Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Enhancing GOP in CTC-Based Mispronunciation Detection with Phonological Knowledge

Aditya Kamlesh Parikh, Cristian Tejedor-Garcia, Catia Cucchiarini, Helmer Strik

Centre for Language Studies, Radboud University, the Netherlands

aditya.parikh@ru.nl, cristian.tejedorgarcia@ru.nl, catia.cucchiarini@ru.nl, helmer.strik@ru.nl

Abstract

Computer-Assisted Pronunciation Training (CAPT) systems employ automatic measures of pronunciation quality, such as the goodness of pronunciation (GOP) metric. GOP relies on forced alignments, which are prone to labeling and segmentation errors due to acoustic variability. While alignment-free methods address these challenges, they are computationally expensive and scale poorly with phoneme sequence length and inventory size. To enhance efficiency, we introduce a substitution-aware alignment-free GOP that restricts phoneme substitutions based on phoneme clusters and common learner errors. We evaluated our GOP on two L2 English speech datasets, one with child speech, My Pronunciation Coach (MPC), and SpeechOcean762, which includes child and adult speech. We compared RPS (restricted phoneme substitutions) and UPS (unrestricted phoneme substitutions) setups within alignment-free methods, which outperformed the baseline. We discuss our results and outline avenues for future research.

Index Terms: goodness of pronunciation, GOP, phoneme recognition, Computer-Assisted Pronunciation Training

1. Introduction

Language is a fundamental skill that shapes communication, cognitive development, and cultural integration [1] and learning languages other than the native one is essential in our globalized society. Traditional classroom settings often make it difficult for teachers to provide the degree of individualized attention that is required for high-quality language learning, especially when it comes to speaking and pronunciation [2, 3]. However, early and effective pronunciation training can help language learners improve their ability to master the language [4].

CAPT systems that incorporate Automatic Speech Recognition (ASR) technology can offer promising solutions to these challenges by providing personalized and scalable learning experiences [4, 5]. These systems enable learners to practice independently through "read-aloud" exercises, incorporating Mispronunciation Detection and/or Diagnosis (MDD) to deliver corrective feedback at the phoneme, word, and sentence levels [6, 7]. Such targeted feedback can help learners bridge the gap between their L1 and L2 while promoting accurate pronunciation and confident communication [8, 9].

The goodness of pronunciation (GOP) is a widely used measure in CAPT research that quantifies pronunciation quality by analyzing posterior probabilities from an ASR system's acoustic model [10, 11]. Initially developed using Hidden Markov Models (HMMs), GOP has since evolved based on deep neural networks (DNNs). Variants such as weighted-GOP [12], lattice-based GOP [13], and context-aware GOP [14] have further improved accuracy and robustness. More recently, the emergence of Self-Supervised Learning (SSL) models, such as Wav2vec2.0, Hubert, and WavLM, has significantly advanced MDD and phoneme recognition tasks [15, 16, 17]. These pretrained models extract rich speech embeddings and require less labeled data for fine-tuning compared to the traditional supervised approaches. In MDD, they are used to compare canonical transcriptions with phoneme recognition outputs to identify pronunciation errors. However, this process involves challenges related to sequence alignment and the need for explicitly labeled mispronunciation data, which is expensive to produce.

In the Connectionist Temporal Classification (CTC) framework, GOP scores for SSL fine-tuned phoneme recognition models can be computed using either forced alignment or alignment-free approaches [18, 19]. Forced alignment can be unreliable in non-native speech due to acoustic variability, potentially overlooking pronunciation errors [20]. The alignment-free method by Cao et al. [19] computes GOP features using scalar and multi-dimensional vector representations, improving the handling of substitution and deletion errors. However, in multilingual phoneme recognition with large phoneme inventories, this becomes computationally expensive, since phonemes in the canonical transcript must be substituted or inserted with others in the phoneme inventory, exponentially increasing computational costs, memory demands, and false positives. As the phoneme inventory expands, the CTC graph complexity grows quadratically [21], making real-time pronunciation learning impractical since immediate feedback is required.

To address these challenges in alignment-free CTC-based MDD for large phoneme inventories, we propose incorporating phoneme clustering [22, 23, 24] and learner-specific error modeling [25, 26]. By grouping acoustically and articulatory similar phonemes, phoneme clustering reduces computational complexity while preserving essential phonemic distinctions. Integrating knowledge of common substitution and deletion errors made by non-native speakers can further refine MDD.

To the best of our knowledge, no prior work has applied phoneme clustering and learner-specific error modeling in an alignment-free CTC-based MDD system. To investigate the potential of our novel approach, we conducted a study that addressed the following research question (RQ): How do phoneme clustering and learner-specific error modeling affect the performance of an alignment-free CTC-based MDD system compared to unrestricted phoneme substitutions?

2. Methodology

2.1. GOP Definitions

First, we follow the definition of GOP by Witt and Young [11]. They compute GOP using the sequence of feature vectors $\mathbf{O}_1^T = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$ of length T and the corresponding

10.21437/Interspeech.2025-829

canonical phoneme transcription $L_{\text{cano}} = \{l_1, \ldots, l_{|L_{\text{cano}}|}\}$. For a given phoneme $l_i \in L_{\text{cano}}$, the original definition of GOP is based on the log-posterior probability of that phoneme:

$$\text{GOP}_{\text{classical}}(l_i) = \log \left(\frac{p(\mathbf{O}_{t_1}^{t_2} \mid l_i) P(l_i)}{\sum_{q \in \mathcal{Q}} p(\mathbf{O}_{t_1}^{t_2} \mid q) P(q)} \right) \Big/ (t_2 - t_1), \tag{1}$$

where $\mathbf{O}_{t_1}^{t_2}$ represents the feature frames corresponding to the canonical phoneme l_i, and $\mathcal{Q}$ is the set of all phonemes in the target language. The term $p(\mathbf{O}_{t_1}^{t_2} \mid l_i)$ is typically modeled by HMM-GMM-based acoustic models. The score is then normalized by the duration of the segment, $t_2 - t_1$. This method relies on forced alignment using ASR to align the canonical transcription to the speech recording. The resulting GOP scores indicate deviations in pronunciation for each phoneme.

Second, with DNNs, GOP has been recently reformulated to utilize frame-level posterior probabilities directly from DNN-based acoustic models. The DNN-based GOP for the phoneme l_i is defined as:

$$\text{GOP}_{\text{DNN}}(l_i) = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} \log \left(P(l_i \mid \mathbf{o}_t) \right) \tag{2}$$

where $P(l_i \mid \mathbf{o}_t)$ is the posterior probability of phoneme l_i given the acoustic observation $\mathbf{o}_t$. This formulation eliminates the need for explicit likelihood computation, relying instead on frame-level outputs of the DNN model.

2.2. Alignment-Free CTC Based GOP

From [19], we adapt the CTC-based alignment-free approach. This approach operates in two stages, taking speech features ($\mathbf{O}_T$) and the canonical transcription ($L_{\text{canonical}}$) as input. In this framework, the computation of GOP relies on the probability of the complete canonical transcription as well as that of the individual phonemes within the sequence. The alignment-free GOP is formulated as:

$$\text{GOP}_{\text{alignment-free}} = \log \left(\frac{P(L_{\text{canonical}} \mid \mathbf{O}_T)}{P(L(i) \mid \mathbf{O}_T)} \right) \tag{3}$$

This method works by first calculating the probability of the full canonical sequence $P(L_{\text{canonical}} \mid \mathbf{O}_T)$. Then, we compute the probability of each phoneme in the canonical sequence by substituting it with other phonemes from the vocabulary or deleting the phoneme entirely to form a perturbed sequence. While this approach can deal with deletion and substitution errors in pronunciation assessment, it can impose a significant computational burden. This raises the question of whether it would be a good method at all.

For a phoneme inventory of $V = 39$ and the canonical transcription of "*Would you like wine*" /w ʊ tʃ ʊ l aɪ k w aɪ n/ with $n = 10$ phonemes, computing the GOP score involves both substitution and deletion. Each phoneme can be substituted by $V - 1$ alternatives or deleted once, leading to:

$$\text{Total calculations} \sim n(V - 1) + n = 390$$

This demonstrates how computational cost scales with transcript length and vocabulary size.

2.3. Substitution-Aware Alignment-Free GOP

In a Substitution-Aware CTC framework, GOP computation limits substitutions to a predefined set of phonetically confusable alternatives. For instance, if a learner is likely to pronounce /ð/ as /d/, the alignment process allows /d/ as a valid alternative at the position originally labeled as /ð/, rather than evaluating all possible substitutions. This is achieved using substitution mappings—sets of potential confusions—derived from phoneme clusters and common pronunciation errors made by non-native children learning a language.

Revisiting our example sentence, "*Would you like wine*," if each phoneme is associated with at least three confusable phoneme pairs, the total number of calculations is reduced from 390 to 40, that is an approximate 90% reduction in computation.

2.3.1. Substitution Mapping Construction

A key aspect of our pronunciation assessment framework is the management of confusing phoneme pairs (or confusion sets) [23, 24]. The substitution mapping mechanism is a linguistically informed approach that limits potential phoneme replacements to acoustically or articulatorily similar pairs, preventing arbitrary substitutions. A handcrafted Phoneme Confusion Map is developed based on three main criteria: (1) Phonetic Proximity, to capture natural articulatory relationships, stops, fricatives, and nasals are only substituted with phonemes that share similar places or manners of articulation (e.g., bilabial stops: /p/ → [/b/, /m/]); (2) common L2 learner errors, based on empirical observations of non-native speech patterns, reflecting frequent pronunciation mistakes (e.g., dental fricative substitutions: /θ/ → [/ ð/, /f/]); and (3) phonological rules, prioritizing allophonic variants (e.g., the flap /ɾ/ substituting for /t/ or /d/) and vowel mergers (e.g., /ɪ/ → [/I/]), which account for common phonetic shifts across different learner populations [25, 27, 26, 28]. We applied this substitution mapping in two alignment-free GOP methods, which are described below.

2.3.2. Phoneme-Adaptive Alignment-Free GOP (PA-AF GOP)

In the numerator of Equation 3, the CTC loss is calculated by measuring how well the model's acoustic frames align with the original canonical phoneme sequence. This calculation involves the conventional forward-pass computation of α-values (forward probabilities) to obtain the sequence likelihood. We denote this function as ctc_loss(p, y), where: p is a $(V \times T)$ matrix of per-frame posterior distributions, and y is the ground-truth phoneme sequence of length N.

In the denominator of Equation 3, we introduce a substitution-aware extension to the CTC forward pass, which computes CTC loss to evaluate the denominator term in GOP scoring while accounting for phonetically plausible mispronunciations. Unlike standard CTC loss, which estimates the likelihood of the reference phoneme sequence, this function incorporates two key adaptations. First, position-specific perturbations modify the target phoneme at a given position by either deleting it to model omission errors or substituting it with acoustically confusable phonemes from a predefined *Phoneme Confusion Map*. Second, state-dependent token masking enforces substitution rules derived from linguistic knowledge by masking transitions to non-confusable phonemes during the dynamic programming computation of the forward-pass variable α.

We define this modified function as ctc_loss(p, y, pos, M), where pos denotes the index of the phoneme in y that can be altered, and M is a dictionary mapping each phoneme ID to its allowable substitutions. At the chosen index pos, the algorithm allows alignment to any phoneme in $M(y_{\text{pos}})$, where y_{pos} is the original phoneme at position pos. In the equation, a high GOP score indicates that substituting or deleting significantly decreases the log-likelihood, which means that the

phoneme was correctly pronounced; a low GOP score indicates that substitutions or deletions have minimal impact on the log-likelihood, potentially indicating mispronunciation.

2.3.3. Phoneme-Perturbed Alignment-Free GOP(PP-AF GOP)

Unlike PA-AF GOP, which integrates substitution mechanisms directly within the CTC loss computation, PP-AF GOP handles phoneme substitutions and deletions externally by modifying the label sequences before computing the standard CTC loss. As in PA-AF GOP, we utilize the *Phoneme Confusion Map* (Section 2.3.1) to guide these modifications.

The GOP score for each phoneme is computed based on the CTC loss difference between the original phoneme sequence and the perturbed phoneme sequence. For each phoneme, a set of perturbation sequences is created by (1) replacing the phoneme with mapped phonemes from the *Phoneme Confusion Map*, generating a new phoneme sequence; and (2) removing the phoneme from the sequence, creating an alternative sequence by omitting one phoneme. Finally, each perturbed sequence is evaluated based on the CTC loss of the acoustic model. The GOP score for each phoneme is computed as:

$$\text{GOP}(p) = \min(L_{\text{perturbed}}) - L_{\text{original}} \quad (4)$$

where L_{original} is the CTC loss for the original phoneme sequence, and $\min(L_{\text{perturbed}})$ is the minimum CTC loss obtained across all perturbed phoneme sequences.

A higher GOP score ($\geq$ 0) indicates that the original phoneme is more suited, while a negative GOP score suggests that a perturbation resulted in a lower CTC loss, meaning the phoneme is suboptimal or mispronounced. As in PA-AF GOP, we conducted our experiments using both Unrestricted Phoneme Substitutions (UPS) and Restricted Phoneme Substitutions (RPS) configurations.

An illustrative example of phoneme transitions in the substitution mapping is shown in Figure 1. This diagram provides a conceptual understanding of how substitution mappings function within the alignment-free method. Each red target phoneme has possible substitutions (yellow), while deletions are indicated by blue lines. Given the canonical sequence /bæt/, possible sequences include /pæt/, /mæt/, and /bæd/, among others. Additionally, deletion-based variations, such as /at/ (removal of /b/), /bt/ (removal of /a/), and /ba/ (removal of /t/), illustrate how phoneme deletion is handled in alignment-free.

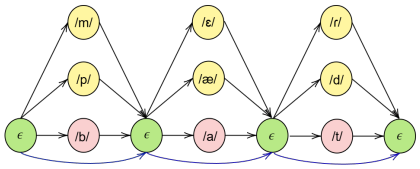


Figure 1: *An illustrative example of the transition of the phonemes in the substitution mapping construction.*

3. Experimental Procedure

3.1. Datasets

3.1.1. My Pronunciation Coach Dataset

To answer our RQ, we conducted experiments with two datasets of L2 English speech. The first one, the MPC speech database [27], is particularly challenging as it contains L2 speech of children (124 in total) learning English in Dutch secondary schools. Child speech presents specific difficulties in terms of ASR over and above those related to L2 speech. Each recording in MPC includes 53 words and 53 sentences covering various English phonemes. Sessions are classified into four quality groups: Doubtful, Overloud, OK and Excellent. For this study, we selected OK and Excellent sessions (71 speakers: 38 males, 33 females), for a total of 3,130 utterances. As MPC lacks annotated mispronunciations, we introduced artificial errors by modifying phoneme sequences. These include replacing /ð/ with /d/, /θ/ with /s/, /æ/ with /e/, /ʌ/ with /ɑ/, and simplifying diphthongs (e.g., /eɪ/ → /eː/, /əʊ/ → /o/).

3.1.2. SpeechOcean762

To allow comparisons with previous research, we also used the SpeechOcean762 dataset [29], an open-source corpus for pronunciation assessment that consists of 5,000 English utterances from 250 Mandarin-native speakers (125 adults, 125 children), with expert annotations at sentence, word and phoneme levels. Of 91,044 phoneme realizations, 3,401 were mispronunciations. We used all 5,000 utterances in our experiments.

3.2. GOP Calculations

We began our experiments by calculating GOP scores using the classical approach, as outlined in Equation 2. To obtain forced alignment at the phoneme level, we employed a Kaldi-based Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) system [30], trained on the LibriSpeech [31] 100-hour training dataset.

Regarding the forced alignment and alignment-free approaches, as described in Equations 3 and 4, we generated pronunciation lexicons using representations of International Phonetic Alphabet (IPA) with the Phonemizer toolkit [32]. For the acoustic model, we utilized an openly available fine-tuned phoneme recognition model based on [33], specifically `facebook/wav2vec2-xlsr-53-espeak-cv-ft`, hosted on HuggingFace. This multilingual model is built on the pretrained checkpoint `wav2vec2-large-xlsr-53` and has been fine-tuned on the CommonVoice dataset [34] to recognize phonetic labels across multiple languages. The phoneme inventory of this model consists of 387 phonetic labels, excluding special tokens such as `<pad>`, `<unk>`, and sentence boundary markers (`<s>` and `</s>`).

We conducted experiments using both forced alignment and alignment-free methods, employing UPS and RPS configurations, to evaluate their effectiveness in mispronunciation detection. Experiments were performed with and without substitution mapping: UPS allows any phoneme to be replaced by another without constraints, while RPS incorporates substitution mapping to restrict phoneme replacements[1].

3.3. Evaluation Metrics

We evaluated model performance using accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC). Due to class imbalance in both MPC and SpeechOcean762, where correctly pronounced phonemes dominate, we optimized the threshold by selecting the GOP percentile that maximized MCC. Additionally, we reported the ROC AUC score at this threshold to assess classification effectiveness.

For SpeechOcean762, which includes human-annotated phoneme accuracy scores, we followed [29] and used second-order polynomial regression to model the relationship between

[1] https://github.com/Aditya3107/GOP_MDD_Phonological.git

GOP scores and human ratings. To evaluate predictive performance, we reported Pearson Correlation Coefficient (PCC) with confidence intervals and Mean Squared Error (MSE) for quantifying prediction accuracy.

4. Results

Table 1 shows the experimental results on the MPC dataset. We report results for both forced alignment (FA) as the baseline, and two alignment-free approaches (PA-AF GOP, PP-AF GOP), with the latter evaluated under both RPS and UPS setups. Both PA-AF GOP and PP-AF GOP outperform FA GOP scores in most metrics, except recall, where FA achieves the highest value (0.929). However, FA has the lowest precision (0.165) due to the overclassification of correct pronunciations as mispronunciations, leading to a precision-recall trade-off. The UPS setup, which considers all phonemes for substitution, generally results in higher recall and MCC values compared to RPS. MCC improves significantly, reaching 0.587 for PA-AF GOP and 0.595 for PP-AF GOP. However, the broader search space in UPS leads to a slight reduction in precision. The precision score for RPS in PP-AF GOP (0.509) is higher than in UPS and outperforms both PA-AF GOP setups, as well as FA. However, this comes with the cost of recall, which is significantly lower than in the UPS setup.

Table 1: *Performance evaluation with MPC*

	FA	**Alignment Free approaches**			
	Baseline	**PA-AF GOP**		**PP-AF GOP**	
		RPS	**UPS**	**RPS**	**UPS**
AUC	0.747	0.869	0.941	0.883	**0.949**
Accuracy	0.514	0.888	0.898	**0.902**	0.900
Precision	0.165	0.447	0.495	**0.509**	0.500
Recall	**0.929**	0.511	0.822	0.581	0.831
F1	0.281	0.477	0.618	0.543	**0.624**
MCC	0.242	0.416	0.587	0.489	**0.595**
AUC MCC_{max}	0.698	0.720	0.864	0.759	**0.869**

Second, the experimental results of the SpeechOcean762 dataset are summarized in Table 2. Similar to the MPC dataset, FA achieves the highest recall (0.605). Among the alignment-free methods, only PP-AF GOP with UPS surpasses FA in key metrics such as accuracy, precision, F1, and MCC, achieving the best overall performance among all setups. Despite FA's strong AUC, it performs the worst in PCC compared to all alignment-free approaches. The highest PCC scores are achieved by the PP-AF GOP UPS setup (0.502 for high confidence and 0.488 for low confidence), followed by RPS (0.476 for high confidence and 0.461 for low confidence). In contrast, PA-AF GOP, both UPS and RPS, has lower PCC scores than PP-AF GOP in both setups but still outperforms the FA baseline. These results indicate that while FA retains an advantage in AUC at $MCC_{\max}$, alignment-free approaches—especially PP-AF GOP UPS—consistently achieve higher PCC scores and the lowest MSE (0.104), suggesting that they provide a more reliable phoneme-level pronunciation assessment by aligning better with human raters' phoneme scores.

While previous work on similar methods reports a PCC of 0.56 [19], those approaches involve building a complete MDD model using multidimensional GOP scores. In contrast, our method is optimized for real-time pronunciation error detection, emphasizing efficiency and simplicity. To the best of our knowledge, the highest PCC reported for phoneme-level performance on SpeechOcean762 is 0.69 [35]. However, this methodology cannot be directly compared to ours, as it includes additional contextual information as input and employs a more advanced multi-task training framework.

Table 2: *Performance evaluation with SpeechOcean762*

	FA	**Alignment Free approaches**			
	Baseline	**PA-AF GOP**		**PP-AF GOP**	
		RPS	**UPS**	**RPS**	**UPS**
AUC	0.882	0.853	0.859	0.887	**0.916**
Accuracy	0.924	0.932	0.924	0.936	**0.942**
Precision	0.262	0.239	0.261	0.273	**0.322**
Recall	**0.605**	0.412	0.452	0.470	0.556
F1	0.366	0.302	0.331	0.345	**0.408**
MCC	0.365	0.280	0.310	0.327	**0.395**
AUC MCC_{max}	**0.771**	0.681	0.701	0.712	0.756
PCC (low conf)	0.279	0.404	0.437	0.461	**0.488**
PCC (high conf)	0.297	0.419	0.424	0.476	**0.502**
MSE	0.125	0.113	0.112	0.107	**0.104**

5. Discussion and Conclusion

The results obtained in this work answered our RQ, demonstrating that phoneme clustering and learner-specific error modeling can reduce computational costs in an alignment-free CTC-based MDD system. However, it is important to note that these approaches may also lead to a decline in performance compared to unrestricted phoneme substitutions. This decline occurs because phoneme recognition models are not optimized for predefined phoneme clusters, limiting their ability to generalize pronunciation variations. Our results (Table 1 and Table 2) confirm that Substitution-Aware Alignment-Free GOP methods can provide a balance between efficiency and accuracy, making them more suitable for use in CAPT applications for users, in which instantaneous feedback is needed.

We observed that PA-AF GOP is more affected by RPS compared to PP-AF GOP. A possible explanation is that PA-AF GOP modifies the CTC forward pass internally, meaning phoneme substitutions are constrained within the CTC alignment process. When using RPS, if the correct mispronunciation is not in the predefined confusing phoneme pairs, the CTC function cannot align it correctly, leading to higher errors. This also suggests that PP-AF GOP is more robust to RPS.

We also observed that the FA baseline has high recall values in both datasets. A potential reason for this could be errors in forced alignment due to acoustic variability in non-native speech in both of our datasets. Minor acoustic variations can lead to over-detection of mispronunciations, where correctly pronounced phonemes are misclassified as errors. This results in a high recall but low precision.

One important aspect of our work is that we considered common substitution and deletion errors made by non-native speakers, but no insertion errors. This is because, unlike substitutions and deletions, insertions are often unpredictable and may include non-lexical sounds such as "umm" or "hmm". Furthermore, insertions are generally less frequent in read-aloud tasks with short utterances, as in our study, than in spontaneous conversations.

The methods that we proposed in this work can also be easily applied to other languages. Expanding the phoneme search space could enhance performance and flexibility in language learning applications. Future research could refine phoneme selection strategies to balance accuracy and efficiency. Additionally, our model-agnostic approach allows integration into any CTC-based phoneme recognition system.

6. Acknowledgements

This publication is part of the project Responsible AI for Voice Diagnostics (RAIVD) with file number NGF.1607.22.013 of the research programme NGF AiNed Fellowship Grants which is financed by the Dutch Research Council (NWO).

7. References

[1] L. Nardon, R. Steers, and C. Stone, "Language, culture, and cognition in cross-cultural communication." in *Proceedings of the new frontiers in management and organizational cognition conference*. National University of Maynooth, 2012.

[2] C. N. Onyishi and M. M. Sefotho, "Teachers' perspectives on the use of differentiated instruction in inclusive classrooms: Implication for teacher education." *International Journal of Higher Education*, vol. 9, no. 6, pp. 136–150, 2020.

[3] A. Neri, C. Cucchiarini, and H. Strik, "Feedback in computer assisted pronunciation training: When technology meets pedagogy," in *10th International CALL Conference on CALL professionals and the future of CALL research*. Antwerpen: Universiteit Antwerpen, 2002, pp. 179–188.

[4] C. Cucchiarini, A. Neri, and H. Strik, "Oral proficiency training in dutch l2: The contribution of asr-based corrective feedback," *Speech Communication*, vol. 51, no. 10, pp. 853–863, 2009.

[5] A. Neri, C. Cucchiarini, and H. Strik, "ASR corrective feedback on pronunciation: Does it really work?" in *Interspeech 2006*. ISCA, 2006, pp. 1372–Wed3A3O.2.

[6] M. Amrate and P.-h. Tsai, "Computer-assisted pronunciation training: A systematic review," *ReCALL*, vol. 37, no. 1, p. 22–42, 2025.

[7] S.-C. Liu and P.-Y. Hung, "Teaching pronunciation with computer assisted pronunciation instruction in a technological university." *Universal Journal of Educational Research*, vol. 4, no. 9, pp. 1939–1943, 2016.

[8] A. Silpachai, R. Neiriz, M. Novotny, R. Gutierrez-Osuna, J. Levis, and E. Chukharev-Hudilainen, "Corrective feedback accuracy and pronunciation improvement: Feedback that is 'good enough'," *Language Learning & Technology*, vol. 28, pp. 1–16, 2024.

[9] J. v. Doremalen, C. Cucchiarini, and H. Strik, "Automatic pronunciation error detection in non-native speech: The case of vowel errors in dutch," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. 1336–1347, 2013.

[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. M. Witt, "Use of speech recognition in computer-assisted language learning." Ph.D. dissertation, University of Cambridge, 2000.

[12] J. van Doremalen, C. Cucchiarini, and H. Strik, "Using non-native error patterns to improve pronunciation verification," in *Interspeech 2010*. ISCA, 2010, pp. 590–593.

[13] Y. Song, W. Liang, and R. Liu, "Lattice-based gop in automatic pronunciation evaluation," in *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, vol. 3. IEEE, 2010, pp. 598–602.

[14] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," in *Interspeech 2020*. ISCA, 2020, pp. 3057–3061.

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[16] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[17] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[18] L. B. Medin, T. Pellegrini, and L. Gelin, "Self-supervised models for phoneme recognition: Applications in children's speech for reading learning," in *Interspeech 2024*. ISCA, 2024, pp. 5168–5172.

[19] X. Cao, Z. Fan, T. Svendsen, and G. Salvi, "A framework for phoneme-level pronunciation assessment using ctc," in *Interspeech 2024*. ISCA, 2024, pp. 302–306.

[20] 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.

[21] A. Laptev, S. Majumdar, and B. Ginsburg, "CTC Variations Through New WFST Topologies," in *Interspeech 2022*, 2022, pp. 1041–1045.

[22] G. K. Tak and V. Bhargava, "Clustering approach in speech phoneme recognition based on statistical analysis," in *RTNSA, CNSA 2010, Chennai, India, July 23-25, 2010. Proceedings 3*. Springer, 2010, pp. 483–489.

[23] M. Meng, J. Liang, and B. Xu, "Multilingual acoustic modeling method based on phoneme clustering," *Pattern Recognition and Artificial Intelligence*, 2009.

[24] D. Oh, J.-S. Park, J.-H. Kim, and G.-J. Jang, "Hierarchical phoneme classification for improved speech recognition," *Applied Sciences*, vol. 11, no. 1, p. 428, 2021.

[25] C. Cucchiarini, H. v. d. Heuvel, E. Sanders, and H. Strik, "Error selection for asr-based english pronunciation training in'my pronunciation coach'," in *Interspeech 2021*. Florence, Italy: sn, 2011.

[26] V. Kruitbosch, "Pronunciation errors made by Dutch secondary school students in English," Master's thesis, Radboud University, 2020.

[27] C. Cucchiarini, W. Nejjari, and H. Strik, "My pronunciation coach: Improving english pronunciation with an automatic coach that listens," *Language Learning in Higher Education*, vol. 1, no. 2, pp. 365–376, 2012.

[28] A. Wheelock, "Phonological difficulties encountered by italian learners of english: An error analysis," *Hawaii Pacific University TESOL Working Paper Series*, vol. 14, pp. 41–61, 2016.

[29] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," in *Interspeech 2021*. ISCA, 2021, pp. 3710–3714.

[30] 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*. IEEE Signal Processing Society, 2011.

[31] 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.

[32] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021.

[33] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," in *Interspeech 2022*, 2022, pp. 2113–2117.

[34] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC 2020*. Marseille, France: ELRA, May 2020, pp. 4218–4222.

[35] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "A hierarchical context-aware modeling approach for multi-aspect and multi-granular pronunciation assessment," in *Interspeech 2023*. ISCA, 2023, pp. 974–978.