Interspeech 2024
1-5 September 2024, Kos, Greece

A Dataset and Two-pass System for Reading Miscue Detection

Raj Gothi[1], *Rahul Kumar*[2], *Mildred Pereira*[2], *Nagesh Nayak*[2], *Preeti Rao*[1,2]

[1]Centre for Machine Intelligence and Data Science, IIT Bombay
[2]Department of Electrical Engineering, IIT Bombay
{22m2160,22m1163,mildredp,nageshnayak,prao}@iitb.ac.in

Abstract

Automatic speech recognition (ASR) has long been viewed as a promising solution to the resource-intensive task of oral reading fluency assessment. The demands on ASR accuracy, however, tend to be high, especially when applied to obtaining reliable reading diagnostics. The prior knowledge of reading prompts is typically used to limit the system WER. The accurate detection of mispronounced words, which can be relatively few in number, while limiting false positives, remains challenging. In this work, we present a new manually transcribed dataset of 1,110 elementary school children reading connected text in L2 English with wide-ranging proficiencies. Apart from local features derived from alternate decodings under different linguistic context constraints, we use an additional deep acoustic model. We discuss the performance gains achieved in a second pass over initial hybrid ASR hypotheses.
Index Terms: reading miscues, automatic assessment, speech recognition

1. Introduction

Foundational literacy has been a target of heightened attention since 2022, when global coalitions were formed to support country-led action to address the learning crisis [1]. The monitoring of outcomes and generation of learning data are among the key components of the effort, making objective and scalable methods for literacy measurement highly desirable. Even as the estimated words read correctly per minute (WCPM) is utilised to obtain fluency benchmarks, the precise nature of the reading errors is necessary to determine a child's reading level and the appropriate intervention. Motivated by the time and resource intensive nature of manual assessments, research in the automatic evaluation of oral reading has been pursued for close to three decades by dedicated academic groups in Europe and the U.S. This has given rise to children's oral reading corpora for isolated words, sentences and, to a lesser extent, connected texts, in languages such as Dutch, French, Italian, English and Portuguese by between 40 to 500 speakers [2–5]. Closest to the dataset presented in this work, in terms of type of reading prompts and number of speakers, is CHOREC, comprising of 300 Dutch-speaking children of Grades 1-4 reading stories [6].

In the related field of CAPT (computer-aided pronunciation training), the focus has been on detecting phonetic errors in the adult second language learning context, as opposed to word reading miscues [7]. Hybrid ASR systems have been the dominant choice of solution for the oral reading accuracy task, given the ease of integrating the acoustic model with a language model (LM) of specific context constraints. Constraining the ASR with the knowledge of the reading prompts (via a task-specific LM) is helpful in limiting the WER and therefore improving the prediction of correctly uttered words, followed possibly by a rescoring of the segmented words using local features derived using less constrained LMs. To briefly review representative earlier work, Bolanos et al. [8] used a GMM-HMM system trained on children's speech corpora to evaluate reading fluency of 313 students of Grades 1-4 reading text passages using a trigram LM for each passage. They obtained a high correlation for the estimated words correct per minute (WCPM) with human raters but did not report the miscue detection accuracy. Duchateau et al. [9] use the careful design of an FST with all expected text-dependent reading miscues. Cheng et al. [10] used a Kaldi DNN-HMM system with an item-specific LM to achieve the reliable prediction of WCPM. Proenca et al. [11] applied a second pass to the decoded output from a DNN-HMM ASR to flag reading miscues based on GOP-like features computed on word segments. They showed a further improvement with the use of phone edit distance with reference to the output of free phone recognition, although this was limited by the accuracy of phone recognition.

Very recently, Molenaar et al. [12] compared Kaldi TDNN systems trained on adult speech combined with 4 different LMs and two general-purpose Whisper-based ASR systems (without and with prompts). The Kaldi system outperformed Whisper, particularly when an LM enriched by the manually transcribed reading was used. Similarly Piton et al. [13] evaluated a few commercial ASR systems on French and Italian children's reading assessment to conclude that the analysis results are not sufficiently fine-grained and recommend a second pass of the ASR output. The challenges lie in that reading miscues (typically a small fraction of the total words read) must be reliably flagged while limiting false positives. The accuracy of identifying mispronounced words tends to be poor with reference to manually detected reading errors due to the mostly out-of-vocabulary substitutions made by beginning readers who are not yet fully familiar with the letter-to-sound rules of the language.

In this work, we present a dataset[1] that considerably enriches the available set of children's oral reading corpora with its large number of unique speakers reading connected text in L2 English, manually transcribed and labeled for use in reading miscue studies. We test the performance of a hybrid ASR, the Kaldi based TDNN (a type of DNN-HMM acoustic model or AM), trained on adult Indian English speech, with two distinct LM constraints. Taking forward the LM that is constrained only on the reading prompts (i.e. the canonical text), as the scenario most commonly arising in practice, we present results for a variety of second-pass local features, and their fusion, that serve to enhance the miscue detection performance, measured in terms of miscue miss-rate for a fixed false positive rate (FPR). Apart

[1]https://github.com/DAP-Lab/mps_dataset

10.21437/Interspeech.2024-1639

from the widely used goodness of pronunciation (GOP), we experiment with phonetic edit distance with the use of wav2vec2 XLSR (pretrained via self-supervised learning (SSL) on 53 languages) and fine-tuned by us to produce relatively accurate phonetic transcripts. In the next section, we describe our new test dataset, as well as the training datasets used in this work, followed by the presentation of the methods and experiments.

2. Datasets

As part of a benchmarking exercise for reading levels in elementary school, audio recordings of oral reading of L2 English grade-appropriate texts were collected for close to 2,500 students in 10 government schools across the Indian states of Maharashtra and Goa in Grades 3, 4 and 5 (age 7-11 years) in the summer of 2023. Ethics clearance was obtained for the audio recording with anonymised speaker information but for grade and gender. The students, who come with diverse home languages, are introduced to both Hindi and English reading and writing in Grade 1. Unlike L1 English learners of reading, they have a non-existent (or very limited) vocabulary for English, and are therefore encountering new words and their written forms simultaneously. They cannot draw on any aural memory to sound the written form and hence make a variety of errors in pronunciation owing to the opaque orthography of English as well as phonotactic constraints of their home language [14]. For instance, the word 'shepherd' in the Grade 3 text is uttered as 'seperd' or 'sheep hard' among other variations, and 'bears' as 'beers' in Grade 4. Although we find some L1 accent-caused variations as well, we compensate for this with our phoneme-based lexicon with alternate pronunciations when applicable.

The text prompts comprise of the 2 paragraphs of a single story, each between 60-80 words, with a unique story assigned to each grade. Manual transcription at the word level is underway. The reading errors are observed to comprise a number of non-English words, which are transcribed phonemically. Of the already transcribed dataset, we select for this work those utterances (i.e. the recording of one story paragraph) where at least 70% of the canonical words have been attempted. These amount to 1600 utterances across 1110 unique speakers, distributed close to uniformly across the 3 grades, for a total audio duration of 19 hours. We term this the MPS (for Maharashtra Primary Schools) dataset. The background noise varies from barely audible to classroom noise in the vicinity of the speaker. This dataset serves as the test dataset in this work with ground-truth miscue labels (Cor/Sub/Del/Ins) derived by comparison of the manual transcription with suitably aligned text prompts (using the phonetically oriented alignment discussed in Sec. 3.1).

Table 1 shows the distribution of attempted words across the dataset of 1600 prompts, where we note that about 10% of the total attempted words constitute reading miscues (i.e. DEL + SUB, including different types of SUB), while the per-utterance miscues range from 0 to 24 words. The manual transcription is detailed enough to label the type of reading disfluency. The test dataset is divided into 6 folds with non-overlapping speakers and uniform distribution of grades and reading levels (in terms of number of miscues per utterance) to facilitate the cross-validation (CV) training and testing of miscue detection classifier of the second pass.

Our AM and LM training datasets in this work summarised in Table 2 are (i) IITM: Indian English adult speech [15], (ii) WAP: Indian children (11-15 age group) reading a variety of English text prompts, collected by us over WhatsApp voice messaging during the pandemic (2020-2021). This dataset brings in the context of read speech in children's voices (even if the group is older in age compared with the MPS test data cohort).

Table 1: *Distribution of word reading errors across the 1,10,898 attempted words in MPS dataset*

Tags	Number (%)	Description
COR	99,554 (89.77%)	Prompt word is correctly pronounced
INS	3269 (2.94%)	Inserted word, not part of the prompt
DEL	1331 (1.20%)	Prompt word omitted
SUB-1	3452 (3.11%)	Substitution: one-phone difference
SUB-2	6561 (5.91%)	Gross substitution (>1 phone variation)

3. Method

Our two-pass system incorporates a hybrid ASR in the first pass with a strongly constrained LM, i.e. word trigram trained on the text prompts. Anticipating the consequent low recall of miscues, the second pass re-evaluates the canonical words labeled 'correct' in Pass 1 via local features computed on the corresponding acoustic segment. In this section, we describe the initial segmentation and labeling achieved by the first pass hybrid ASR and the computation of features for the second pass.

3.1. Segmentation and alignment with text prompt

The hybrid system is based on the 'nnet3' Time Delay Neural Network (TDNN) (no-chain model, which is more suited to GOP computation than is the chain model [16]) trained with the Kaldi Librispeech recipe on the IITM dataset. Frame labels for TDNN model training were obtained by forced alignment using a GMM-HMM model trained beforehand. The number of modelled phones were 40 (39 speech and a silence phone). For the LM, we use 3-gram models on training text as detailed in the experiments section.

For the reading assessment task, the alignment of the prompt words with the ASR hypothesis words (or with the manual transcript) needs to be more nuanced than that achievable with conventional word-level edit distance, i.e. where exact match is considered and errors comprise substitutions, deletions, or insertions, based on the Levenshtein distance. In order to predict the precise types of reading miscues accurately and also obtain the correct word boundaries for the subsequent stage of local feature computation, the ASR output sequence of words needs to be aligned with the reading prompt sequence of words in a manner that is phonetically informed. This exploits the expected phonetic similarity between the uttered word and attempted prompt word. We modify the alignment algorithm of Ruiz et al. [17] to accommodate the peculiarities of the reading

Table 2: *Datasets used in this work (all Indian English).*

Dataset	Dur. (hours)	Speakers	Type (unique words)
IITM [15]	184	1585	Adult speech: read, conversational (23103)
WAP	42	1164 (Grades 6-8)	Child speech: read (8093)
MPS	19.1	1110 (Grades 3-5)	Child speech:read

application such as the occurrence of word splitting or merging. False starts, a common event in oral reading, are accommodated with bias towards right alignment. We also account for the possibly multiple valid pronunciations of a prompt word. Eventually, we associate a decoder output word with each canonical word (unless it is found deleted by the alignment algorithm). The canonical word is labeled Correct if it matches the decoder word and Substituted otherwise. Insertions are ignored in oral reading accuracy measurement.

3.2. Word-level features for miscue detection

From the Pass 1, we obtain the acoustic word boundaries associated with each canonical word labeled correct. Next, local features as described below are extracted. In the Pass 2, we control the trade-off between miss-rate and FPR by applying a threshold to the feature to relabel those ASR-predicted Correct words with score below the threshold as Substitutions.

3.2.1. Lattice-based confidence scores

Lattice based confidence scores are obtained from the word graph generated during the decoding of the utterance. The word graph serves as a condensed representation of the hybrid decoder search space. Word posterior, likelihood ratio test (LRT) and hypothesis density are different measures of the dominance of the decoded word over the alternative hypotheses in the word graph, and represent decoder confidence [18].

3.2.2. Goodness of Pronunciation (GOP)

GOP is an established approach to segmental error detection [19], that measures the acoustic quality of phone realization by its posterior probability, which is subsequently normalized and aggregated to get a word-level pronunciation score. A GOP score for a specific phone segment is calculated by taking the difference of the log probability of the forced alignment of the hypothesis and the log probability of a less constrained recognition phase. Common word-level aggregates are across-phones minimum and mean, and across-frames mean (which weights the phones by duration). We use the Kaldi implementation of GOP obtained from DNN AM senone probabilities [20]. GOP, while being highly effective, is also limited due to the considerable overlap of values across poor and correct pronunciations [21].

3.2.3. Phone recognition features

Phone recognition, achieved with a less constrained LM such as a phone bigram, is a promising approach to identifying reading miscues corresponding to substitutions that are otherwise masked by the word decoding with strongly constrained LMs. The mismatch between the uttered word and the corresponding canonical word pronunciation is measured by the Levenshtein edit distance between the two phone sequences [11]. However, this requires reliable phone recognition which is typically hard to achieve. We consider using a phone bigram LM with the hybrid AM to obtain the phone sequence of the utterance under weak LM constraints. Further, based on the growing research on SSL pretrained models that shows their effectiveness in capturing the characteristics of basic acoustic units more accurately, and yielding correspondingly superior PER in within- and cross-language scenarios [22], we investigate wav2vec2 based phone recognition. The pretrained model is fine-tuned on our Indian English training datasets using phone alignments of the manual transcripts as obtained from the hybrid ASR system. We added a linear layer on top of the wav2vec2-model, aimed at mapping the final encoder layer's 1024 dimensions to 46 tokens. These tokens include all the phones, special tokens, and an additional token (*) denoting word boundary, which serves to group phone sequences within words. This facilitates the comparison of the corresponding word-level phone sequences.

4. Experiments and Results

We evaluate our first-pass system, followed by the gains achieved for each of the considered second-pass features applied to the decoder segments labeled Correct in the first pass. Finally, we report the performance with the fusion of features via a trained classifier. Performance is reported in terms of miscue detection rate at FPR = 5%. We show DET curves (suitable for imbalanced data like ours, [23]) in Figure 1 and also report the AUC (area under the ROC) as another measure of performance.

4.1. First pass performance

We obtain on our test set, the WER with the two LMs: word trigram trained on (i) manual transcript of the test data, and (ii) canonical text prompts only. In all cases LM interpolation is applied. The case (i) has the prior knowledge of the reading errors and therefore provides an upper bound on the first pass performance as a reference. The case (ii) is the one expected in practice. From Table 3, we see a close to 3% absolute increase in WER in the second case and a much larger increase in the miss rate with most reading miscues getting decoded as the attempted canonical word (which happens to match the observations of [12]). Our aim in this work is to process the first pass output in case (ii) to improve the recall of miscues to the extent possible while limiting the FPR to a reasonable value.

Table 3: *WER and miscue detection performance of Hybrid ASR with different LM constraints on the test dataset.*

Pass 1 LM	WER	Miss-Rate	FPR
Trigram on Transcripts	10.17%	0.21	0.046
Trigram on Prompts	13.01%	0.72	0.011

4.2. GOP and lattice based features

Table 5 shows the performance of the word-level features derived from the hybrid decoder outputs. Of the confidence scores, the posterior performs the best. The GOP measures (including the related LLR measure [20]) do much better, with minimum phone GOP being the best word level feature.

4.3. Phone recognition features

We start with a comparison of the phone error rate (PER) of the hybrid and wav2vec2 systems under matching training data to the extent possible (given that the latter is already SSL pretrained). We use both the training datasets to fine-tune the wav2vec2 using the CTC loss function [24]. The CNN feature encoder layer was frozen, while all transformer encoder layers were fine-tuned and supplemented with an additional linear layer. The fine-tuning procedure consisted of 2 epochs using the IITM dataset, followed by an additional 10 epochs using the WAP dataset. The hybrid system is trained with a phone bigram LM. We test it with the original IITM trained AM, as well as a

further trained version using the WAP dataset in order to make the training data used comparable with that for the wav2vec2.

Table 4: *PER on MPS dataset for Hybrid and Wav2vec2 systems with various context constraints and training datasets, where Hybrid ASR's LM phone-bigram is trained on WAP transcripts*

Model (Training data)	Test PER(%)
Hybrid (IITM AM, WAP LM)	24.73
Hybrid (IITM + WAP AM, WAP LM)	18.15
Wav2Vec2-XLSR (IITM+WAP)	9.67

Table 4 shows the PER on the test set of the phone recognition by the hybrid system and wav2vec2. The wav2vec2 provides 20 ms frame-level logits. The phone sequence is obtained from the non-blank symbol frames across the word using the boundaries obtained in the first pass hybrid decoder. We note that the PER is substantially lower with the wav2vec2 decoding with the same training data, reinforcing past observations that pretrained models learn better discriminant representations in fine-tuning compared to the same applied to the TDNN hybrid [25].

We observed that the wav2vec2 decoded phone sequence was by and large an accurate match to the pronunciation as recorded in the manual transcript. Errors sometimes appear in terms of short subsequences of ground-truth phones replaced entirely with blank symbols in the wav2vec2 output. It is difficult to explain these errors in view of the black-box nature of the model and its weak and implicit language context constraints. Given the observed superior PER of wav2vec2, we employ this output for the phone edit distance feature. Apart from the simple Levenshtein distance (Lev), we leveraged the phone confusion matrix obtained from the WAP training data to compute a cost-weighted phone distance (CostLev) where pairs of confusable phones received a lower substitution cost. Table 5 shows the resulting decrease in the miss rate.

With a view to diminishing the influence of wav2vec2 phone recognition errors on system performance, we investigated the utility of confidence scoring, an area of active research for end-to-end (E2E) systems [26]. While the probability of the best hypothesis is a natural way of estimating confidence, its effectiveness is limited when the probability distribution is skewed towards the best hypothesis (the prediction overconfidence of E2E). We therefore normalize the raw probability values of wav2vec2 frame outputs using temperature scaling and then summarise at frame-level with one of log-max or negative entropy [27]. Finally, a word-level confidence is computed by summing frames across the word. Table 5 shows that the phonetic distance clearly benefits from the confidence measures, with temperature-scaled entropy sum being most successful.

4.4. Feature fusion

In the interest of combining features, we use the logistic regression classifier with balanced bagging for ensembling. The MPS dataset was tested in 6-fold CV to obtain the results reported for feature fusion in Table 5. We note that feature fusion is clearly helpful in lowering miss-rate over any one feature category.

5. Summary and Conclusion

Figure 1 summarises the performance of the miscue detection methods of this work on our new dataset. The leftmost starting point of the curves is the achieved metric of the hybrid system employed for segmentation and initial labeling of canonical words. Our goal was to explore other attainable performance points in the 2d space with the most favourable trade-off between miss-rate and FPR. The posterior, directly derived from the hybrid system word lattice, gives a potential trajectory but one that falls too slowly with increase in FPR. This is not surprising given that the word lattice is expected to be very sparse and therefore not informative in the context of the strongly constrained LM of reading prompts. We see that GOP improves the trade-off significantly. The similar, or slightly superior, performance is delivered by the wav2vec2 based phonetic distance, especially when accompanied with a confidence score computed on the scaled probabilities. Further, that the features actually contain complementary information about the local acoustic characteristics is borne out by the clear superiority of the fusion curves.

Future work could explore the deeper integration of SSL pretrained models with hybrid networks for this task where both word-level segmentation of the utterance and phone recognition accuracy play equally critical roles.

Table 5: *Miss rate at FPR=5% and AUC for features across different classes and their combinations. Bold font: best in class, (R: raw; T: temperature, L: logmax, E: entropy; S: sum)*

Feature	Miss rate(%)	AUC
Lattice-based		
LRT	61.04	0.670
Hypotheses Density	58.37	0.721
Posterior	**58.14**	0.712
GOP-based		
LLR	49.80	0.874
Mean GOP	47.01	0.886
Min GOP	**39.66**	0.896
Phone edit distance (wav2vec2)		
Lev	40.74	0.890
CostLev	39.22	0.894
CostLev + RLS	39.39	0.919
CostLev + RES	38.63	0.922
CostLev + TLS	37.02	0.923
CostLev + TES	**36.84**	0.920
Feature fusion		
Fusion-best 3	32.36	0.928
Fusion-All	**31.07**	0.935

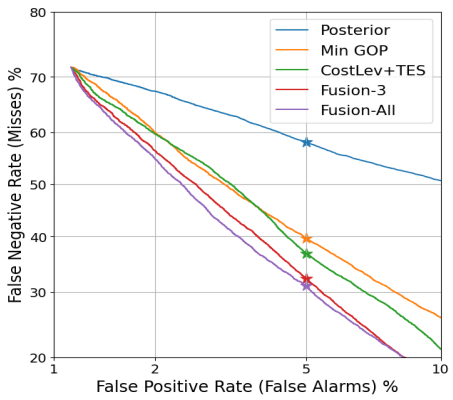

Figure 1: *DET curves from Pass 2 for selected features*

6. Acknowledgement

The authors gratefully acknowledge the Centre for Machine Intelligence and Data Science (C-MInDS) at the Indian Institute of Technology Bombay for financially supporting this research work.

7. References

[1] "The global coalition for foundational learning," 2022, https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2023/05/The-Global-Coalition-for-Foundational-Learning-Narrative.pdf.

[2] C. Cucchiarini and H. Van hamme, "The jasmin speech corpus: recordings of children, non-natives and elderly people," *Essential Speech and Language Technology for Dutch: Results by the STEVIN programme*, pp. 43–59, 2013.

[3] J. Proença, D. Celorico, S. Candeias, C. Lopes, and F. Perdigão, "The letsread corpus of portuguese children reading aloud for performance evaluation," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 781–785.

[4] M. P. Black, J. Tepperman, and S. S. Narayanan, "Automatic prediction of children's reading ability for high-level literacy assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 1015–1028, 2010.

[5] C. Hacker, A. Batliner, S. Steidl, E. Nöth, H. Niemann, and T. Cincarek, "Assessment of non-native children's pronunciation: Human marking and automatic scoring," *Proceedings of the 10th International Conference on SPEECH and COMPUTER*, vol. 1, pp. 123–126, 2005.

[6] L. Cleuren, J. Duchateau, P. Ghesquiere *et al.*, "Children's oral reading corpus (chorec): description and assessment of annotator agreement," *LREC 2008 Proceedings*, pp. 998–1005, 2008.

[7] B.-C. Yan, M.-C. Wu, H.-T. Hung, and B. Chen, "An End-to-End Mispronunciation Detection System for L2 English Speech Leveraging Novel Anti-Phone Modeling," in *Proc. Interspeech*, 2020, pp. 3032–3036.

[8] D. Bolaños, R. A. Cole, W. Ward, E. Borts, and E. Svirsky, "Flora: Fluent oral reading assessment of children's speech," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 4, pp. 1–19, 2011.

[9] J. Duchateau, M. Wigham, K. Demuynck, and H. van Hamme, "A flexible recogniser architecture in a reading tutor for children," in *Proc. ITRW on Speech Recognition and Intrinsic Variation*, 2006, pp. 59–64.

[10] J. Cheng, "Real-Time Scoring of an Oral Reading Assessment on Mobile Devices," in *Proc. Interspeech*, 2018, pp. 1621–1625.

[11] J. Proença, C. Lopes, M. Tjalve, A. Stolcke, S. Candeias, and F. Perdigao, "Mispronunciation detection in children's reading of sentences," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1207–1219, 2018.

[12] B. Molenaar, C. Tejedor-Garcia, C. Cucchiarini, and H. Strik, "Automatic Assessment of Oral Reading Accuracy for Reading Diagnostics," in *Proc. INTERSPEECH*, 2023, pp. 5232–5236.

[13] T. Piton, E. Hermann, A. Pasqualotto, M. Cohen, M. Magimai.-Doss, and D. Bavelier, "Using Commercial ASR Solutions to Assess Reading Skills in Children: A Case Report," in *Proc. INTERSPEECH*, 2023, pp. 4573–4577.

[14] M. Ellis, "Supporting young l2 english learners with word recognition: design of early reading materials," *Neofilolog*, vol. 59, no. 2, pp. 126–143, 2022.

[15] IIT Madras Speech Lab, "IIT-Madras Hindi-Tamil-English ASR Challenge," 2021, https://sites.google.com/view/indian-language-asrchallenge/home.

[16] K. group, "Discussion about failure of chain models for gop," 2020, "See jimbozhang's comment on May 22, 2020". [Online]. Available: https://github.com/kaldi-asr/kaldi/issues/3675

[17] N. Ruiz and M. Federico, "Phonetically-oriented word error alignment for speech recognition error analysis in speech translation," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 296–302.

[18] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech communication*, vol. 45, no. 4, pp. 455–470, 2005.

[19] 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.

[20] W. Hu, Y. Qian, and F. K. Soong, "An improved dnn-based approach to mispronunciation detection and diagnosis of l2 learners' speech." *SLaTE*, vol. 5, pp. 71–76, 2015.

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

[22] H. Ji, T. Patel, and O. Scharenborg, "Predicting within and across language phoneme recognition performance of self-supervised learning speech pre-trained models," 2022.

[23] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The det curve in assessment of detection task performance." in *Eurospeech*, vol. 4, 1997, pp. 1895–1898.

[24] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[25] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore wav2vec 2.0 for Mispronunciation Detection," in *Proc. Interspeech*, 2021, pp. 4428–4432.

[26] D. Oneaţă, A. Caranica, A. Stan, and H. Cucu, "An evaluation of word-level confidence estimation for end-to-end automatic speech recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 258–265.

[27] A. Laptev and B. Ginsburg, "Fast entropy-based methods of word-level confidence estimation for end-to-end automatic speech recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 152–159.