INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

The Impact of Forced-Alignment Errors on Automatic Pronunciation Evaluation

Vikram C. Mathad[1], *Tristan J. Mahr*[2], *Nancy Scherer*[1], *Kathy Chapman*[3], *Katherine C. Hustad*[2], *Julie Liss*[1], *Visar Berisha*[1,4]

[1]College of Health Solutions, Arizona State University, USA
[2]Department of Communication Sciences and Disorders, University of Wisconsin–Madison, USA
[3]Department of Communication Sciences and Disorders, University of Utah, USA
[4]School of Electrical, Computer, & Energy Engineering, Arizona State University, USA
{vchikkaw, nancy.scherer, julie.liss, visar}@asu.edu,
{tristan.mahr, katie.hustad}@wisc.edu, kathy.chapman@health.utah.edu

Abstract

Automatic evaluation of phone-level pronunciation scores typically involves two stages: (1) automatic phonetic segmentation via text-constrained phoneme alignment and (2) quantification of acoustic deviation for each phoneme-level relative to a database of correctly-pronounced speech. It's clear that the second stage depends on the first. That is, if there is misalignment, the acoustic deviation will also be impacted. In this paper, we analyzed the impact of alignment error on a measure of goodness of pronunciation. We computed (1) automatic pronunciation scores using force-aligned samples, (2) the forced-alignment error rate, and (3) acoustic deviation using manually-aligned samples. We used a bivariate linear regression model to characterize the contributions of forced alignment errors and acoustic deviation on the automatic pronunciation scores. This was done across two different children speech databases, namely children with cleft lip/palate and typically developing children between the ages of 3-6 years. The analysis shows that, for speech from typically-developing children, most of the variation in the automatic pronunciation scores is explained by acoustic deviation, with the errors in forced alignment playing a relatively minor role. The forced alignment errors have a small but significant downstream impact on pronunciation assessment for children with cleft lip/palate.

Index Terms: alignment error, force alignments, goodness of pronunciation, manual alignments.

1. Introduction

Pronunciation errors are the result of a deviation in speech acoustics relative to typical speech production. They are present in second (L_2) language learners and speech sound disorders caused due to neurological disorders, structural anomalies (e.g. cleft lip and palate), hearing impairment, etc. Automatic phone-level pronunciation scoring systems evaluate the acoustic deviation in a spoken utterance relative to a corpus of typical speech on a phoneme-by-phoneme basis. This is useful for several downstream applications. For example, computerized automatic pronunciation training (CAPT) systems have been developed to improve pronunciation errors in the second language learning [1]. Clinically they can be used to evaluate speech sound disorders such as cleft lip and palate (CLP) [2], dysarthria [3], unilateral cerebral palsy [4], and childhood apraxia [5].

One of the most common systems for pronunciation evaluation is the goodness of pronunciation (GOP) algorithm [1, 6, 4]. This score is computed using phoneme-level posterior probabilities estimated using Gaussian mixture model-hidden Markov model (GMM-HMM) or deep neural network-hidden Markov model (DNN-HMM)-based acoustic model. While computing the GOP, the target phoneme boundaries are located using a forced-alignment algorithm. Other pronunciation assessment systems similarly rely on forced alignment. For example, classifiers, such as support vector machines or deep neural networks, trained to classify between correctly and incorrectly-pronounced speech have been proposed in the context of pronunciation assessment [2, 5]. These systems rely on an initial segmentation at the phoneme level, followed by classification. For both GOP-type algorithms and classification-based algorithms, forced-alignment errors can have downstream consequences. As an example, consider a correctly-pronounced utterance that has been incorrectly segmented. In this case, both types of algorithms will incorrectly assess the speech as having low articulatory precision due to phonetic mismatch.

Forced-alignment is the most commonly used approach to phonetic segmentation. In forced-alignment, speech and the corresponding orthographic transcription are passed through a pre-trained acoustic model, trained using a large amount of representative speech, typically from publicly available databases [7, 8]. These models are susceptible to errors when used on atypical speech due to acoustic mismatch between the utterance to be aligned and the acoustic model's training set. Studies have analyzed these alignment errors under different conditions, for example with children's speech [9], with speech from different languages [10], or with dysarthric speech [11, 6]. These studies show heterogeneity in performance across different age ranges in children and severity levels in dysarthria.

In light of this existing body of work, it stands to reason that forced-alignment errors likely impact automatic pronunciation when used in these populations. Since the alignment error results in reduced pronunciation scores, it becomes difficult to interpret whether the reduced pronunciation score is due to alignment error, acoustic deviation, or a combination of both. Hence, there is a need to analyze the effect of alignment errors on these scores.

In this study, we analyze the impact of alignment error on

http://dx.doi.org/10.21437/Interspeech.2021-1403

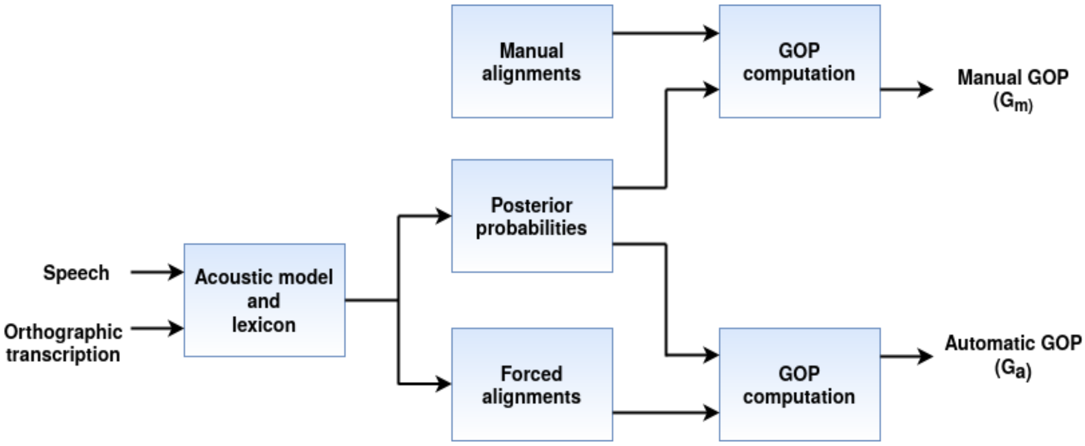

Figure 1: *An overview of the experimental design implementation.*

pronunciation scores. We focus our analysis on the goodness GOP algorithm in [1]. First, we computed the GOP scores using manual alignments. The manual GOP score is free from alignment errors and results only from acoustic deviation. Next, we computed the automatic GOP score using first forced-alignment then assessment of acoustic deviation. Simultaneously, we estimated the forced-alignment error rate relative to the manually-aligned speech. We used a bivariate linear model to study the relationship between acoustic deviation, alignment error, and automatic GOP. This was done for two different databases, namely, typically-developing children ages 3-6 and children with cleft lip/palate.

2. Methods

2.1. Speech corpora

In this work we used two different speech databases, each of which had manual phone level alignments. These are described below.

Typically-developing children's (TDC) speech database: We used the children's speech database originally used in [9]. This database contains word and sentence-level recordings from 42 (21 boys and 21 girls) typically developing children between the ages of 3 and 6. The children are native American English speakers and have normal speech and hearing characteristics. The manual alignments were hand-corrected after initial forced-alignment by the prosody lab aligner [12]. This was done by listening to each speech sample and performing manual boundary adjustments on Praat textgrids [13].

Cleft palate (CP) speech database: The CP database contains speech samples from 60 children with cleft palate recorded at the University of Utah. All participants are between ages 5 and 7 years old and each child provided 24 English sentences, recorded according to the Americleft protocol [14]. Speakers with CP produce obligatory (weak and nasalized consonants) and compensatory errors (glottal, pharyngeal, palatalized, and nasal fricative sounds) [15]. Manual alignments were obtained via manual correction of forced alignments from the Montreal aligner [16] by Author-1 of this paper.

2.2. Algorithm Implementation Details

The overall implementation of the approach to analyze the effect of alignment errors on automatic GOP (G_a) is shown in Fig. 1. The speech samples and the corresponding orthographic transcriptions are passed through a DNN-HMM acoustic model. We used the pre-trained DNN-HMM acoustic model [1] trained on 1000 hours of the Librispeech database and uses the Librispeech lexicon[7]. The model uses i-vector based speaker adaptation. The forced-alignments and phoneme posteriors computed from the DNN-HMM are used to calculate G_a We used the Kaldi recipe [2] to calculate G_a.

As shown in Fig. 1, the manual GOP (G_m) is computed by using hand-corrected manual alignments and the phoneme posteriors are obtained. The manual alignments are generated by using the hand-corrected Praat-text grids. Note that we used the same set of posteriors but different sets of phone alignments when computing the G_m and G_a

The alignment error is computed for each phoneme by comparing the manual and automatic alignments. For the manual alignment case, let $t^{(i)}_{m_1}$ and $t^{(i)}_{m_2}$ be the onset and offset instants for the i^{th} phoneme, respectively. Similarly, let $t^{(i)}_{a_1}$ and $t^{(i)}_{a_2}$ be the onset and offset instants for automatic alignments for i^{th} phoneme. The onset alignment error for i^{th} phoneme ($e^{(i)}_{on}$) is computed as

$$e^{(i)}_{on} = |t^{(i)}_{a_1} - t^{(i)}_{m_1}| \tag{1}$$

The offset alignment error for i^{th} phoneme ($e^{(i)}_{off}$) is computed as

$$e^{(i)}_{off} = |t^{(i)}_{a_2} - t^{(i)}_{m_2}| \tag{2}$$

The average alignment error ($t^{(i)}_{e}$) is computed as the mean of onset and offset errors.

$$t^{(i)}_{e} = \frac{e^{(i)}_{on} + e^{(i)}_{off}}{2} \tag{3}$$

The automatic GOP, manual GOP, and the alignment error are computed for each phoneme instance. For each speaker, we grouped and averaged across broad phoneme categories: vowels, approximants, nasals, plosives, and fricatives. In addition, we computed the average across all phonemes. Using these speaker-level scores, we analyzed the dependency of G_a scores on G_m and alignment error.

2.3. Statistical Modeling

The graphical model that underlies the statistical analysis models we develop is shown in Fig 2. Imprecise articulation impacts

[1] http://kaldi-asr.org/models/m13

[2] https://github.com/kaldi-asr/kaldi/tree/master/egs/gop_speechocean762/s5

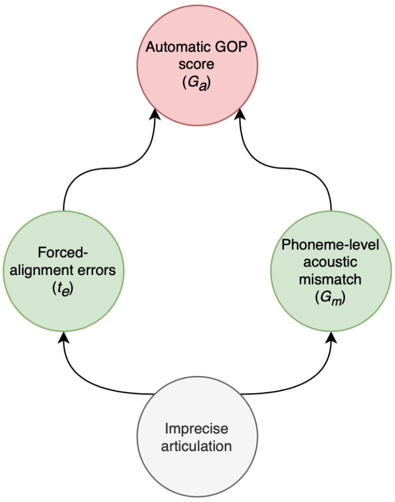
Automatic GOP score (G_a)
Forced-alignment errors (t_e)
Phoneme-level acoustic mismatch (G_m)
Imprecise articulation

Figure 2: *Imprecise articulation impacts both forced alignment errors and phoneme-level acoustic mismatch. These variables, in turn, impact the automatic GOP calculation.*

both forced alignment errors (t_e) and phoneme-level acoustic mismatch, as measured by manually-aligned GOP (G_m). These variables, in turn, impact the automatic GOP calculation G_a. To evaluate the impact of alignment errors and acoustic mismatch on automatic GOP, both individually and combined, we fit three models to the data from each database. Model 1 assesses the relationship between alignment errors and automatic GOP. Model 2 evaluates the relationship between acoustic mismatch and GOP. Model 3, evaluates the relationship between both variables and GOP.

$$\text{Model } 1: \quad G_a \sim t_e \tag{4}$$
$$\text{Model } 2: \quad G_a \sim G_m \tag{5}$$
$$\text{Model } 3: \quad G_a \sim t_e + G_m \tag{6}$$

3. Results

Herein we present the results of our analysis. We first provide an overview of the alignment error statistics for each of the two datasets; then we present our linear regression model results that assess the impact of alignment errors and acoustic mismatch on the automatic goodness of pronunciation measure.

Table 1: *Mean and standard deviation of alignment errors by phoneme category for each dataset (t_e)*

Category	TDC		CP	
	Count	mean±dev. (ms)	Count	mean±dev. (ms)
Vowels	138	35.396±36.966	222	54.739±99.108
Approximants	17	30.786±33.101	56	61.881±108.186
Nasals	22	30.787±32.273	30	59.646±111.066
Stops	118	36.769±39.615	75	42.168±88.602
Fricatives	61	31.562±48.616	70	48.339±93.973
All phonemes	356	34.394±39.685	453	53.416±99.480

3.1. Alignment error statistics

The mean and standard deviations of the average alignment error, t_e, are shown in Table 1. The table shows that, across all phoneme categories, the alignment error is larger for the CP database than the TDC group. These results are consistent with expectations as the TDC database contains typically-developing children, with fewer articulation errors relative to the clinical group.

Table 2: *Percent of variance explained in automatic GOP by Models 1, 2, and 3 for different phoneme categories across three different databases.*

		Databases	
		TDC	CP
Vowels	Model 1	0.4687	0.699
	Model 2	0.9709	0.9312
	Model 3	0.9709	0.9368
Approximants	Model 1	0.1235	0.3674
	Model 2	0.9637	0.9111
	Model 3	0.9646	0.9128
Nasals	Model 1	0.4944	0.1764
	Model 2	0.8769	0.7281
	Model 3	0.8824	0.7449
Stops	Model 1	0.36	0.4729
	Model 2	0.9534	0.9412
	Model 3	0.9576	0.9413
Fricatives	Model 1	0.2943	0.4133
	Model 2	0.9667	0.9594
	Model 3	0.9726	0.9596
All phonemes	Model 1	0.4713	0.5794
	Model 2	0.9867	0.9577
	Model 3	0.9879	0.9590

3.2. Statistical analysis results

The goodness-of-fit for Models 1, 2, and 3 was measured using model R^2. The R^2 values for different phoneme categories and databases are shown in Table 2. Evaluated individually in Models 1 and 2, considerably more of the variance in G_a is explained by acoustic mismatch than alignment error. Interestingly, across the board, the marginal increase in R^2 from adding alignment error is incremental over the univariate model that only includes the acoustic mismatch variable.

As shown in Table 2, the acoustic deviation is the main driver in the estimation of the automatic GOP, with alignment error explaining only a negligible amount of variance in most cases. The exception is only in CP children, where the combination of t_e and G_m explains more of the variance for the vowel and nasal phoneme categories.

In addition to the R^2, we also analyzed the relative contributions of t_e and G_m to G_a in Model 3. The complete statistical results for Model 3 for different phoneme categories and databases are shown in Table 3. For the TDC case, the forced-alignment error has a significant impact only for stops and fricatives. It's important to note that although this contribution is significant, its effect size (as measured by the t-statistic) is an order of magnitude lower than it is for the acoustic mismatch variable i.e., G_m. This indicates that the automatic GOP assessment is relatively robust to alignment errors. For example, in our analysis, the alignment errors on the order of 35ms (from the TDC corpus) have only a small impact on GOP and, for most phoneme categories, do not manifest as statistically significant contributions.

In the CP case, the addition of forced alignment errors explains more of the variability in G_a for nasals and vowels. For nasals, it is interesting to note that the G_m explains only 72.81% of the variance in G_a. We originally hypothesized that this difference between G_a and G_m is caused due to the presence of

Table 3: *Statistical analysis of Model 3.*

Category	Variable	TDC				CP			
		coeff	std	t	p	coeff	std	t	p
Vowels	t_e	-0.0001	0.0004	-0.2229	0.8248	-0.0002	0.0001	-2.2507	p<0.05
	G_m	1.0961	0.0422	25.9451	p<0.001	0.9617	0.0657	14.6356	p<0.001
Approximants	t_e	-0.0003	0.0003	-0.9812	0.3325	0.0002	0.0002	1.0512	0.2976
	G_m	0.9878	0.0325	30.4362	p<0.001	0.9271	0.0491	18.8785	p<0.001
Nasals	t_e	-0.0016	0.0012	-1.3519	0.1842	0.0006	0.0003	1.9397	p<0.05
	G_m	1.0372	0.0914	11.3439	p<0.001	0.9191	0.0816	11.2706	p<0.001
Stops	t_e	-0.001	0.0005	-1.9636	p<0.05	0.0001	0.0002	0.239	0.8119
	G_m	1.0169	0.0434	23.4347	p<0.001	0.9886	0.0464	21.3278	p<0.001
Fricatives	t_e	-0.0011	0.0004	-2.8945	p<0.05	-0.0001	0.0003	-0.4651	0.6437
	G_m	1.0616	0.0342	31.0808	p<0.001	1.0197	0.0367	27.7601	p<0.001
All phonemes	t_e	-0.0006	0.0003	-1.923	0.0618	0.0002	0.0002	1.3583	0.1797
	G_m	1.1075	0.0272	40.7346	p<0.001	1.0871	0.0473	22.9867	p<0.001

alignment error. However, the addition of forced alignment error (Model 3) does not improve the R^2 value. Even when evaluated independently (Model 1), the variable t_e only explains a small amount of the variability in G_a ($R^2 = 0.1764$). This could be due to inherent variability in the data as the sample size of nasal phonemes is smaller when compared to other phoneme categories (Table 1). The presence of CP impacts the production of high-pressure consonants, i.e., stops and fricatives which require build-up of oral air pressure; however, this is more difficult if the air is escaping through the nasal cavity [15]. The automatic GOP scores for phonemes that are correctly produced are likely to be largely explained by poor alignment, as the acoustic deviation would be low. That is, in presence of normal articulation and if an alignment error exists, then the alignment is responsible for the reduced GOP. However, the situation is more complicated if the phonemes are mis-articulated. The GOP scores are low for pressure consonants due to their deviation from normal production. These results indicate that the alignment errors do not greatly impact G_a and the change in G_a is driven by acoustic deviation. In contrast to high-pressure consonants, the production mechanism for nasal consonants is less impacted in the presence of CP [15]. Since the nasal consonants are produced with normal characteristics, lower G_a values are more likely to be caused by alignment error. Similarly, the vowels are also not as distorted as high-pressure consonants in CP speakers (although they are nasalized). Hence, the vowels also showed some sensitivity to alignment error.

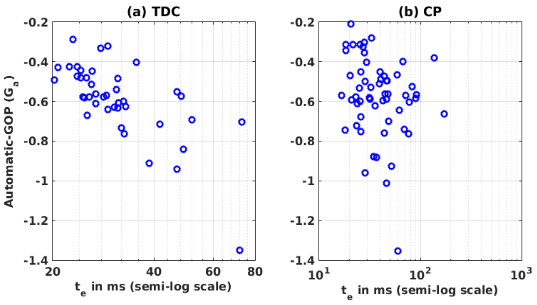

Figure 3: *Scatter plot between G_a and t_e for (a) TDC and (b) CP cases for nasals.*

The most likely cause of mismatch between the G_a and G_m is the presence of alignment errors. But the alignment error does not always show a significant impact on G_a. In our models, we assumed a linear relationship between the alignment error and the GOP. However, this assumption may not be valid in all cases. Fig. 3 (a) and (b) shows the scatter plot between t_e and G_a for the TDC and CP cases, respectively, for the nasal consonant category. The scatter plots shows a non-linear relationship between t_e and G_a. This is especially evident in the CP case where the logarithmic relationship is more apparent. The existence of a non-linear relationship between alignment error and GOP scores for some of the phoneme categories may be another reason that the forced alignment error did not always improve the fit in Model 3.

4. Summary and Conclusion

The present work analyzes the impact of forced-alignment errors on automatic pronunciation assessment algorithms, namely the GOP. We developed linear models to analyze the impact of alignment error, acoustic-deviation from normal, and their combination on GOP scores. The analysis was carried on TDC and CP children's speech corpora. The alignment error was found to be higher for the CP case than the TDC group due to the presence of articulation errors. Most of the variability in G_a was explained by acoustic deviation, i.e., G_m for the TDC case. Surprisingly, forced alignment error had a negligible impact in most cases on automatic pronunciation scores for the TDC corpus. In the case of CP, there was a significant impact of alignment error on nasals and vowels. However, the alignment error did not explain all the variability in G_a; this is likely due to the assumption of a linear relationship in our model. Future work will focus on modeling the non-linear relationship between alignment error and GOP. Future work will also explore how the data used to train the acoustic models impacts automatic pronunciation scores.

5. References

[1] S. Witt and S. Young, "Computer-assisted pronunciation teaching based on automatic speech recognition," *Language Teaching and Language Technology Groningen, The Netherlands*, 1997.

[2] C. Vikram, N. Adiga, and S. M. Prasanna, "Detection of nasalized voiced stops in cleft palate speech using epoch-synchronous features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1189–1200, 2019.

[3] L. Fontan, T. Pellegrini, J. Olcoz, and A. Abad, "Predicting disordered speech comprehensibility from goodness of pronunciation scores," in *Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2015) satellite workshop of Interspeech 2015*, 2015, pp. pp–1.

[4] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, and M. Robert, "The goodness of pronunciation algorithm applied to disordered speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[5] M. A. Shahin, B. Ahmed, J. X. Ji, and K. J. Ballard, "Anomaly detection approach for pronunciation verification of disordered speech using speech attribute features." in *INTERSPEECH*, 2018, pp. 1671–1675.

[6] I. Laaridh, C. Fredouille, and C. Meunier, "Evaluation of a phone-based anomaly detection approach for dysarthric speech," in *Interspeech 2016*, 2016, pp. 223–227.

[7] 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)*. IEEE, 2015, pp. 5206–5210.

[8] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990.

[9] 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*, pp. 1–10, 2021.

[10] L. MacKenzie and D. Turton, "Assessing the accuracy of existing forced alignment software on varieties of british english," *Linguistics Vanguard*, vol. 6, no. s1, 2020.

[11] I. Laaridh, C. Fredouille, and C. Meunier, "Automatic speech processing for dysarthria: A study of inter-pathology variability."

[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] P. Boersma, "Praat: doing phonetics by computer," *http://www.praat. org/*, 2006.

[14] K. L. Chapman, A. Baylis, J. Trost-Cardamone, K. N. Cordero, A. Dixon, C. Dobbelsteyn, A. Thurmes, K. Wilson, A. Harding-Bell, T. Sweeney *et al.*, "The americleft speech project: a training and reliability study," *The Cleft Palate-Craniofacial Journal*, vol. 53, no. 1, pp. 93–108, 2016.

[15] G. Henningsson, D. P. Kuehn, D. Sell, T. Sweeney, J. E. Trost-Cardamone, and T. L. Whitehill, "Universal parameters for reporting speech outcomes in individuals with cleft palate," *The Cleft Palate-Craniofacial Journal*, vol. 45, no. 1, pp. 1–17, 2008.

[16] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.