Evaluating Logit-Based GOP Scores for Mispronunciation Detection

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

Pronunciation assessment relies on goodness of pronunciation (GOP) scores, traditionally derived from softmax-based posterior probabilities. However, posterior probabilities may suffer from overconfidence and poor phoneme separation, limiting their effectiveness. This study compares logit-based GOP scores with probability-based GOP scores for mispronunciation detection. We conducted our experiment on two L2 English speech datasets spoken by Dutch and Mandarin speakers, assessing classification performance and correlation with human ratings. Logit-based methods outperform probability-based GOP in classification, but their effectiveness depends on dataset characteristics. The maximum logit GOP shows the strongest alignment with human perception, while a combination of different GOP scores balances probability and logit features. The findings suggest that hybrid GOP methods incorporating uncertainty modeling and phoneme-specific weighting improve pronunciation assessment.

Index Terms: GOP, logit-based GOP, mispronunciation detection, pronunciation assessment, softmax posterior probabilities

1. Introduction

In today's interconnected world, globalization has led to increased movement across borders for work, education, and other opportunities. For individuals who are adapting to a new linguistic environment, learning the local language is essential for social integration, career advancement, and overall well-being [1]. Effective communication goes beyond knowing the vocabulary and grammar: A clear pronunciation is equally important, as it impacts intelligibility, confidence, and the ability to engage in meaningful conversations [2, 3]. Poor pronunciation can lead to misunderstandings, impede effective social interactions and even create barriers in academic and professional settings [4]. However, mastering pronunciation in a second language (L2) can be challenging. Differences between the first language (L1) and the target language often result in persistent pronunciation errors. These difficulties are further compounded by the limited tools and resources available to language instructors for providing personalized pronunciation feedback.

To address these challenges, Computer-Assisted Pronunciation Training (CAPT) systems have gained popularity [5, 6]. A key component of these systems is Mispronunciation Detection and Diagnosis (MDD), which helps learners identify and correct pronunciation errors in real time [7]. Phoneme-level assessment, in particular, provides more precise feedback than broader word- or sentence-level evaluations, allowing learners to focus on specific areas for improvement [8]. One of the most widely used methods for detecting phoneme-level mispronunciations is the goodness of pronunciation (GOP) score [9].

GOP was initially introduced as a measure of pronunciation quality, estimating the probability of a phoneme and comparing it against a predefined threshold to flag mispronunciations [10]. Over time, several enhancements have improved its accuracy. Weighted-GOP [11] adjusts phoneme scores based on linguistic and acoustic factors, prioritizing phonemes prone to mispronunciation. Lattice-based GOP [12] considers multiple pronunciation possibilities using phoneme lattices, yielding more robust confidence scores. Context-aware GOP [13] incorporates phoneme transitions and durations to better capture natural pronunciation variations. More recently, multidimensional GOP features [14, 8, 15] have been introduced, leveraging richer feature representations beyond simple probability thresholds for more precise mispronunciation detection.

Despite these advancements, GOP methods face challenges in data availability and computational efficiency [16]. Annotated pronunciation data, including prosody and fluency scores, requires expert evaluation, making large-scale collection expensive. Additionally, traditional GOP methods scale poorly with phoneme size because of an increasing number of features and increasing computational costs. This highlights the need for a fast, robust, and non-trainable GOP computation method.

More recently, foundation models [17, 18] trained on massive amounts of data have been used to improve MDD. These models can be fine-tuned with significantly less data, addressing some of the data availability constraints faced by traditional GOP-based methods. GOP scores can be derived from a forced alignment using Connectionist Temporal Classification (CTC)-based phoneme recognition models, which rely on posterior probabilities generated by acoustic models. These probabilities are obtained through softmax normalization of model logits, which is a widely adopted method [19]. However, softmax-based probability estimates suffer from overconfidence [20, 21], inflating confidence in incorrect phoneme predictions and reducing the granularity needed to detect subtle mispronunciations. This issue is particularly problematic in phoneme recognition for children's speech and non-native speakers, where articulatory deviations are very common [22, 23].

To address these limitations, we propose a logit-based GOP method that directly utilizes raw logits from CTC-based models rather than softmax-normalized probabilities. Logits retain more discriminative information and avoid the gradient saturation problem inherent in softmax-based scoring [24]. We explore four logit-based metrics to enhance mispronunciation detection: Maximum Logit ($\text{GOP}_{\text{MaxLogit}}$), Mean Logit Margin [25] ($\text{GOP}_{\text{Margin}}$), Logit Variance ($\text{GOP}_{\text{LogitVariance}}$), and combined (hybrid) logit-probability $\text{GOP}_{\text{Combined}}$ scores. This novel approach provides a fast, robust, and non-trainable solution, crucial for real-time phoneme assessment.

Our method builds upon prior work on uncertainty quantifi-

cation in pronunciation assessment, such as [26], which applies GOP-based scores to dysarthric speech. However, prior studies have not explicitly investigated the use of raw logits in GOP calculations. Our approach fills this gap by utilizing logit-based metrics to improve accuracy and reliability in pronunciation assessment. Our leading research question (RQ) is: To what extent does a logit-based GOP score enhance mispronunciation detection and improve correlation with human rater scores compared to traditional softmax-based GOP scores?

2. Methodology

2.1. Definition of GOP

The GOP score, first introduced by Witt and Young [9], quantifies pronunciation quality by comparing the likelihood of a hypothesized phoneme to competing alternatives. For a phoneme p aligned to an audio segment, the original GOP formulation computes:

$$\text{GOP}_{\text{original}}(p) = \log \frac{P(\mathbf{X}|p)}{\frac{1}{N}\sum_{q\in\mathcal{Q}} P(\mathbf{X}|q)} \tag{1}$$

where $P(\mathbf{X}|p)$ is the likelihood of the acoustic features $\mathbf{X}$ given phoneme p, and $\mathcal{Q}$ represents competing phonemes.

With deep neural networks (DNNs), GOP is derived from posterior probabilities using the negative log of the mean softmax output over aligned frames [27]:

$$\text{GOP}_{\text{DNN}}(p) = -\log\left(\frac{1}{T}\sum_{t=1}^{T} P(p|\mathbf{x}_t)\right) \tag{2}$$

where $P(p|\mathbf{x}_t)$ is the softmax probability of phoneme p at frame t and T is the total number of frames in the phoneme segment. This equation interprets the mean softmax probability of the target phoneme as a probabilistic measure of pronunciation quality. This approach inherits softmax limitations such as overconfidence and gradient saturation.

2.2. Logit-Based GOPS

To address softmax limitations, we propose four novel metrics using raw logits, defined below.

2.2.1. $GOP_{MaxLogit}$

This metric captures the model's peak confidence in the target phoneme p across aligned frames t_1 to t_2:

$$\text{GOP}_{\text{MaxLogit}}(p) = \max_{t\in[t_1,t_2]} \mathbf{l}_t^{(p)} \tag{3}$$

where $\mathbf{l}_t^{(p)}$ is the logit for phoneme p at frame t. It identifies unambiguous articulations but may emphasize transient spikes.

2.2.2. GOP_{Margin}

This measure quantifies the average superiority of the target phoneme over its strongest competitor. For each frame, we compute the difference (or margin) between the target logit and the highest competing logit. The average of these margins over the segment indicates how well-separated the target phoneme is from other phonemes. This helps in cases where pronunciation errors cause phoneme confusion, which may not always be reflected in probability scores.

$$\text{GOP}_{\text{Margin}}(p) = \frac{1}{T}\sum_{t=t_1}^{t_2}\left(\mathbf{l}_t^{(p)} - \max_{k\neq p}\mathbf{l}_t^{(k)}\right), \tag{4}$$

2.2.3. $GOP_{VarLogit}$

This metric measures the *variability* of the model's confidence in predicting the target phoneme across timeframes.

$$\text{GOP}_{\text{VarLogit}}(p) = \frac{1}{T}\sum_{t=t_1}^{t_2}\left(\mathbf{l}_t^{(p)} - \mu_p\right)^2, \quad \mu_p = \frac{1}{T}\sum_{t=t_1}^{t_2}\mathbf{l}_t^{(p)}, \tag{5}$$

It is computed as the variance of the raw logit values associated with the target phoneme. A low logit variance suggests that the model consistently assigns similar confidence levels to the phoneme across frames, indicating a stable and confident recognition. Conversely, a high logit variance implies fluctuating confidence, which may occur due to acoustic distortions, coarticulation effects, or phonetic ambiguity.

2.2.4. $GOP_{Combined}$

This hybrid metric is designed to use the strengths of both logit-based and probability-based approaches to pronunciation assessment. It integrates the Mean Logit Margin, which quantifies the relative confidence of the target phoneme against competing phonemes, and the traditional GOP_{DNN}.

$$\text{GOP}_{\text{Combined}}(p) = \alpha\cdot\text{GOP}_{\text{Margin}}(p) - (1-\alpha)\cdot\text{GOP}_{\text{DNN}}(p) \tag{6}$$

where $\alpha \in [0,1]$ balances contributions. By combining these two metrics, the combined score may provide a more balanced assessment of pronunciation quality, mitigating the weaknesses of each individual measure. This hybrid approach ensures that both posterior probability (via GOP_{DNN}) and phoneme separability (via $\text{GOP}_{\text{margin}}$) contribute to the final score, making it more sensitive to pronunciation deviations while reducing the impact of softmax-related limitations.

2.3. GOP Calculations

In our study for forced alignment, we utilized the CTC segmentation algorithm [28], which uses an end-to-end CTC-based phoneme recognition model to determine phoneme boundaries. For the CTC-based acoustic model for phoneme recognition, we utilized an open-source fine-tuned phoneme recognition Wav2vec2.0 model[1] based on [29].

2.4. Datasets

To address our RQ, we conducted experiments using two L2 English speech datasets: My Pronunciation Coach (MPC) [30] and SpeechOcean762 [31]. MPC comprises recordings of Dutch children speaking English, while the latter includes speech from Mandarin-speaking adults and children. Given its high acoustic variability—resulting from L1 transfer and inconsistent phoneme realizations [32]—non-native children's speech was a primary focus, as it presents a particularly challenging testbed for pronunciation assessment.

MPC [30] contains speech from 124 Dutch secondary school students. Each recording includes 53 English words and 53 sentences covering a wide range of phonemes. Recordings are categorized into quality groups: Excellent, OK, Doubtful and Overloud. For this study, we used 50 OK and 21 Excellent sessions, totalling 3,130 utterances from 71 speakers (38 males, 33 females). Since MPC lacks annotated mispronunciations, we introduced simulated pronunciation errors by modify-

[1] https://huggingface.co/facebook/wav2vec2-xlsr-53-espeak-cv-ft

ing phoneme sequences. Common substitutions include replacing /ð/ → /d/, /θ/ → /s/, /æ/ → /e/, and diphthong simplifications such as /eɪ/ → /eː/.

SpeechOcean762 [31] is an open-source corpus for pronunciation assessment, containing 5,000 English utterances from 250 native Mandarin speakers (125 adults, 125 children). Each utterance is annotated by five experts at the sentence, word, and phoneme levels, with 3,401 phonemes labelled as mispronunciations. We used all 5,000 utterances in our experiments.[2]

2.5. Evaluation Metrics

We assessed model performance using accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC). Given the class imbalance in both datasets, we optimized the GOP threshold by selecting the percentile that maximized MCC. Additionally, we reported the ROC AUC score at this threshold to evaluate classification effectiveness.

In addition, in order to analyze GOP score distributions, we used violin plots to compare posterior probability-based, logit-based, and hybrid GOP scores across correct and mispronounced phonemes. This visualization helps determine whether a given GOP scoring method provides a clear phoneme separation, a key factor in pronunciation assessment.

The Speechocean762 dataset includes human-annotated phoneme accuracy scores. Following prior research [31], we applied a second-order polynomial regression to model the relationship between GOP and human phoneme accuracy ratings. Performance was evaluated using Pearson Correlation Coefficient (PCC) and Mean Squared Error (MSE) to quantify prediction accuracy on the test set. Finally, phoneme-level mispronunciation error rates of the Speechocean762 dataset were also analyzed using a bar plot, comparing the GOP method with the highest PCC correlation to human-rated phoneme accuracy.

3. Results

Table 1 presents the evaluation scores for posterior probability based (first column), logit-based (second to fourth columns) and hybrid GOP scores (last column) on the MPC dataset. Of all these measures, $\text{GOP}_{\text{Margin}}$ achieves the highest accuracy (0.851), MCC (0.347). It also outperforms other approaches in F1-score (0.415) and precision (0.347), ensuring better mispronunciation detection while maintaining precision. However, $\text{GOP}_{\text{MaxLogit}}$ achieves the highest AUC at MCC_{max} (0.736), making it the most effective in distinguishing correctly pronounced and mispronounced phonemes. The measure GOP_{DNN} shows the highest recall (0.929) but has the lowest precision (0.184), indicating it detects most mispronunciations but lacks specificity. $\text{GOP}_{\text{MaxLogit}}$ shows moderate performance, achieving an accuracy of 0.590 and an MCC of 0.286. Its performance is slightly better than GOP_{DNN} but still lower than $\text{GOP}_{\text{Margin}}$ in most metrics. The $\text{GOP}_{\text{Combined}}$ score achieves an accuracy of 0.82 and MCC of 0.314, showing a good balance. Its AUC at MCCmax (0.704) is competitive, suggesting it may be a promising hybrid approach.

Table 2 shows the evaluation results of different GOP-based pronunciation assessment scores of the SpeechOcean762 dataset. This table also includes PCC and MSE scores since the SpeechOcean762 dataset includes human-annotated phoneme accuracy ratings, allowing us to evaluate the correlation between GOP and expert scores. GOP_{DNN} achieves the highest accuracy (0.947), precision (0.333), and MCC (0.367), showing that it effectively separates correctly pronounced and mispronounced phonemes. However, its PCC scores (0.278 for low confidence and 0.295 for high confidence) are significantly lower than other logit-based approaches. This suggests that while GOP_{DNN} can classify pronunciation errors well, it does not align well with human-annotated phoneme scores, making it less reliable for subjective pronunciation assessment. $\text{GOP}_{\text{MaxLogit}}$ achieves the highest PCC scores (0.442 for low confidence and 0.456 for high confidence), outperforming all other GOP metrics in correlating with human ratings. This indicates that maximum logit values capture pronunciation quality in a way that aligns better with human perception compared to posterior probability-based GOP. It also achieves a strong AUC at MCC_{max} (0.754), reinforcing its reliability as a GOP metric. $\text{GOP}_{\text{Margin}}$ shows the weakest overall performance, with an MCC of only 0.174 and lower PCC scores (0.173 for low confidence and 0.191 for high confidence).

Table 1: *Performance analysis on the MPC dataset*

	GOP_{DNN}	$\text{GOP}_{\text{MaxLogit}}$	$\text{GOP}_{\text{Margin}}$	$\text{GOP}_{\text{VarLogit}}$	$\text{GOP}_{\text{Combined}}$
Accuracy	0.572	0.590	**0.851**	0.461	0.820
Precision	0.184	0.189	**0.347**	0.146	0.297
Recall	**0.929**	0.919	0.515	0.882	0.559
F1	0.307	0.314	**0.415**	0.250	0.388
MCC	0.279	0.286	**0.342**	0.184	0.314
AUC MCC_{max}	0.730	**0.736**	0.702	0.648	0.704

Table 2: *Performance analysis on the SpeechOcean762 dataset*

	GOP_{DNN}	$\text{GOP}_{\text{MaxLogit}}$	$\text{GOP}_{\text{Margin}}$	$\text{GOP}_{\text{VarLogit}}$	$\text{GOP}_{\text{Combined}}$
Accuracy	**0.947**	0.925	0.741	0.894	0.843
Precision	**0.333**	0.257	0.089	0.195	0.139
Recall	0.466	0.571	**0.672**	0.621	0.642
F1	**0.388**	0.354	0.157	0.297	0.228
MCC	**0.367**	0.350	0.174	0.308	0.247
AUC MCC_{max}	0.715	**0.754**	0.708	0.763	0.747
PCC (low conf)	0.278	**0.442**	0.173	0.341	0.303
PCC (high conf)	0.295	**0.456**	0.191	0.357	0.319
MSE	0.124	**0.109**	0.131	0.120	0.123

While margin-based GOP obtained most of the best metric scores in the MPC dataset, it does not generalize well to SpeechOcean762, possibly due to differences in speaker demographics and phoneme variability. In Table 2 we see that $\text{GOP}_{\text{VarLogit}}$ achieves high recall (0.621) but has weak MCC (0.308) and low precision (0.195), suggesting it works well for detecting mispronunciations but lacks robustness in classification. $\text{GOP}_{\text{Combined}}$ achieves an MCC of 0.247 and PCC scores (0.303 for low confidence and 0.319 for high confidence), indicating a balance between probability and logit-based features.

To better interpret the performance results, Figure 2 visualizes the distribution of correctly pronounced and mispronounced phonemes across both datasets. GOP_{DNN} shows a wide distribution overlap between correct and mispronounced phonemes in both cases (both graphs of Figure 2), with close means and high variance, making it less effective for error distinction. In contrast, $\text{GOP}_{\text{MaxLogit}}$ achieves better separation, especially in MPC (first graph of Figure 2), though overlap increases in SpeechOcean762 (second graph of Figure 2), leading to higher variability. The best-performing score in MPC, $\text{GOP}_{\text{Margin}}$, effectively classifies children's speech (first graph of Figure 2), but struggles with greater distribution overlap in SpeechOcean762 (second graph of Figure 2). $\text{GOP}_{\text{VarLogit}}$ shows high variability and significant overlap in both cases, making it unreliable. $\text{GOP}_{\text{Combined}}$ balances probability- and

[2] https://github.com/Aditya3107/GOP_logit.git

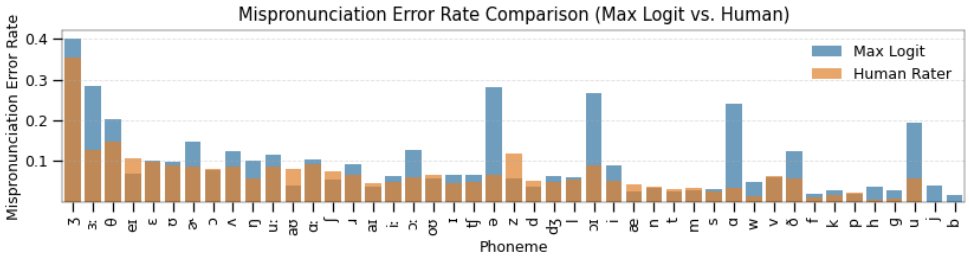

Figure 1: *Comparison of mispronunciation error rates by phoneme (Max Logit vs. human rater) in the SpeechOcean762 dataset. Blue bars show $GOP_{MaxLogit}$-predicted error rates, while red bars indicate human-rated phoneme accuracy.*

logit-based approaches across both datasets, reducing overlap compared to GOP_{DNN}. It presents a trade-off between classification accuracy from posterior probabilities and human correlation from logit-based scores. While we incorporated GOP_{Margin} in $GOP_{Combined}$, $GOP_{MaxLogit}$ could also be considered, leaving the optimal choice undecided.

Figure 2: *Comparison of GOP score distributions across MPC and Speechocean762 datasets.*

Finally, to analyze the alignment between GOP-based mispronunciation detection and human-rated phoneme accuracy, we investigated whether the GOP scoring method with the highest correlation to human ratings, $GOP_{MaxLogit}$, effectively identifies mispronounced phonemes. Figure 1 compares phoneme-level mispronunciation error rates predicted by $GOP_{MaxLogit}$ with human annotator ratings from the SpeechOcean762 dataset. Discrepancies were further examined by computing the difference in mispronunciation rates between $GOP_{MaxLogit}$ predictions and human ratings, highlighting phonemes where the model was overconfident, underconfident, or well-aligned (Figure 1).

Phonemes where the $GOP_{MaxLogit}$ overestimates mispronunciations (meaning the model assigns significantly higher mispronunciation rates than human raters) include (top 5): /ə/, /ɑ/, /ɔɪ/, /ɜː/, and /u/. These phonemes are frequently flagged as mispronounced by the model, despite human raters considering them correctly pronounced in most instances. This suggests that $GOP_{MaxLogit}$ is overly confident for these phonemes, possibly misinterpreting slight articulatory variations as errors. Conversely, phonemes where $GOP_{MaxLogit}$ *under*estimates mispronunciations (bottom 5), meaning human raters perceive more errors than the model detects, include /æ/, /ʃ/, /eɪ/, /aʊ/, and /z/. Some phonemes exhibit strong agreement between $GOP_{MaxLogit}$ predictions and human-rated error rates, indicating well-aligned phoneme classification. These phonemes are /b/, /tʃ/, /iː/, /dʒ/, and /ɑː/.

4. Discussion and Conclusion

In this work, we have analyzed differences in probability-based and logit-based GOP for pronunciation assessment across two datasets, MPC and SpeechOcean762. To answer our RQ, our findings indicate that logit-based methods achieve a better classification performance than probability-based GOP; however, their effectiveness depends on the characteristics of the dataset. GOP_{DNN} consistently obtains high recall, but low precision, indicating its tendency to over-detect mispronunciations, as seen in its high overlap between correctly pronounced and mispronounced phonemes. In contrast, logit-based methods ($GOP_{MaxLogit}$) demonstrate better phoneme separation (Fig. 2).

A key insight is that $GOP_{MaxLogit}$ aligns best with human ratings, achieving the highest PCC scores (Table 1), while GOP_{DNN}, despite strong classification performance, does not correlate well with expert judgments (Table 2). This suggests that maximum logit values better capture pronunciation quality from a perceptual standpoint. $GOP_{VarLogit}$ shows high variability (Table 1 and 2), making it less reliable, while $GOP_{Combined}$ balances probability and logit-based information but still shows some overlap.

To the best of our knowledge, the highest PCC score reported on the SpeechOcean762 dataset is 0.69 [33]. However, this was achieved using a multidimensional MDD model that calculates GOP scores while incorporating additional aspects of speech in the SpeechOcean762 dataset. In contrast, our approach focuses solely on logit-based GOP scoring, making direct comparisons with these other methodologies challenging. Future research should focus on reducing reliance on forced alignment, which can introduce errors due to acoustic variability in child and non-native speech, contributing to high recall but low precision.

In conclusion, our logit-based methodology offers a model-agnostic framework for any CTC-based acoustic model using a threshold-based approach. However, results vary across datasets, with GOP_{Margin} performing best on MPC and $GOP_{MaxLogit}$ on SpeechOcean762, underscoring the role of acoustic variability.

5. 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).

6. References

[1] A. Kuschel, N. Hansen, L. Heyse, and R. P. Wittek, "Combining language training and work experience for refugees with low-literacy levels: a mixed-methods case study," *Journal of International Migration and Integration*, vol. 24, no. 4, pp. 1635–1661, 2023.

[2] R. Walker, E.-L. Low, and J. Setter, "English pronunciation for a global world," Oxford, October 2021, Last visited: 2025-02-10. [Online]. Available: https://centaur.reading.ac.uk/101017/

[3] J. Jenkins, *The phonology of English as an international language*. Oxford University Press, 2000.

[4] J. Zoss, "What do adult english learners say about their pronunciation and linguistic self-confidence?" *MinneTESOL Journal*, 2016.

[5] H.-W. Hsu, "An examination of automatic speech recognition (asr)-based computer-assisted pronunciation training (capt) for less-proficient efl students using the technology acceptance model," *International Journal of Technology in Education*, vol. 7, no. 3, pp. 456–473, 2024.

[6] M. Amrate and P. hua Tsai, "Computer-assisted pronunciation training: A systematic review," *ReCALL*, no. 1, pp. 22–42, 2024.

[7] N. Alrashoudi, H. Al-Khalifa, and Y. Alotaibi, "Improving mispronunciation detection and diagnosis for non-native learners of the arabic language," *Discover Computing*, vol. 28, no. 1, p. 1, 2025.

[8] X. Cao, Z. Fan, T. Svendsen, and G. Salvi, "A Framework for Phoneme-Level Pronunciation Assessment Using CTC," in *Interspeech 2024*, 2024, pp. 302–306.

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

[10] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study," in *Speech and Language Technology in Education (SLaTE 2009)*, 2009, pp. 49–52.

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

[12] 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, 2010, pp. 598–602.

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

[14] H. Do, W. Lee, and G. G. Lee, "Acoustic feature mixup for balanced multi-aspect pronunciation assessment," *CoRR*, vol. abs/2406.15723, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2406.15723

[15] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7262–7266.

[16] Y. El Kheir, "Mispronunciation detection with speechblender data augmentation pipeline," PhD Thesis Report, KTH Royal Institute of Technology, Stockholm, Sweden, 2023, Last visited: 2025-02-10. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-339940

[17] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, 2022, pp. 2278–2282.

[18] 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 Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451–3460, Oct. 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

[19] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities." in *INTERSPEECH*, vol. 2, 2019, pp. 954–958.

[20] G. Pereyra, G. Tucker, J. Chorowski, L. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," 2017. [Online]. Available: https://openreview.net/forum?id=HkCjNI5ex

[21] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, "Mitigating neural network overconfidence with logit normalization," in *International conference on machine learning*. PMLR, 2022, pp. 23 631–23 644.

[22] X. Xie and T. F. Jaeger, "Comparing non-native and native speech: Are l2 productions more variable?" *The Journal of the Acoustical Society of America*, vol. 147, no. 5, pp. 3322–3347, 2020.

[23] J. L. Preston, J. R. Irwin, and J. Turcios, "Perception of speech sounds in school-aged children with speech sound disorders," in *Seminars in speech and language*, vol. 36, no. 04. Thieme Medical Publishers, 2015, pp. 224–233.

[24] X. Li, X. Li, D. Pan, and D. Zhu, "On the learning property of logistic and softmax losses for deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 4739–4746.

[25] J. Weng, Z. Luo, S. Li, N. Sebe, and Z. Zhong, "Logit margin matters: Improving transferable targeted adversarial attack by logit calibration," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 3561–3574, 2023.

[26] E. J. Yeo, K. Choi, S. Kim, and M. Chung, "Speech intelligibility assessment of dysarthric speech by using goodness of pronunciation with uncertainty quantification," in *Interspeech 2023*, 2023, pp. 166–170.

[27] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)," in *Interspeech*, 2013, pp. 1886–1890.

[28] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "Ctc-segmentation of large corpora for german end-to-end speech recognition," in *International Conference on Speech and Computer*. Springer, 2020, pp. 267–278.

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

[30] 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.

[31] 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*, 2021, pp. 3710–3714.

[32] R. Gretter, M. Matassoni, D. Falavigna, A. Misra, C. Leong, K. Knill, and L. Wang, "Etlt 2021: Shared task on automatic speech recognition for non-native children's speech," in *Interspeech 2021*, 2021, pp. 3845–3849.

[33] 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*, 2023, pp. 974–978.