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

Assessing Posterior-Based Mispronunciation Detection on Field-Collected Recordings from Child Speech Therapy Sessions

Adam Hair[1], *Guanlong Zhao*[1], *Beena Ahmed*[2], *Kirrie J. Ballard*[3], *Ricardo Gutierrez-Osuna*[1]

[1]Department of Computer Science & Engineering, Texas A&M University, USA
[2]School of Electrical Engineering & Telecommunications, University of New South Wales, Australia
[3]Sydney School of Health Sciences, The University of Sydney, Australia

{adamhair, rgutier}@tamu.edu, beena.ahmed@unsw.edu.au, kirrie.ballard@sydney.edu.au

Abstract

A critical component of child speech therapy is home practice with a caregiver, who can provide feedback. However, caregivers oftentimes struggle with accurately rating speech and with perceiving pronunciation errors. One potential solution for this issue is to embed automatic mispronunciation-detection (MPD) algorithms within digital speech therapy applications. To address the need for MPD within child speech therapy, we investigated posterior-based mispronunciation detection using a custom corpus of disordered speech from children that had been manually annotated by an expert clinician. Namely, we trained a family of phoneme-specific logistic regression classifiers (LRC) and support vector machines (SVM) on log posterior probability and log posterior ratio features. Our results show that these classifiers outperformed baseline Goodness of Pronunciation scoring by 11% and 10%, respectively. Even more importantly, in an offline test, the LRC and SVM classifiers outperformed student clinicians at identifying mispronunciations by 18% and 16%, respectively. These results suggest that posterior-based mispronunciation detection may be suitable to provide at-home therapy feedback for children.

Index Terms: computer-assisted pronunciation training (CAPT), pronunciation verification, child speech

1. Introduction

Children with speech disorders benefit from frequent and high-intensity speech therapy [1] to develop and practice new skills [2]. To increase treatment dosage, children oftentimes practice at home under the supervision of a caregiver [3]. Home practice relies on the caregiver to lead activities and provide pronunciation feedback. However, clinicians have encountered issues with home practice delivered by caregivers, primarily low completion rates, incorrect implementation, and high therapy attrition rates [3, 4]. These problems can be attributed to difficulties making time to complete the therapy practice [5, 6] and an absence of caregiver training; many caregivers feel they lack knowledge or experience to support their child themselves [7], and others report that they sometimes feel unsure how to provide proper feedback [5]. Caregivers have also been found to rate pronunciations leniently [8] or unreliably [9] during home therapy practice, and untrained adults may generally have difficulty perceiving errors in child speech [10]. While caregivers can be trained to deliver effective phonological interventions [11], the training takes time (on the order of two months [11]) and ignores scheduling-related barriers to home practice.

A potential solution to limited caregiver availability and inconsistent pronunciation feedback is to incorporate mispronunciation detection (MPD) algorithms into digital speech therapy applications. This empowers children to practice more independently and allows caregivers to lightly supervise therapy practice, instead of directly administering the activities. MPD algorithms will invariably be less accurate than trained clinicians, but they may still rate productions more accurately and consistently than caregivers. For example, in previous work [8], we found that an MPD algorithm overwhelmingly outperformed caregivers at word-level MPD in-the-field. Although some digital child therapy projects have provided word-level feedback [12, 13], systems like these eventually need phoneme-level feedback so that speech therapy practice can target specific problematic sounds [14]. This is a substantially more challenging task because the system needs to model individual errors, rather than matching whole utterances to a certain word label. Furthermore, even though phoneme-level MPD is an active research area for second-language (L2) learners (e.g., [15, 16]), less attention has been paid to detecting mispronunciations in disordered speech from children.

In this article, we investigate whether existing MPD techniques from the L2 pronunciation training literature could be used for child speech therapy and evaluated them on a limited, but challenging, corpus of disordered speech from children collected during actual speech therapy practice. Specifically, we trained phoneme-specific classifiers to identify mispronunciations arising from inaccurate speech sound production using posterior-probability-based features proposed by Hu et al. [17]. These features are extracted from an off-the-shelf acoustic model (AM) in a manner similar to the traditional Goodness of Pronunciation (GOP) score [18], but have been shown to outperform GOP scores on adult L2 speech [17, 19]. Being able to extract features with a generic speaker-independent AM is especially important in the context of child speech therapy, since there is a scarcity of disordered child speech corpora large enough to build AMs from scratch.

Using a custom corpus with expert annotations, we implemented two types of phoneme-level classifiers based on logistic regression classifiers (LRC) and support vector machines (SVM), and compared their MPD performance against that of a baseline GOP system. The two phoneme-level classifiers predicted mispronunciations significantly better than baseline GOP, even though both systems use features based on the same underlying AMs. More importantly, the two classifiers significantly outperformed student clinicians in an offline pronunciation verification test. These results have practical significance, as they indicate that MPD – based on off-the-shelf AMs – provides more accurate pronunciation feedback than caregivers themselves. Thus, integrating MPD within speech therapy applications may improve the quality of feedback that children receive and, by reducing caregiver burden, may also facilitate

http://dx.doi.org/10.21437/Interspeech.2021-69

more frequent and high-intensity practice.

2. Background

Current approaches to MPD can generally be grouped into three categories: posterior-based, classifier-based, or rule-based. Posterior-based MPD methods score phoneme segments according to the posterior likelihood output of the production matching the target phoneme. These scores are often converted into binary pronunciation classifications by comparing against a predetermined threshold [20, 21], which yields the same output as classifier-based methods. Posterior probabilities are often derived from the output of an AM (i.e., from an automatic speech recognizer) and frequently take the form of a Goodness of Pronunciation (GOP) metric [18]. These methods are commonly used as MPD baselines [14, 22], but have also served as the foundation for novel methods [15, 23]. For child speakers, Dudy et al. [24] combined the GOP with rule-based error modeling and explicit acoustic modeling of the phonetic errors. Saz et al. [25] also deployed posterior-based MPD for child speech and increased likelihood score separation by using speaker normalization and AM adaptation.

Classifier-based approaches treat MPD as a binary classification problem, where a phoneme can either be correct or incorrect (i.e., mispronounced) [26]. Individual phoneme segments are converted into feature vectors, which are passed through a classifier to obtain a pronunciation prediction [19, 22]. Feature vectors may consist of Mel-Frequency Cepstral Coefficients (MFCC) [27], speech attribute scores [28], or posterior probabilities [17, 19]. Researchers have explored a variety of classification methods, including decision trees [14, 29], SVMs [30, 31], and more recently, various deep neural network (DNN) architectures [16, 32]. These methods have also been used with child speech. For example, Shahin et al. [33] explored a classifier-based approach using a one-class SVM trained on phonetic attribute features to detect anomalous phoneme pronunciations. Wang et al. [28] also tested classifier-based MPD for child speech, wherein they trained binary pronunciation classifiers on the distance from the expected phoneme, as measured by a Siamese network.

Rule-based methods take existing knowledge of mispronunciation patterns to identify errors, usually by including these errors in the ASR decoder lattice [34, 35]. Obtaining the necessary error patterns requires expert manual curation [35, 36] or access to large quantities of speech to identify the patterns in a data-driven fashion [37, 38]. Shahin et al. [34] deployed rule-based MPD for child speech by including expected errors as provided by a speech-language pathologist to the decoding path. They later adopted a data-driven approach, where garbage nodes along the decoding lattice collected unexpected or missing phonemes and penalty values were tuned to control garbage node acceptance rates [39].

3. Methods

Our approach to MPD consists of three steps: forced-alignment, feature extraction, and classification. First, recordings are forced-aligned against the canonical pronunciation using a pre-trained aligner [40], which automatically generates the phoneme segments. Features are extracted by passing individual speech frames to the AM, which generates the posterior probabilities, and then transforming them into the final feature vector (see below). Silence segments are discarded, leaving only speech segments for our analysis. Following Hu et al. [17], we represent each phoneme segment by a feature vector containing two types of features: Log Posterior Probabilities (LPP) and Log Posterior Ratios (LPR). The LPP is a log posterior normalized over the phoneme duration:

$$\begin{aligned} LPP(p|\mathbf{o}) &= \log P(p|\mathbf{o}; t_s, t_e) \\ &\approx \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} \log P(p|\mathbf{o}_t). \end{aligned} \tag{1}$$

where the posterior for phoneme p is obtained according to:

$$P(p|\mathbf{o}) = \sum_{s \in p} P(s|\mathbf{o}) \tag{2}$$

for each senone s associated with phoneme p, i.e., a senone shared by a tied-state triphone where the center phoneme is p. The posterior $P(s|\mathbf{o})$ comes directly from a DNN-based AM [41]. The LPR is the difference of the LPPs for phonemes p_i and p_j, given the same observation $\mathbf{o}$:

$$LPR(p_j|p_i, \mathbf{o}) = LPP(p_j|\mathbf{o}) - LPP(p_i|\mathbf{o}). \tag{3}$$

For each phoneme segment, we compute a series of LPPs and LPRs to form a feature vector. LPPs are calculated for all N phoneme classes and LPRs are calculated for all pairs p_i,p_j where p_i is the expected phoneme class and $j \in N$. The final feature vector $f(p_i|\mathbf{o}; t_s, t_e)$ concatenates LPPs and LPRs:

$$\begin{aligned} f(p_i|\mathbf{o}; t_s, t_e) = [&LPP(p_1|\mathbf{o}), LPP(p_2|\mathbf{o}), ..., LPP(p_N|\mathbf{o}), \\ &LPR(p_1|p_i, \mathbf{o}), LPR(p_2|p_i, \mathbf{o}), ..., LPR(p_N|p_i, \mathbf{o})]^T. \end{aligned} \tag{4}$$

After the individual segments are transformed into the final feature vectors, they are used to train supervised phoneme-specific classifiers with examples of correct and incorrect phoneme pronunciations. For classification, we used SVMs and LRCs; SVMs are commonly deployed for MPD (e.g., [30, 31]) and neural LRCs have also been used successfully for this task [17, 19]. Given the limited amount of data available, we use a traditional LRC instead of a neural-network-based classifier.

3.1. Goodness of pronunciation baseline

We use a GOP system as a baseline for automatic MPD. Originally, the GOP was defined as the normalized log posterior of phoneme p, which was computed as the ratio between the likelihood of the expected phoneme and the most probable phoneme [18]. These likelihoods traditionally came from a GMM-HMM AM. However, given that we use a DNN AM that directly outputs senone posteriors, the original GOP equation needs to be modified slightly. Therefore, we use the GOP computation proposed by Hu et al. [19], where the score is the ratio between the LPPs for the expected phoneme and the highest posterior across the set of all phonemes (denoted as Q):

$$GOP(p|\mathbf{o}) = LPP(p|\mathbf{o}) - \max_{q \in Q} LPP(q|\mathbf{o}). \tag{5}$$

Following calculation, GOP scores are converted into a binary evaluation by comparing against a threshold. If the score is greater than the threshold, the phoneme segment is labeled as correctly pronounced, otherwise, the segment is labeled as incorrectly pronounced. To avoid biasing the threshold due to the imbalance of positive and negative phoneme samples, we determine each phoneme-specific threshold by searching the training set for the one that yields the maximum combined F1 score; this value is the average of the F1 scores calculated for correct and incorrect pronunciation detection:

$$F1_{comb} = \frac{TN}{2TN + FN + FP} + \frac{TP}{2TP + FN + FP}. \quad (6)$$

4. Experiment

We trained the DNN AM on the Librispeech corpus [42], which contains 960 hours of adult English speech, mostly American English. This corpus is not used for any other training or testing. Specifically, we use the Kaldi Librispeech recipe[1] to train a DNN with five fully-connected hidden layers (5,000 neurons) using the p-norm non-linearity ($p = 2$). After the final hidden layer, there is a 14,000-node softmax layer that is group-summed to produce the final output across 5,816 senones. We extract 13-dimension MFCCs with 7-frame context, which are transformed with linear discriminants analysis (LDA) to form a 40-dimension feature vector, and these vectors are concatenated into nine-frame inputs (40×9) for the DNN; final input features are decorrelated using a fixed linear transform, which is computed as a modified LDA without dimensionality reduction followed by variance reduction along output dimensions with low between-class variance (see the Kaldi *dnn2* documentation[2]). The DNN AM output represents the senone posterior probabilities conditioned on an input observation, i.e., $P(s|\mathbf{o})$.

We implemented the LRC and SVM using the Scikit-learn Python library [43]. The LRC used an L2 penalty and iterated until the model converged during per-phoneme training. The SVM used a fourth-degree polynomial kernel. These hyperparameters were determined empirically. Forced alignment was performed with the Montreal Forced Aligner [40]. We used a 40-phoneme set, so each feature vector contained 80 features: 40 LPPs and 40 LPRs.

For our MPD tests, we used a custom corpus of disordered speech from children. This corpus is an expert-annotated subset of a larger collection of speech therapy audio recordings, which were gathered as part of a longitudinal evaluation of a tablet-based speech therapy game [8]. This corpus contains 2,336 recordings of prompted single or compound word utterances from nine children with speech sound disorders (native Australian-English speakers), each practicing 20 words; these words were selected to address the children's speech therapy needs, rather than to maximize phonetic coverage. Recordings were captured at 16kHz on a tablet at the children's homes, and contain some distortions and excited speech. Children generally spoke at a normal volume. The corpus contains 10,059 non-silence phonemes, 27% of which are mispronounced; mispronunciations are not evenly distributed across phoneme classes. For example, the phoneme /ʒ/ is not represented in this corpus and /w/ only has correct samples; all other phonemes have samples with mispronunciations. Table 1 shows the 16 most common phonemes in the corpus. Each utterance was annotated for phoneme-level errors by a speech-language pathologist at the University of Sydney. All annotations were collected offline and are binary labels of correctness for each phoneme; they do not provide the actual sound produced for substitution errors.

In this article, we define a true positive (TP) as a pronunciation error that was correctly labeled as a pronunciation error, and a true negative (TN) as a correct pronunciation labeled as correct. Additionally, since the proposed MPD systems cannot handle insertion errors, we focus only on substitution and deletion errors.

Table 1: *Top 16 phonemes in the corpus as percent of total non-silence phonemes. Bold indicates used in experiments*

Phoneme	Frequency	Phoneme	Frequency
/l/	8.7%	**/s/**	3.8%
/ʌ/	7.6%	**/i/**	3.7%
/ɝ/	6.8%	/m/	3.5%
/t/	6.1%	/n/	3.4%
/k/	5.6%	**/ʃ/**	3.4%
/ɪ/	4.8%	**/ɛ/**	3.2%
/p/	4.8%	/b/	3.0%
/æ/	4.1%	**/ʧ/**	2.6%

Table 2: *Average combined F1 score from 5-fold cross validation (standard error)*

	LRC	SVM	GOP	Chance
/ʌ/	57.2 (2.0)	**58.0 (2.3)**	54.6 (1.4)	48.5 (0.0)
/ʧ/	**55.7 (2.4)**	43.6 (3.1)	58.4 (1.1)	47.6 (0.0)
/ɛ/	77.7 (1.1)	**78.3 (1.6)**	69.1 (1.5)	49.7 (0.0)
/ɝ/	58.8 (1.9)	**62.9 (2.0)**	44.6 (1.6)	49.3 (0.0)
/i/	61.0 (4.4)	**62.7 (2.7)**	50.9 (1.2)	46.7 (0.0)
/l/	50.7 (1.5)	**52.0 (2.3)**	50.4 (1.6)	42.1 (0.0)
/s/	75.0 (2.0)	**77.4 (2.0)**	52.4 (3.6)	49.9 (0.0)
/ʃ/	**59.8 (3.4)**	57.6 (2.0)	66.0 (1.4)	49.9 (0.0)
All	62.0 (1.6)	61.6 (1.9)	55.8 (1.4)	48.0 (0.4)

5. Results

5.1. Comparison against automated baseline

To ensure that there were enough samples to train and test the classifiers, we only examined phonemes which had at least 60 samples of correct and incorrect pronunciations in the child corpus ($n = 8$ phonemes). For each phoneme, we trained two phoneme-specific classifiers: an LRC and an SVM. To accommodate our small corpus, we used 5-fold stratified cross-validation (each fold contains the same class distribution) when evaluating classifier performance. For each fold, both classifiers were trained, labels were predicted for the test data, and the predictions from each classifier were scored against the expert labels. As a baseline, we also computed the performance of GOP scoring at each fold. Phoneme-specific GOP thresholds were used to convert test segment scores into labels, which were compared against the expert labels. Because each phoneme has a different correct/incorrect class distribution, we also calculated the performance of a random binary classifier as a measure of chance level. The average combined F1 scores for all phonemes are shown in Table 2. All three methods performed above chance level ($p < 0.05$, paired t-test). Both the LRC and SVM achieved significantly higher combined F1 scores than the GOP baseline ($p < 0.05$, paired t-test). The LRC and SVM demonstrated 11.1% and 10.4% relative increases for average combined F1 score, respectively, compared to GOP. Although the SVM outperformed the LRC for six phonemes, on average, there was no significant difference between the LRC and SVM ($p > 0.05$, paired t-test). The SVM and GOP systems each

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/librispeech

[2] https://kaldi-asr.org/doc/dnn2.html

Table 3: *Average combined F1 score across the set of student clinician annotations (standard error)*

Student Clinician	LRC	SVM	Chance
69.0 (1.6)	**81.5 (0.4)**	80.1 (0.4)	48.9 (0.1)

failed to classify one phoneme correctly: the SVM had problems with /ʧ/ and the GOP struggled with /ɝ/.

5.2. Comparison against human raters

To put our MPD results in context for at-home speech therapy with an untrained caregiver, we also compared our performance against that of an independent set of human evaluators. For this purpose, we asked 32 student clinicians to annotate a randomly-selected subset of 154 recordings from our corpus, 27.6% of which are mispronounced; again, mispronunciations are not evenly distributed across phonemes. Due to the annotation process, each evaluator labeled a slightly different quantity of the 154 recordings. As the final step in our analysis, we compared these student clinician labels against classifier predictions, only considering the eight phonemes analyzed above. Phoneme-specific SVMs and LRCs were trained using phoneme samples from all recordings in the corpus not annotated by the student clinicians. We treated evaluator annotations as another set of predictions and scored them against the expert annotations, which we used as ground truth. For each evaluator, we calculated their performance and the chance level for the phoneme set they annotated. Additionally, each set of student-annotated phonemes was labeled by the LRCs and SVMs; these predictions were also compared against the expert annotations.

Average performance on the 154-recording subset for student evaluators and classifiers is displayed in Table 3. The student clinicians labeled the phoneme segments well above chance level ($p << 0.05$, paired t-test). However, both automated approaches significantly outperformed the students ($p << 0.05$, paired t-test). The LRC and SVM obtained combined F1 scores 18.1% and 16.1% higher, respectively, relative to the student clinicians. On this subset, the LRC achieved a significantly higher combined F1 score than the SVM ($p < 0.05$, paired t-test). Differences in combined F1 scores between the automated baseline test (Table 2) and the human raters test (Table 3) are partially explained by the class distributions of their respective training data; on average, the classifiers trained for the GOP baseline test had more incorrect samples per phoneme ($\mu = 39.2\%$, $\sigma = 18.1\%$) than the classifiers trained for the human raters test ($\mu = 18.6\%$, $\sigma = 7.30\%$).

6. Discussion and conclusion

Our results show that phoneme-specific classifiers trained using posterior-probability-based features identify mispronunciations in field-collected disordered speech from children significantly better than a baseline GOP system. This follows results presented by Hu et al. [17], even though they used a neural-network-based classifier and we used traditional classifiers. We found no significant difference between LRC and SVM MPD on the entire corpus. Notably, both types of phoneme-level classifiers significantly outperformed student clinicians at identifying mispronunciations in a subset of our corpus. This suggests that the presented methods may approximate expert clinician evaluations better than students with some training and, as such, may also outperform caregivers. These results further strengthen the argument that child speech therapy systems should include automated MPD to improve the quality of pronunciation feedback received at home.

Though our classifiers were trained with phoneme-specific data, we used global classifier hyperparameters (e.g., SVM kernel, LRC penalty). Thus, further improvements may be achieved by setting hyperparameters on a per phoneme basis. Speech production is a complex process, with many variables contributing to the final sound (place, manner, voicing, etc.). Accordingly, phoneme-specific hyperparameters may help classifiers better identify pronunciation errors. Another interesting next step would be further investigating one-shot learning techniques, such as the Siamese network used by Wang et al. [28]; these methods may be able to extract meaningful pronunciation information using limited corpora of disordered speech from children, such as the one presented in this manuscript.

Our goal with this type of system is not to replace clinicians or clinic visits, but to better approximate clinician evaluations for home-based speech therapy practice. This is especially important given the difficulty some adult carers have identifying errors in their child's speech [4, 10]; some caregivers have been shown to evaluate word-level pronunciation below chance level [8]. Additionally, even though caregivers are motivated to help their child, some are reluctant to take the lead and want clinicians to do the decision making during therapy practice [9, 44]; an automated system that imitates clinician ratings helps to fill this desire. Although significant work remains in the speech therapy mispronunciation domain, the results presented in this article suggest that phoneme-level classifiers perform well over chance level and can even outperform student clinicians when comparing against expert evaluations. As such, child speech therapy application designers could use these methods to provide automated feedback in their systems. Significantly, this can reduce caregiver scheduling burdens by allowing them to lightly supervise and support, instead of directly managing home practice as an evaluator, thereby increasing the quality and quantity of speech therapy children receive.

7. Acknowledgements

This work was made possible by NPRP Grant # [8-293-2-124] from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

8. References

[1] E. Maas, C. Gildersleeve-Neumann, K. J. Jakielski, and R. Stoeckel, "Motor-based intervention protocols in treatment of childhood apraxia of speech (cas)," *Current developmental disorders reports*, vol. 1, no. 3, pp. 197–206, 2014.

[2] D. C. Thomas, P. McCabe, and K. J. Ballard, "Rapid syllable transitions (rest) treatment for childhood apraxia of speech: The effect of lower dose-frequency," *Journal of Communication Disorders*, vol. 51, pp. 29–42, 2014.

[3] E. Sugden, E. Baker *et al.*, "An australian survey of parent involvement in intervention for childhood speech sound disorders," *IJSLP*, vol. 20, no. 7, pp. 766–778, 2018.

[4] D. C. Thomas, P. McCabe, and K. J. Ballard, "Combined clinician-parent delivery of rapid syllable transition (rest) treatment for childhood apraxia of speech," *IJSLP*, vol. 20, no. 7, pp. 683–698, 2018.

[5] E. Sugden, N. Munro *et al.*, "Parents' experiences of completing home practice for speech sound disorders," *Journal of Early Intervention*, vol. 41, no. 2, pp. 159–181, 2019.

[6] R. Goodhue, M. Onslow, S. Quine, S. O'Brian, and A. Hearne, "The lidcombe program of early stuttering intervention: mothers' experiences," *Journal of Fluency Disorders*, vol. 35, no. 1, pp. 70–84, 2010.

[7] K. E. Davies, J. Marshall, L. J. Brown, and J. Goldbart, "Co-working: Parents' conception of roles in supporting their children's speech and language development," *Child Language Teaching and Therapy*, vol. 33, no. 2, pp. 171–185, 2017.

[8] A. Hair, K. J. Ballard, C. Markoulli, P. Monroe, J. McKechnie, B. Ahmed, and R. Gutierrez-Osuna, "A longitudinal evaluation of tablet-based child speech therapy with apraxia world," *TACCESS*, vol. 14, no. 1, pp. 14–59, 2021.

[9] D. C. Thomas, P. McCabe, K. J. Ballard, and G. Bricker-Katz, "Parent experiences of variations in service delivery of rapid syllable transition (rest) treatment for childhood apraxia of speech," *Developmental neurorehabilitation*, vol. 21, no. 6, pp. 391–401, 2018.

[10] B. Munson, J. M. Johnson, and J. Edwards, "The role of experience in the perception of phonetic detail in children's speech: A comparison between speech-language pathologists and clinically untrained listeners," *AJSLP*, vol. 21, pp. 124–139, 2012.

[11] E. Sugden, E. Baker *et al.*, "Evaluation of parent-and speech-language pathologist–delivered multiple oppositions intervention for children with phonological impairment: A multiple-baseline design study," *AJSLP*, vol. 29, no. 1, pp. 111–126, 2020.

[12] B. Ahmed, P. Monroe, A. Hair, C. T. Tan, R. Gutierrez-Osuna, and K. J. Ballard, "Speech-driven mobile games for speech therapy: User experiences and feasibility," *IJSLP*, vol. 20, no. 6, pp. 644–658, 2018.

[13] J. S. Duval, E. Márquez Segura, and S. Kurniawan, "Spokeit: A co-created speech therapy experience," in *CHI Extended Abstracts*, 2018, pp. 1–4.

[14] B. Munson, J. M. Johnson, and J. Edwards, "Comparing different approaches for automatic pronunciation error detection," *Speech communication*, vol. 51, no. 10, pp. 845–852, 2009.

[15] H. Ryu and M. Chung, "Mispronunciation diagnosis of l2 english at articulatory level using articulatory goodness-of-pronunciation features." in *SLaTE*, 2017, pp. 65–70.

[16] W.-K. Leung, X. Liu, and H. Meng, "Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis," in *ICASSP*, 2019, pp. 8132–8136.

[17] 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 *ISCSLP*, 2014, pp. 245–249.

[18] 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.

[19] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[20] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in *EUROSPEECH*, 1997, pp. 645–648.

[21] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *EUROSPEECH*, 1999, pp. 851–854.

[22] V. Arora, A. Lahiri, and H. Reetz, "Phonological feature-based speech recognition system for pronunciation training in non-native language learning," *The Journal of the Acoustical Society of America*, vol. 143, no. 1, pp. 98–108, 2018.

[23] H. Huang, H. Xu, X. Wang, and W. Silamu, "Maximum f1-score discriminative training criterion for automatic mispronunciation detection," *IEEE/ACM TASLP*, vol. 23, no. 4, pp. 787–797, 2015.

[24] S. Dudy, M. Asgari, and A. Kain, "Pronunciation analysis for children with speech sound disorders," in *EMBC*, 2015, pp. 5573–5576.

[25] O. Saz, E. Lleida, and W.-R. Rodríguez, "Avoiding speaker variability in pronunciation verification of children's disordered speech," in *WOCCI*, 2009, pp. 1–5.

[26] M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," in *International Symposium on automatic detection on errors in pronunciation training*, 2012, pp. 1–8.

[27] J. Van Doremalen, C. Cucchiarini, and H. Strik, "Automatic detection of vowel pronunciation errors using multiple information sources," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, 2009, pp. 580–585.

[28] J. Wang, Y. Qin, Z. Peng, and T. Lee, "Child speech disorder detection with siamese recurrent network using speech attribute features," in *Interspeech*, 2019, pp. 3885–3889.

[29] A. Ito, Y.-L. Lim, M. Suzuki, and S. Makino, "Pronunciation error detection method based on error rule clustering using a decision tree," in *EUROSPEECH*, 2005, pp. 173–176.

[30] 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.

[31] H. Franco, L. Ferrer, and H. Bratt, "Adaptive and discriminative modeling for improved mispronunciation detection," in *ICASSP*, 2014, pp. 7709–7713.

[32] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *TASLP*, vol. 25, no. 1, 2017.

[33] M. Shahin and B. Ahmed, "Anomaly detection based pronunciation verification approach using speech attribute features," *Speech Communication*, vol. 111, pp. 29–43, 2019.

[34] M. Shahin, B. Ahmed, J. McKechnie, K. Ballard, and R. Gutierrez-Osuna, "A comparison of gmm-hmm and dnn-hmm based pronunciation verification techniques for use in the assessment of childhood apraxia of speech," in *Interspeech*, 2014, pp. 1583–1587.

[35] A. M. Harrison, W. Y. Lau, H. M. Meng, and L. Wang, "Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer," in *Interspeech*, 2008, pp. 2787–2790.

[36] H. Meng, Y. Y. Lo, L. Wang, and W. Y. Lau, "Deriving salient learners' mispronunciations from cross-language phonological comparisons," in *ASRU*, 2007, pp. 437–442.

[37] Y.-B. Wang and L.-s. Lee, "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning," *IEEE/ACM TASLP*, vol. 23, no. 3, pp. 564–579, 2015.

[38] S. Mao, X. Li *et al.*, "Unsupervised discovery of an extended phoneme set in l2 english speech for mispronunciation detection and diagnosis," in *ICASSP*, 2018, pp. 6244–6248.

[39] M. Shahin, B. Ahmed, A. Parnandi, V. Karappa, J. McKechnie, K. J. Ballard, and R. Gutierrez-Osuna, "Tabby talks: An automated tool for the assessment of childhood apraxia of speech," *Speech Communication*, vol. 70, pp. 49–64, 2015.

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

[41] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[42] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[43] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[44] N. Watts Pappas, L. McAllister, and S. McLeod, "Parental beliefs and experiences regarding involvement in intervention for their child with speech sound disorder," *Child Language Teaching and Therapy*, vol. 32, no. 2, pp. 223–239, 2016.