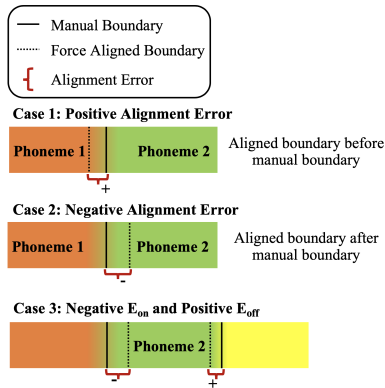


Figure 4: Alignment error visualized. Regions with color gradient correspond to coarticulation between phonemes.

increase in automatic PLLR. The increase happens when onset error is negative (manual onset before FA onset) or when offset error is positive (FA offset before manual offset). These scenarios are depicted in Case 3 of Figure 4. In these cases, the manual onset boundary is closer to the preceding phoneme (negative onset) and the manual offset boundary is closer to the subsequent phoneme (positive offset), so acoustic frames in the manual interval would include more coarticulation and hence lower PLLR-valued frames from adjacent phonemes. Put differently, if the automatic interval omits coarticulated frames, it has “cleaner” acoustics for PLLR-scoring. Follow-on analysis with error-by-utterance position interactions found that onset/offset error did not improve PLLR scores for utterance-initial and -final phones where there is not coarticulation effects.

3.3. Implications for Practical Use of $PLL R_a$

Our preceding statistical analysis level examined differences in PLLR scores of individual phonemes. That level of analysis provides useful, fine-grained information about measurement error. But in practice, a speaker will not be evaluated on a single token’s (a specific phoneme occurrence) score but rather on the speaker’s average level of performance (and perhaps the variance of productions). Moreover, a speaker’s performance will be assessed relative to a distribution of other speakers, so the actual PLLR value and its units are not as important as a standardized score or percentile rating. So, it is important to examine the consistency of utterance-averaged per-speaker PLLR scores. These would assess whether discrepancies in PLLR scores at the token level propagate through to intended use cases.

Figure 5 shows a strong correspondence between the two sets of standardized PLLR scores. In this case, phonemes scores were averaged for each speaker, and then for each phoneme, the speaker means were standardized. The standardized $PLL R_a$ and $PLL R_m$ for each phoneme were highly correlated with one another, and the scores in each phoneme class followed similar distributions. The correlation values and density curves provide evidence for the reliability of these PLLR scores, but further work is needed to validate speech assessment via PLLR against traditional assessments. The aggregated comparison is also repeated each utterance position. PLLR scores were averaged per-speaker, across-phonemes, and per-utterance position to generate the bottom row of plots in Figure 5. High similarity between automatic and manual PLLR distributions is again observed.

Our study demonstrates that forced alignment errors, phoneme position, and phoneme type have moderate effects when calculating PLLR scores. Other studies have also found that FA errors do not impact phonetic analysis in child speech

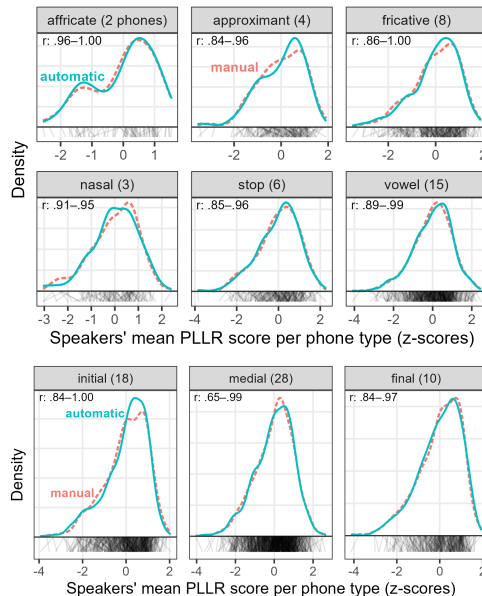


Figure 5: Density of standardized speaker-average PLLR scores by phoneme class and then by phoneme position. For each child and phoneme, we computed the child average PLLR score and then z-score standardized the scores within each phoneme (or by position). We plotted the distribution of these scores by phoneme class. The range of Pearson correlations for phonemes in each category is reported in each subplot corner. Line segments below subplots represent the differences in mean PLLR scores for each child (manual on top and automatic on bottom) and visualize how each child contributes to the density.

when extracting metrics such as center of gravity [24].

Limitations. The relationship between alignment error and $\Delta PLLR$ may not be linear and a nonlinear mixed model may be more appropriate, however those models typically have increased analytic flexibility and we did not have the sample size to support their use herein. Secondly, the analysis was repeated for only two alignment methods, our own method adapted from [18] and the MFA [13]. It is unclear if relative magnitude of effects in predicting $\Delta PLLR$ would be similar for the aligners not assayed in this study. Additionally, other variants of the GOP were not considered in this study (e.g., GOP variants that only use the numerator of the PLLR in (1), directly using the phoneme posterior or likelihood rather than using a ratio). We also expect the influence of alignment error on these GOP variants to be modest.

4. Conclusion

We demonstrate that differences in pronunciation scores calculated with manual alignments and forced alignments are not substantially impacted by phoneme position in utterance or by phoneme class. Evaluating our corpus collected from typically developing child speech, alignment error had a moderate effect on the difference between the manual and automatic scores, $\Delta PLLR$. Despite this moderate effect, the distributions of standardized $PLL R_a$ and standardized $PLL R_m$ were nearly identical across all phoneme classes after utterance-wise averaging. Our results justify the use of PLLR computed using forced alignment to measure pronunciation across phoneme types. Future work will focus on further validating $PLL R_a$ against traditional pronunciation assessments from human raters.

5. References

- [1] Y. Kheir, A. Ali, and S. Chowdhury, "Automatic pronunciation assessment - a review," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8304–8324. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.557>
- [2] G. M. Stegmann, S. Hahn, J. Liss, J. Shefner, S. Rutkove, K. Shelton, C. J. Duncan, and V. Berisha, "Early detection and tracking of bulbar changes in ALS via frequent and remote speech analysis," *NPJ digital medicine*, vol. 3, no. 1, p. 132, 2020.
- [3] L. Fontan, T. Pellegrini, J. Olcoz, and A. Abad, "Predicting disordered speech comprehensibility from goodness of pronunciation scores," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*. Dresden, Germany: Association for Computational Linguistics, Sep. 2015, pp. 42–46. [Online]. Available: <https://aclanthology.org/W15-5108>
- [4] B. Lin and L. Wang, "Deep feature transfer learning for automatic pronunciation assessment," in *Interspeech*, vol. 2021, 2021, pp. 4438–4442.
- [5] M. Tu, A. Grabek, J. Liss, and V. Berisha, "Investigating the role of I1 in automatic pronunciation evaluation of I2 speech," *arXiv preprint arXiv:1807.01738*, 2018.
- [6] J. Vidal, C. Bonomi, M. Sancinetti, and L. Ferrer, "Phone-level pronunciation scoring for Spanish speakers learning English using a gop-dnn system," in *Interspeech*, 2021, pp. 4423–4427.
- [7] J. Proença, C. Lopes, M. Tjalve, A. Stolcke, S. Candeias, and F. Perdigão, "Detection of mispronunciations and disfluencies in children reading aloud," in *Interspeech*, 2017, pp. 1437–1441.
- [8] V. C. Mathad, T. J. Mahr, N. Scherer, K. Chapman, K. C. Hustad, J. Liss, and V. Berisha, "The impact of forced-alignment errors on automatic pronunciation evaluation," in *Interspeech*, 2021, pp. 1922–1926.
- [9] J. A. Gierut, "Treatment efficacy: Functional phonological disorders in children," *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 1, pp. S85–S100, 1998.
- [10] 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.
- [11] S. Witt and S. Young, "Computer-assisted pronunciation teaching based on automatic speech recognition," in *Language teaching and language technology*. Routledge, 2014, pp. 25–35.
- [12] K. Gorman, J. Howell, and M. Wagner, "Prosodylab-aligner: A tool for forced alignment of laboratory speech," *Canadian Acoustics*, vol. 39, no. 3, pp. 192–193, 2011.
- [13] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldii," in *Interspeech*, vol. 2017, 2017, pp. 498–502.
- [14] G. Yeung and A. Alwan, "On the difficulties of automatic speech recognition for kindergarten-aged children," *Interspeech 2018*, 2018.
- [15] T. J. Mahr, V. Berisha, K. Kawabata, J. Liss, and K. C. Hustad, "Performance of forced-alignment algorithms on children's speech," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 6S, pp. 2213–2222, 2021.
- [16] M. Hodge, J. Daniels, and C. Gotzke, "Tocs+ intelligibility measures," *Edmonton, AB: University of Alberta*, 2007.
- [17] P. Boersma, "Praat: doing phonetics by computer [computer program]," <http://www.praat.org/>, 2011.
- [18] J. Zhu, C. Zhang, and D. Jurgens, "Phone-to-audio alignment without text: A semi-supervised approach," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8167–8171.
- [19] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.
- [20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldii speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.
- [21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5206–5210. [Online]. Available: <http://ieeexplore.ieee.org/document/7178964/>
- [22] D. M. Bates, "lme4: Mixed-effects modeling with R," 2010.
- [23] K. Crowe and S. McLeod, "Children's English consonant acquisition in the United States: A review," *American Journal of Speech-Language Pathology*, vol. 29, no. 4, pp. 2155–2169, 2020.
- [24] T. Knowles, M. Clayards, and M. Sonderegger, "Examining factors influencing the viability of automatic acoustic analysis of child speech," *Journal of Speech, Language, and Hearing Research*, vol. 61, no. 10, pp. 2487–2501, 2018.