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

Phone-Level Pronunciation Scoring for Spanish Speakers Learning English Using a GOP-DNN System

Jazmín Vidal[1,2], *Cyntia Bonomi*[1], *Marcelo Sancinetti*[1], *Luciana Ferrer*[2]

[1]Departamento de Computación, FCEyN, Universidad de Buenos Aires (UBA), Argentina
[2]Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina
{jvidal,gbonomi,msancinetti,lferrer}@dc.uba.ar

Abstract

In today's globalized world being able to communicate in English is crucial to many people. Computer assisted pronunciation training (CAPT) systems can help students achieve English proficiency by providing an accessible way to practice, offering personalized feedback. However, phone-level pronunciation scoring is still a very challenging task, with performance far from that of human annotators. In this paper we compare and present results on the Spanish subset of the L2-ARCTIC corpus and the new Epa-DB database, both containing non-native English speech by native Spanish speakers and intended for the development of pronunciation scoring systems. We show the most frequent errors in each database and compare performance of a state-of-the-art goodness of pronunciation (GOP) system. Results show that both databases have similar error patterns and that performance is similar for most phones, despite differences in recording conditions. For the EpaDB database we also present an analysis of the errors per target phone. This study validates the EpaDB collection and annotations, providing initial results and contributing to the advancement of a challenging low-resource task.

Index Terms: computer assisted language learning, phone-level pronunciation scoring, goodness of pronunciation, speech corpora, low-resources

1. Introduction

Computer-assisted language learning (CALL) programs have been available for decades to help students learn and practice a new language. Most CALL programs focus on improving grammar skills and vocabulary. A few also teach oral skills, giving the student feedback about the quality of their pronunciation at different levels. Many systems have been proposed in the last decades that produce pronunciation scores for each paragraph, phrase, word or phone pronounced by the student [1, 2, 3, 4]. Some of them reach performance levels that are comparable to the agreement across humans when scores are computed over long chunks of speech [1, 5]. Yet, word- and phone-level scoring are still challenging tasks with much lower level of performance; and experts still claim for the need of more accurate solutions [6].

The ultimate goal of our project is to develop a pronunciation training system for Spanish speakers learning English. To this end, in a previous work [7], we introduced Epa-DB, a database of 3200 English short utterances produced by 50 Spanish speakers from Argentina annotated at a detailed phonetic level. In this paper we present a comparison between Epa-DB and the Spanish subset of the L2-Arctic corpus [8], which is, to our knowledge, the only other database of Spanish speakers speaking English labelled for pronunciation scoring.

First, we briefly analyze the most frequent error patterns in both databases to establish a common ground of mistakes for Spanish learners of English in the two databases, confirming that they mostly overlap and coincide with the expected problematic phones for this target population reported in the literature [9]. Then, we evaluate performance on both databases using a freely available state-of-the art DNN-HMM goodness of pronunciation (GOP) system based on [10]. Results show that the system performance is comparable across databases on most phones. Finally, our results show that, for many phones, the performance of the GOP system is unacceptably poor for practical applications, specially for this task where it is important to avoid frustrating the student with frequent incorrect interventions [11]. We show the distribution of scores for the different variants of each target phone, showing that the system gets confused mostly on certain variants, while others are correctly distinguished. The hard variants, however, correspond to salient errors in pronunciation which we wish to correct, exposing a major limitation in the GOP algorithm.

The results and analysis in this paper validate the Epa-DB collection and provide an initial baseline for future developments on this data, which is publicly and freely available for research use.

2. Non-native Datasets

In this section we briefly present L2-ARCTIC and EpaDB databases and show an analysis of the incorrect variants used for different target phones across the two databases.

2.1. Epa-DB dataset

Epa-DB [7] is a database of 3200 English short utterances produced by 50 Spanish speakers from Argentina annotated at a detailed phonetic level. Each speaker recorded 64 short English phrases phonetically balanced and designed to contain at least one example of every phone difficult to pronounce by the target population [9]. It also includes manually assigned phoneme boundaries for every phone along with an overall score of the perceived non-nativeness of each utterance according to the annotators. Annotations are in ARPAbet using an ARPAbet-like extension to account for English allophones and sounds not present in the English language. A deviation "*" symbol is used to annotate sounds that are difficult to classify. More information on the ARPA-bet extension can be found in the Epa-DB documentation.

Further, Epa-DB provides a set of possible reference transcriptions for each of the 64 phrases, corresponding to the possible ways a native speaker could pronounce each phrase. The longest reference transcription for a phrase (i.e., that which does not include possible phone-deletions), was used as the canoni-

  http://dx.doi.org/10.21437/Interspeech.2021-745

cal pronunciation with respect to which the manual annotations were made, marking deletions as "0" and additions as being substitutions of one of the canonical phones to two concatenated phones.

The speech data was recorded on the personal computers of each participant through an online application in order to mimic the envisioned use scenario where users will be practicing their pronunciation at home with their own computers. Annotations were made by three annotators, a Spanish-native linguist, a Spanish-native English professor, and an English-native English professor. To date, the last two annotators have only labelled part of the database. For this reason, in this work, only the annotations from the first annotator are used.

A link to download Epa-DB along with the code to run the experiments in this paper is available at `https://github.com/JazminVidal/gop-dnn-epadb`.

2.2. L2-ARCTIC dataset

L2-ARCTIC [8] is a speech corpus of non-native English intended for research in voice conversion, accent conversion, and mispronunciation detection. The speech was recorded in a controlled scenario, under quiet conditions using quality microphones. Recordings contain approximately 27.1 hours of read speech from the Carnegie Mellon University ARCTIC prompts [12]. In our experiments we use only the Spanish L1 subset of this dataset, which includes 600 manually-annotated utterances from 4 speakers.

The dataset provides orthographic and forced-aligned phonemic transcriptions and 150 manually-annotated utterances per speaker that identify three types of mispronunciation errors: substitutions, deletions, and additions. Annotations are placed in a tier that contains, for the correctly pronounced phones, the forced-alignment label in ARPAbet symbols, and for the incorrectly pronounced ones a three-part label, "CPL,PPL,e", where "CPL" is the correct phoneme label (i.e., what should have been produced), "PPL" is the perceived phoneme label (i.e., what was actually produced), and "e" stands for the corresponding type of error, namely "s" for substitutions and "d" for deletions. Additions are labeled as "sil,PPL,a", where "sil" stands for silence and "a" for addition. In this tier, a deviation ("*") symbol is used to mark "PPL" phones with non-native pronunciation. For those phones, additional information specifying the nature of the error is provided in a different tier using IPA symbols at the allophonic level. Note that, whereas Epa-DB annotates different allophones of the same phoneme separately, L2-ARCTIC only provides allophonic information in those cases where phonemes are incorrectly pronounced.

2.3. Frequent Error Patterns

In this section, we compare the most frequent errors in both databases to establish a common ground of mistakes for non-native Spanish speakers of English and to compare the criteria used for annotation. For the comparison, we map the phones in the IPA annotation tier of the L2-ARCTIC to the set of ARPAbet and ARPAbet like extension symbols designed for Epa-DB.

Figure 1 shows, for each target phone in the English language, the percent of each manually labeled error normalized with respect to the total number of errors for that phone for each database. The last pair of bars to the left represents the most frequent addition errors in L2-ARCTIC. These errors are shown in a separate group since, in this database, additions are not attributed to any specific target phone. Only target phones with at least 15 errors are included and bars for the errors with frequency below 10% are omitted. Percentage of incorrect cases for each target phone and database are shown under the x-axis.

We can see that, for most phones, frequent variants coincide across databases and with the expected problematic phones for this target population reported in the literature [9]. For a minority of phones, though, the percentage of incorrect cases differs significantly across the two databases. These differences can be explained mostly in three ways. First, the systematic appearance of specific erroneous variants in one database and not in the other, such as Sh ([h]) instead of S in Epa-DB and LL ([ʎ]) instead of Y in L2-ARCTIC could be indicating that Spanish speakers from different regions produce different error patterns. A closer look at those substitutions [13] suggests a dependency on the speakers dialect. Second, other differences are likely simply due to the fact that the Spanish subset of L2-ARCTIC includes only 4 speakers, with similar proficiency level, and who might have some particular error patterns that are not necessarily common across the population, like the F and T variants for the target P phone in words like *sympathy* and *up* in two of the speakers. Lastly, some of the discrepancies are explained by differences in labeling criteria. For example, as mentioned above, in L2-ARCTIC, additions are not assigned to a specific target phones and, hence, are listed separately in the figure, while for Epa-DB, additions are shown as a substitution of a target phone to two concatenated phones, e.g., S pronounced as E+S (/e+s/).

3. GOP-based Pronunciation Scoring

Pronunciation scoring systems can be divided into those that use only native data and those that require non-native data with pronunciation labels for training. Those of the first kind, usually rely on automatic speech recognition (ASR) systems trained with native speech. Pronunciation scores for a given test utterance are generated with respect to this model using different probabilistic measures: log-likelihood, log-posterior probabilities, and goodness of pronunciation scores [2, 14, 15]. These scores measure, in slightly different ways, the similarity between the student's speech and native-sounding speech, represented by the ASR model. The second family of systems is based on models trained with non-native data to distinguish correctly- from incorrectly-pronounced segments. They usually perform better than the ones described above, but require non-native training data annotated with pronunciation labels [16, 3, 17, 18, 19].

In this paper we implement a standard method of the first family, the GOP method [15]. This method relies on a first stage of alignment of the audio to its transcription to obtain the location of each phone. Then, a score for each phone in the alignments is computed as

$$GOP(p) = -\frac{1}{N} \log P(p|O^{(p)}) \qquad (1)$$

where $P(p(O^{(p)}|p)$ is the posterior probability of the target phone p given the sequence of features corresponding to that phone $O^{(p)}$ and N is the number of frames corresponding to that phone. Given a GOP score, the final decision is made by comparing it with a threshold, which is usually determined separately for each phone [16, 20, 21]. In the original work by Witt, the posterior was computed using the likelihoods obtained from a GMM-HMM ASR system and several assumptions and approximations to speed up calculation. In recent years, several papers have proposed to obtain these posteriors using a DNN-HMM ASR system [10, 21], where the posterior in (1) is approximated as a product of the posteriors given by the DNN

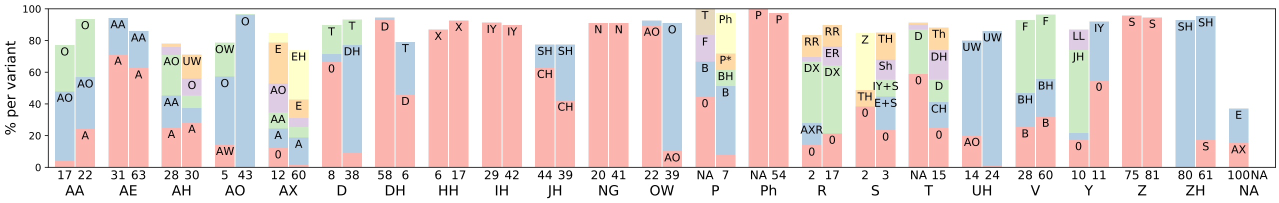

Figure 1: *Plot of the most common errors per target phone for Epa-DB (right bar in each group) and L2-ARCTIC (left bar) databases. Only target phones with more than 15 substitutions are shown. Zeros ("0") correspond to deletions and, phones concatenated by a "+" symbol to additions (in Epa-DB). The bar labeled "NA" show the addition errors in L2-ARCTIC. The number at the bottom of each bar corresponds to the percentage of incorrect instances for each target phone for each database. "NA" is used in cases where this percentage can not be estimated due to the lack of allophonic annotations for correct cases in L2-ARCTIC.*

over all the frames corresponding to the target phone. These papers show significant improvements in performance with respect to the GMM-based GOP scores and can be considered the state-of-the-art for methods of the first family. For this reason, we select this approach for the analyses in this paper.

4. Experimental Setup

For computing GOP, we use the official Kaldi [22] recipe along with the Kaldi Librispeech ASR model, a TDNN-F [23] acoustic model trained with 960 hours of native English speech from the LibriSpeech [24] collection. The acoustic features are 13-dimensional Mel frequency cepstral coefficients (MFCCs). I-vectors [25] are used to represent the speaker's characteristics. The alignments for the GOP computation are created with respect to the manual transcriptions using the pipeline provided by the toolkit with the same acoustic model mentioned above. For each phone in these alignments, the GOP system generates a score. In order to determine whether each of these phones corresponds to a correctly or incorrectly pronounced phone, we need to find a mapping between the phones in the alignments and those in the manual transcriptions. For Epa-DB, we use a custom algorithm that uses the reference transcriptions provided in the release of the database. For L2-Arctic, we use a modified version of the ALINE algorithm [26] adapted to work with ARPAbet representation of phones.

Results in this section are shown separately for each target phone in the Kaldi alignments. Note that allophones of the English language such as Ph ([p^h]), Th ([t^h]), Kh ([k^h]), are merged in Kaldi's phoneset with the unaspirated variants P, T, K. For this reasons, in these results aspirated and unaspirated variants of each plosive are included as a single target phone. Yet, the pronunciation label for those instances are kept as in the original database. Hence, an instance of a target P pronounced as P might be labelled as incorrect if the target P had to be aspirated.

5. Results

In this section we evaluate the GOP system described above on the L2-ARCTIC and EpaDB databases.

5.1. Comparison of Epa-DB and L2-ARCTIC results

We test the GOP system over a subset of 30 speakers, 15 male and 15 female, of the Epa-DB database (20 other speakers are left as held-out for future experiments) and over the 4 Spanish speakers, 2 male and 2 female, in the L2-ARCTIC corpus. Epa-DB includes 31366 phone instances of which 19.67% are incorrectly pronounced, while the Spanish subset of L2-ARCTIC includes 19781 phone instances of which 14.68% are incorrectly pronounced.

We present results in terms of equal error rate (EER), the false positive rate when the threshold is selected so that this rate is equal to the false negative rate. In this work, incorrectly pronounced phones are labelled as positive and correctly pronounced ones as negative. The EER reflects the discrimination performance of the system, ignoring the issue of threshold selection while also being immune to changes in the proportions between positive and negative labels. Given that the two datasets have different proportion of pronunciation errors per phone, the EER is an appropriate measure to compare performance across the two sets.

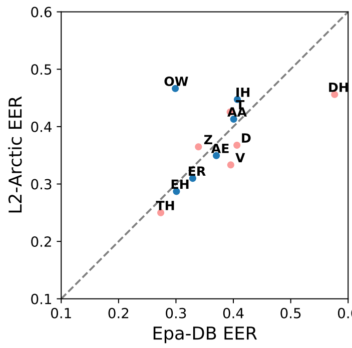

Figure 2: *EER per phone for the two databases, for phones with more than 50 samples of each class in each database. Vowels are shown in blue and consonants in pink.*

Figure 2 shows a scatter plot of the results obtained on the two databases using the GOP system. We only show the phones that appear more than 50 times per class in each database. The diagonal line across the plot indicates the points where the system would have the same performance on both databases. We can see that performance for most phones fall close to this line. This is somewhat surprising given the fact that the Epa-DB waveforms are much noisier than those in L2-ARCTIC, since they were collected through the internet by people in their homes, using their own microphones and with some level of background noise. They are also less clean than the LibriSpeech data used to train the acoustic model used by the GOP system. Yet, the GOP scores seem to be robust to this mismatch, resulting in similar EER values for Epa-DB and L2-ARCTIC for most of the phones. The two cases where EERs differ significantly are probably explained by the statistics in Figure 1, where we can see that the percent of incorrect cases (for DH) or the frequent variants (for OW) are different across datasets.

5.2. Epa-DB Analysis

Table 1 shows the performance of the GOP system on Epa-DB in terms of EER and F1 score, which is a commonly used metric in pronunciation scoring. In the case of F1 score, we select the threshold that leads to the best value for this score on the test data, leading to a somewhat optimistic estimate. Results in Table 1 show that the GOP system has very poor performance in terms of EER and F1 score for many phones. This result is inline with previous papers using the GOP approach [10, 20], which generally show that this algorithm works poorly on some subset of phones.

Table 1: *Number of total phones, percentage of incorrectly pronounced, maximum F1 score and EER results for all target phones in the automatic alignments with more than 50 instances of each class, in ascending order of EER.*

Phone	Total	% Errors	F1	EER
EY	441	13.83	0.80	0.12
JH	178	39.89	0.81	0.18
AY	1040	5.96	0.45	0.21
R	1298	18.34	0.53	0.27
TH	298	21.48	0.56	0.27
OW	417	37.41	0.64	0.30
EH	812	11.58	0.43	0.30
HH	473	17.12	0.44	0.32
NG	299	41.14	0.66	0.33
ER	933	30.98	0.58	0.33
ZH	178	61.80	0.82	0.34
Z	848	80.66	0.91	0.34
K	920	15.65	0.43	0.34
G	443	16.03	0.41	0.35
UH	352	18.47	0.41	0.37
AE	799	62.83	0.79	0.37
T	2162	18.41	0.37	0.39
V	445	60.22	0.78	0.40
AA	587	36.63	0.57	0.40
D	846	28.13	0.50	0.41
IH	1549	40.61	0.59	0.41
B	414	16.67	0.39	0.42
AO	491	41.75	0.59	0.44
P	724	20.03	0.36	0.46
UW	712	10.11	0.22	0.50
DH	1048	8.30	0.16	0.58

Finally, Figure 3 shows the score distributions for the correctly and incorrectly pronounced instances in Table 1 that have at least 200 samples in each class. The T phone is excluded because, as mentioned above, it is, in fact, composed of two target allophones, T and Th, which prevents any useful analysis. The black curves correspond to the score distribution for the correct (dashed) and incorrect (solid) instances. These curves are normalized to have area under the curve of 1.0. Colored curves show the contribution of each of the incorrect variants to the solid black curve.

We can see that, for some target phones, it is only one of the variants that causes most of the errors. For example, for the D target phone, the T variant is almost always labelled as correct for the EER threshold. A similar effect can be seen on IH, where the IY variant is mostly confused with the correct variant. For AO, most of the errors come from the Spanish O ([o]) variant. These two cases seem to indicate that the system is confusing long with short variants, which is expected given that the GOP score is normalized by the duration of the phones.

One of the main reasons for the poor performance of the GOP algorithm is that it was not trained for the task we aim to solve. This is, of course, a necessary trade off when annotated data for the task of interest is not available. Yet, our new Epa database will allow us to explore techniques where the models are directly trained (or finetuned) for the pronunciation scoring task using labelled non-native data. Several recent works have shown gains from this type of approaches [21, 10, 27]. Further,

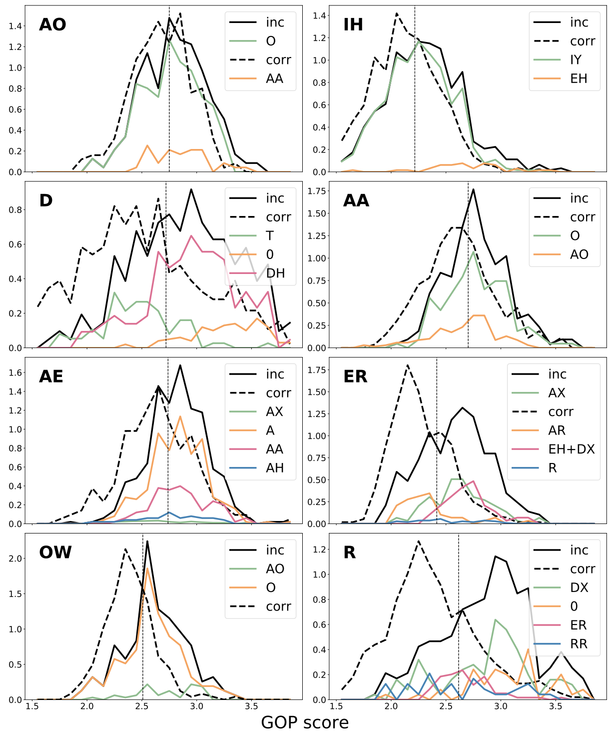

Figure 3: *Distribution of scores for each variant of a certain target phone for correct (black dashed line) and incorrect (black solid line) instances. The target phone is indicated in the title of the figure. Incorrect variants that make up the solid back curve are shown in color. The label "0" corresponds to deletion. The dotted vertical line corresponds to the EER threshold.*

we will consider the use of syllables as the unit for scoring and correcting the student. This might allow for better performing models without loosing much precision in the feedback to the student.

6. Conclusions

In this paper we compare the new Epa-DB with the Spanish-speaker subset of the L2-ARCTIC corpus. We present an analysis of the most frequent substitution errors in both databases, showing that they mostly coincide with each other and with the expected problematic phones for Spanish learners of English reported in the literature. Additionally, we compare results on both databases using a state-of-the-art DNN-HMM Goodness of Pronunciation (GOP) system trained with a large corpus of native English speech and find that, on most phones, results are similar across the two datasets. This is somewhat surprising given the fact that the Epa-DB recordings are much noisier than those in L2-ARCTIC, since they were collected through the internet by people in their homes. The results shown in this paper serve as validation of the new Epa Database and provide a baseline for comparing any future developments using this data. We provide an accompanying repository of code to replicate the results in this paper.

7. Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101007666 / ESPERANTO / H2020-MSCA-RISE-2020.

8. References

[1] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "Automatic evaluation and training in english pronunciation," in *First International Conference on Spoken Language Processing*, 1990.

[2] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1997, pp. 1471–1474.

[3] S. Wei, G. Hu, Y. Hu, and R.-H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communication*, vol. 51, no. 10, pp. 896–905, 2009.

[4] S.-Y. Yoon, M. Hasegawa-Johnson, and R. Sproat, "Landmark-based automated pronunciation error detection," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[5] G. Seed and J. Xu, "Integrating technology with language assessment: Automated speaking assessment," *Learning and Assessment: Making the Connections*, p. 286, 2017.

[6] M. G. O'Brien, T. M. Derwing, C. Cucchiarini, D. M. Hardison, H. Mixdorff, R. I. Thomson, H. Strik, J. M. Levis, M. J. Munro, J. A. Foote *et al.*, "Directions for the future of technology in pronunciation research and teaching," *Journal of Second Language Pronunciation*, vol. 4, no. 2, pp. 182–207, 2018.

[7] J. Vidal, L. Ferrer, and L. Brambilla, "Epadb: a database for development of pronunciation assessment systems," *Proc. Interspeech 2019*, pp. 589–593, 2019.

[8] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-arctic: A non-native english speech corpus," in *Proc. Interspeech*, 2018, p. 2783–2787. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1110

[9] O. L. Uribe Enciso, F. Hernandez, S. Smith, R. Pabón, and A. Steve, "Problematic phonemes for spanish-speakers' learners of english." *GIST Education and Learning Research Journal*, vol. 19, pp. 215–238, 2019.

[10] W. Hu, Y. Qian, and F. K. Soong, "An improved dnn-based approach to mispronunciation detection and diagnosis of l2 learners' speech." in *SLaTE*, 2015, pp. 71–76.

[11] A. Neri, C. Cucchiarini, and H. Strik, "The effectiveness of computer-based speech corrective feedback for improving segmental quality in l2 dutch," *ReCALL: the Journal of EUROCALL*, vol. 20, no. 2, p. 225, 2008.

[12] J. Kominek, A. W. Black, and V. Ver, "Cmu arctic databases for speech synthesis," 2003.

[13] J. M. Lipski, "Geographical and social varieties of spanish: An overview," *The handbook of Hispanic linguistics*, pp. 1–26, 2012.

[14] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[15] 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.

[16] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study," in *SLaTE*, 2009.

[17] H. Franco, L. Ferrer, and H. Bratt, "Adaptive and discriminative modeling for improved mispronunciation detection," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7709–7713.

[18] W. Hu, Y. Qian, and F. K. Soong, "A new neural network based logistic regression classifier for improving mispronunciation detection of l2 language learners," in *The 9th International Symposium on Chinese Spoken Language Processing*. IEEE, 2014, pp. 245–249.

[19] F. Landini, L. Ferrer, and H. Franco, "Adaptation approaches for pronunciation scoring with sparse training data," in *International Conference on Speech and Computer*. Springer, 2017, pp. 87–97.

[20] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," *arXiv preprint arXiv:2008.08647*, 2020.

[21] H. Huang, H. Xu, Y. Hu, and G. Zhou, "A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection," *The Journal of the Acoustical Society of America*, vol. 142, no. 5, pp. 3165–3177, 2017.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[23] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.

[24] 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.

[25] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[26] G. Kondrak, "A new algorithm for the alignment of phonetic sequences," in *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000.

[27] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2016.