Orthography-based pronunciation scoring for better CAPT feedback

Caitlin Richter[1], *Ragnar Pálsson*[2], *Luke O'Brien*[3], *Kolbrún Friðriksdóttir*[4], *Branislav Bédi*[5], *Eydís Huld Magnúsdóttir*[6], *Jón Guðnason*[7]

[1,2,7]Reykjavik University, Iceland
[3,6]Tiro, Iceland
[4]University of Iceland, Iceland
[5]Árni Magnússon Institute, Iceland

{caitlinr,ragnarp,jg}@ru.is, {luke,eydis}@tiro.is, kolbrunf@hi.is,
branislav.bedi@arnastofnun.is

Abstract

We establish the viability of a streamlined architecture for pedagogically appropriate computer assisted pronunciation training (CAPT), to give second language learners automatic feedback about their mispronunciations. This takes advantage of end-to-end speech recognition models to detect mispronunciation in audio segments that correspond directly to orthographic letters, in contrast to standard mispronunciation detection using phone representations. Results in a classification task show the potential for similar sensitivity to non-nativelike phonetic errors in grapheme-aligned segments as in phone-aligned segments. Advantages of this approach over phone-based pronunciation scoring can include providing naturally comprehensible (orthographic, not phonemic) feedback to learners, being inherently open-vocabulary in the target language, and evaluating pronunciations with reference to a full range of target-language acoustic variants rather than a prespecified canonical phone sequence.

Index Terms: computer assisted pronunciation training, comprehensible feedback, forced alignment, phone segmentation, pronunciation error detection

1. Introduction

Computer assisted pronunciation training (CAPT) enables learners of a second language (L2) to become more fluent, comprehensible, and comfortable speaking, through self-study with automated pronunciation assessment [1]. Fine grained feedback on individual speech segments is particularly effective to direct learners' awareness and effort where it will be productive, in keeping with the Noticing Hypothesis [2] established in L2 pedagogy for corrective feedback in CAPT systems [3]. However, sub-word mispronunciation scoring is traditionally based on phones, as in hidden Markov model (HMM) speech recognisers [4], with drawbacks from the perspective of either the average learner or the CAPT designer who must transform the phone scores into more intelligible feedback [5].

We propose a streamlined architecture repurposing end-to-end automatic speech recognition (E2E ASR) to extract pronunciation information corresponding to orthographic letters. This is directly interpretable by literate learners [3, 5], and reduces needs for language-specific resources like pronunciation dictionaries. Evaluations in Norwegian and Icelandic establish feasibility of detecting non-nativelike pronunciations at similar sensitivity in both letter-based segments and phone segments, with either automatic or gold (human-annotated) phone alignments.

1.1. Related work

Two major challenges in sub-word pronunciation scoring are to achieve accurate scores for such short speech segments, and to deliver feedback that learners can understand and improve from.

1.1.1. Mispronunciation detection

Even word-level CAPT scoring has low accuracy. Summarising literature on L2 English speakers, Korzekwa et al. [1] find performance of '60% precision at 40%–80% recall', while their proposal to synthesise additional training speech improves area under the precision/recall curve (AUPR) to at most 0.75 for the task of classifying in/correctly pronounced L2 English words.

Phone-level error detection accuracy is even lower. Goodness of Pronunciation (GoP) exemplifies classical methods, segmenting a spoken word into its expected phone sequence with forced alignment before evaluating similarity of each phone to canonical acoustic models [4]. With further development to refine acoustic modelling [6, 7], representative best results include F1 score of 0.61 for L2 phone error detection in a system requiring relatively modest few hours of expensively labelled training data for fine tuning [8]. Recent methods include dedicated end-to-end pipelines, for example detecting whether a phone sequence decoded from a speech sample matches the sequence given in a pronunciation dictionary [9, 10, 11]. These approaches require increasing amounts of annotated training data, or those specifically aiming to reduce this problem struggle to match a traditional GoP baseline [10, 12, 13, 14, 15]. Alternatively, in previous work we applied dynamic time warping (DTW) to detect non-nativelike pronunciations using only unannotated parallel speech [16], finding area under the receiver operating characteristic curve (AUROC) of 0.88 for the proxy task of classifying phone segments as having been actually produced by a native (L1) or L2 Norwegian speaker.

1.1.2. Usable mispronunciation information

The sub-word mispronunciation detection methods reviewed above implement pronunciation scoring of phone segments,[1] either by forced alignment of the CAPT user's speech to a phone sequence from a pronunciation dictionary, or by directly decoding the speech to its best hypothesised phone sequence.

However, the resulting phone-level pronunciation error feedback is not intelligible to most L2 CAPT users, without specialist training in linguistics [5, 17]. Studies of CAPT pedagogy and learner outcomes indicate that it is often more effective to instead simply highlight (with underline, bold, and/or colour) the letter(s) that intuitively seem to correspond to the sound in which a mispronunciation is detected [2, 5, 3, 18].

From the technical implementation perspective, providing such readable feedback for learners therefore relies on grapheme-to-phoneme (g2p/p2g) conversion, which needs language-specific training data and/or handmade rules, and can

[1] Various scoring procedures are best described in terms of phonetic segments, phonemes, or neither exactly, so we refer generally to phones.

be prone to disruptive levels of error in languages without sufficient resources [3, 19, 20]. Even having reliable p2g input/output at word level does not necessarily imply access to phone-to-letter alignments sufficient to identify the 'mispronounced letters' when a phone pronunciation error is detected. For example, in *action* [æ k ʃ ə n], learners may benefit from attention on '*ac**ti**on*' or '*act**io**n*' when mispronouncing the third or fourth phones respectively. Learners who mispronounce Icelandic *fljúga* [f l j uː a] should receive different feedback for mispronouncing [uː] as [yː] '*flj**ú**ga*' than for inserting a consonant corresponding to the silent letter <g> '*fljú**g**a*'.

Solutions for some cases include: providing training for learners to interpret phone-based feedback; pre-specifying p2g alignments for a closed vocabulary; interpolating approximate alignments from the length of a word's phone sequence and letter sequence; applying alignment heuristics/constraints such as awareness of vowels and consonants; and accommodating for a limited set of anticipated errors (e.g. [f l j uː **g** a]) in an extended recognition network [3, 9, 17, 21, 22, 23, 24].

1.2. Approach and Contributions

All solutions for phone-based CAPT in §1.1.2 have drawbacks; E2E ASR models, like Wav2vec-2.0, could bypass the problem instead. Connectionist temporal classification (CTC) decoding labels frames of audio with characters from the model's output vocabulary, i.e. letters from a language's normal orthography. In this way, the pronunciation of the segment of audio aligned to each letter can be directly evaluated, and the evaluation naturally understood by language learners. While a similar effect may theoretically be derived in shallower ASR architectures using character-based/subword lexicons or byte pair encoding, Wav2vec2 brings a major step forward in performance across many tasks [25]. Many languages can be served by massively multilingual pretrained models like XLS-R [26]. The only strict requirement is a CTC decoding model with full coverage of the target language's alphabet, although best performance may be expected with language-specific decoder fine-tuning and this is necessary in any case for languages with globally unique characters (Icelandic 'þ'). Regardless, no pronunciation dictionary or g2p is required at any point, and unlike creation of those resources, CTC decoder finetuning requires only speech paired with orthographic transcriptions and is then inherently open-vocabulary; the potential advantages for development are clear.

However, time-aligning characters is an unanticipated use of E2E ASR models, where timing information for decoder output is normally irrelevant beyond ordering the characters correctly [27]. In other words, for conventional forced alignments as in §1.1.2, it is expected that the start and end times found for each phone are the times at which the speaker started and finished producing that phone; but CTC decoding carries no such clear expectation about the content of the speech aligned to a given letter. Therefore, it is necessary to empirically evaluate whether CTC segmentation with Wav2vec2 models could be the basis for intelligible fine-grained mispronunciation feedback.

We adopt the relative DTW approach developed in [16]. This provides an evaluation task applicable to Icelandic, which we aim to develop CAPT for, and has shown competitive results while requiring little, albeit specific, in-language data. Most importantly, unlike other methods in §1.1.1, the sub-word segment labels (phone/letter identities) are irrelevant to scoring; this feature means that learners are not constrained towards a few pre-specified dictionary pronunciations when several variants are actually acceptable, but it also facilitates our present experiment with a minimal controlled change in the implementation by performing forced alignment to letters instead of phones.

Contributions of this work include repurposing E2E ASR models for character-level time alignments in place of phone alignments; evaluating the method of [16] in a realistic rather than idealised CAPT scenario, i.e. a crowdsourced speech corpus with automatic segment alignments; and facilitating informative fine-grained learner feedback in open-vocabulary CAPT.

2. Methods

2.1. Corpora

NB Tale[2] is a Norwegian corpus of 260 native (first language; L1) speakers from all dialects, and 117 advanced non-native (second-language; L2) speakers. Recordings were collected in a controlled quiet environment from two microphones, and there are parallel recordings from every speaker for three sentences. Our experiments use the Sennheiser recordings of these sentences totalling 3.5 hours, following [16].

CAPTinI Icelandic recordings come from the Samrómur [28] and Samrómur Unverified [29] corpora, crowdsourced read speech collected from 2019 onwards. Our experiments use recordings of adult speakers gathered through the CAPTinI subcollection[3] for Icelandic pronunciation training [30]. To provide enough parallel data for DTW pronunciation scoring, only sentences with at least 10 recordings each of L1 and L2 Icelandic speakers are included in the final sample. In all this is 3014 recordings (2023 L1 and 991 L2) of 76 unique sentences, in total 3 hours, with different assortments of speakers for each sentence due to the crowdsourced collection.

2.2. Alignments

A recent investigation of word-level forced alignment in Swedish [31] identifies the best options as the Montreal Forced Aligner (MFA) [32] and CTC decoding with Wav2vec-2.0 [27]; we are not aware of an equivalent study for sub-word alignment.

Gold word and phone alignments are provided with NB Tale, as used for pronunciation scoring in [16]. Gold phone transcriptions are not available for CAPTinI.

MFA is a popular toolkit built on Kaldi using Gaussian mixture model-HMM triphone acoustic models to align speech to phone sequences from a pronunciation dictionary [32]. In keeping with a common use of MFA, we train acoustic models on the actual speech data we wish to align, i.e. NB Tale or CAPTinI datasets, along with corresponding orthographic transcripts and a pronunciation dictionary of the respective language. Icelandic alignment uses the General Icelandic Pronunciation Dictionary for ASR, which usually has one or two pronunciation options per word [33]. A Norwegian pronunciation dictionary was induced from the gold NB Tale phone transcripts; after removing stress diacritics, and keeping only pronunciations observed at least 10 times, most Norwegian words had 3-5 variants.

CTC segmentation with Wav2vec2 ASR models aligns audio with letters according to output label probabilities for each frame of audio [27]. This was selected for our E2E ASR experiments because Wav2vec2 is well established as a speech representation for DTW-based pronunciation tasks [34, 16], fine tuned models are available for both relevant languages, and the time resolution of Whisper, a major alternative with competitive general ASR performance, is considerably looser [35].

[2] www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-31/

[3] https://samromur.is/captini

2.3. Speech representations

MFA alignment uses Mel frequency Cepstral coeficients (MFCCs). To perform CTC alignment, and as speech embeddings for DTW score computation regardless of alignment type, we use the Norwegian model wav2vec2-1b-npsc-nst[4] with 48 transformer layers finetuned from XLS-R 1B on the NPSC and NST datasets, and the Icelandic model wav2vec2-large-xlsr-53-icelandic-ep10-1000h[5] with 24 transformer layers finetuned from XLSR-53 on 1000 hours of Icelandic speech.

2.4. Pronunciation scoring

Pronunciation scores following [16] derive from the dynamic time warping path costs for aligning test speech T with two parallel reference speech sets Ref_{L1} and Ref_{L2}, containing respectively native and non-native speech. Equation 1 describes the basic difference-to-sum ratio pronunciation score, which represents how distinctly nativelike or non-nativelike T is.

$$RelDTW(T) = \frac{Cost(Ref_{L2}, T) - Cost(Ref_{L1}, T)}{Cost(Ref_{L2}, T) + Cost(Ref_{L1}, T)} \quad (1)$$

When T is a whole word, the term $Cost(Ref_{L1}, T)$ is the average of DTW costs for aligning T with each $r \in Ref_{L1}$, as previously validated to measure foreign accent strength in English [36, 34]. When T is a segment within a word, the entire word is input to DTW, but the segment's pronunciation costs are extracted from only the portion of the DTW alignment path corresponding to that segment's time span within the test word, as described in [16]. In both cases $Cost(Ref_{L2}, T)$ represents the same computation with a reference set of L2 rather than L1 speakers; details of the method are described in [16, 34]. In evaluations on words and subword segments, classification of test speech is performed according to a threshold for relative DTW on Equation 1, whose value ranges between -1 and 1.

2.5. Experiment

Experiments classify both native and non-native speech samples according to a threshold on the DTW-based measure. In the absence of a labelled corpus of L2 pronunciation errors, classification accuracy in this task indicates the measure's basic ability to identify L2 mispronunciations that are non-nativelike while still accepting all real L1 variants as correct.

Relative DTW is run for every Transformer layer of a language's Wav2vec2 model, because phonetic information is often represented in some intermediate layers [34, 37, 16]. The classification performance is assessed for words and for subword segments, using each of the the two or three available alignment types in Icelandic and Norwegian data.

All evaluations use repeated k-fold cross validation, where k=2 and repeated with 5 random corpus splits; results show averages across these 10 runs. For Norwegian NB Tale, in each fold 130 L1 speakers are used as Ref_{L1} and 58 or 59 L2 speakers are used as Ref_{L2}, while the held-out speakers are classified by comparison to these. For Icelandic CAPTinI, since each sentence has a different set of speakers, each cross-validation fold contains a random half of the L1 speakers per sentence as Ref_{L1} and half of the L2 speakers per the same sentence as Ref_{L2}. The runtime for relative DTW scoring and evaluation was around 1 hour per Transformer layer, on Intel Xeon Gold 6248R CPU; code and evaluation data are also made available.[6]

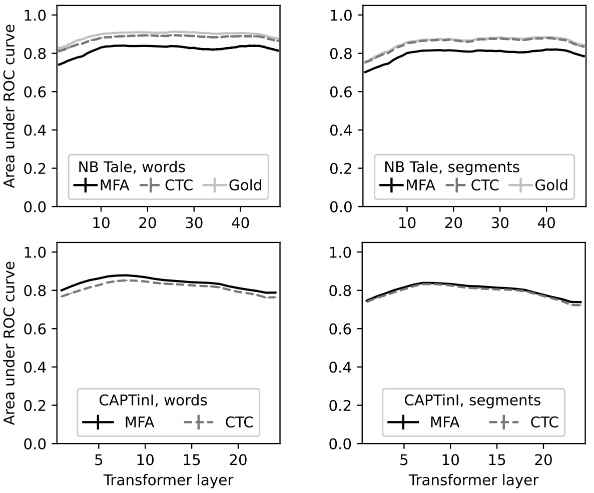

Figure 1: *Area under ROC curve by Transformer layer (1-24 or 1-48), for each available alignment type. Bars show standard error in AUROC across 10 runs.*

The main evaluation metrics are the area under the receiver operating characteristic curve (AUROC) and equal error rate (EER, the false positive rate and false negative rate at the point where these are equal), aiming to provide an overall picture of classifier performance and also convey practical usability for possible learner populations [38].

3. Results and analysis

Figure 1 shows AUROC for words and segments with each alignment type, across all Transformer layers, while Figure 2 reports the same for EER. Table 1 shows quantitative detail for selected sub-word segment results, adding several commonly used measures to facilitate comparison with related literature.

Figure 3 reports the recall for L2 speaker classification when applying the threshold at which precision for L1 speaker classification is at least 0.80, or R@0.8P. Unlike the summary metrics, this gives a snapshot of applied performance reflecting a domain-specific asymmetry, that flagging correct speech as a mispronunciation has a worse impact on the learners' experience than classification errors in the opposite direction [7].

Firstly, results for words provide necessary context to interpret the phone and letter results, by showing the background effects of corpus differences like recording conditions, number of speakers per reference set, and speech diversity. Word-level results also reflect basic aligner accuracy, independent of how suitable phones or characters would be as sub-word units of pronunciation analysis. Gold alignments for NB Tale (Gold-N), replicating [16], show the best performance overall. Performance is almost as good with CTC-N, but considerably worse with MFA-N. For CAPTinI, MFA-C performs slightly better than CTC-C.

A similar pattern of results holds for sub-word segments, with an even smaller performance gap between CTC and the best aligner per language. High performance for Gold-N is un-

[4] Model: https://huggingface.co/NbAiLab/wav2vec2-1b-npsc-nst. NPSC: www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-58/. NST: www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-54/

[5] https://huggingface.co/carlosdanielhernandezmena/wav2vec2-large-xlsr-53-icelandic-ep10-1000h

[6] https://github.com/catiR/orthography-capt

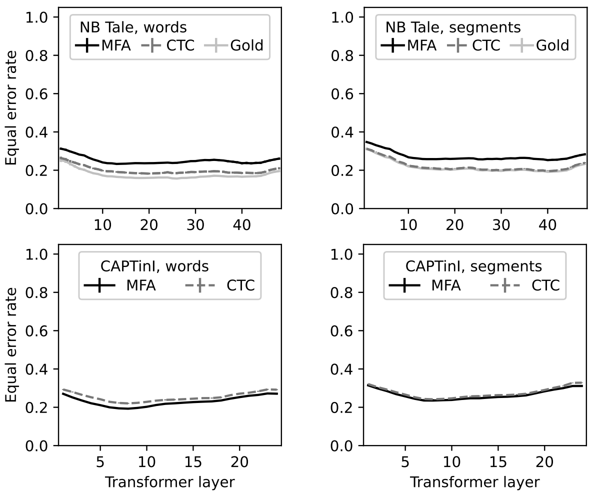

Figure 2: *EER by Transformer layer (1-24 or 1-48) for each alignment type. Bars show standard error across 10 runs.*

Table 1: *Detailed results for segments, selecting the highest performing layers from each corpus and alignment type. Data is specified by alignment type, NB Tale or CAPTinI corpus, and selected Transformer layer in parentheses. Results shown are: area under ROC curve, equal error rate, area under precision-recall curve, L2 recall when L1 precision is 0.8, and best F1 score.*

Data	ROC	EER	PR	R@P	F1
Gold-N (40)	0.89	19.2%	0.81	0.67	0.74
CTC-N (40)	0.88	19.5%	0.80	0.64	0.73
CTC-C (7)	0.83	24.3%	0.69	0.25	0.67
MFA-N (16)	0.82	25.8%	0.69	0.35	0.64
MFA-C (7)	0.84	23.5%	0.71	0.30	0.68

surprising given the lack of alignment error, clean recording conditions, and large number of reference set speakers. CTC-N also performs well. However, as both aligners for CAPTinI exceed MFA-N performance despite these corpus characteristics, MFA alignments were probably quite inaccurate for NB Tale.

Overall, the pattern of results indicates that letter-based segments can be at least as good a unit as phones for non-nativelike pronunciation classification, while the similarity between Gold-N and CTC-N performance in particular is a promising indication for meaningful pronunciation scoring by letter-aligned acoustic segments as well as traditional phones. Figure 3 indicates an especially strong impact on the potential user experience from choosing different Transformer layers.

Finally, all experiments regardless of corpus and alignment type showed better performance than reported baselines of [16]. Overall, MFA-N as well as both MFA-C and CTC-C have performance roughly similar to an experiment of [16] with gold alignments but only 10 reference speakers per set. While the 10-speaker scenario roughly matches the Icelandic data, MFA-N experiments use the entire NB Tale corpus, which reinforces the impression that inaccurate alignments negatively affected MFA-N performance.

In our experience using MFA in a variety of lower-resource language settings, this tool can have fairly high variance in quality, with substantial fluctuations from training acoustic models on different subsets of data or varying aspects of pronunciation dictionaries/phone sets. Since only one MFA training setup was used in this experiment, there is a fair chance that MFA-N performance could be improved with further efforts.

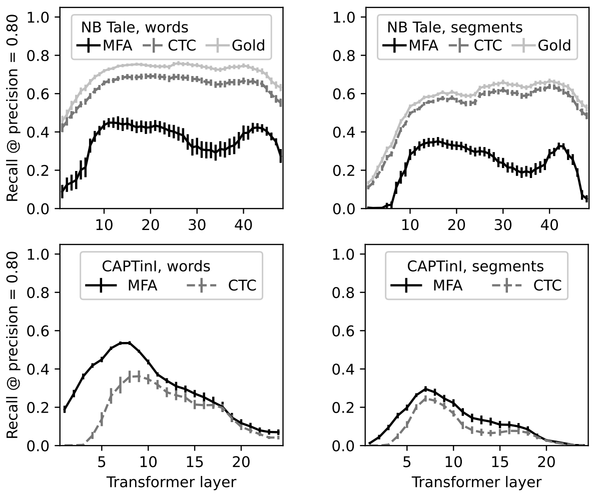

Figure 3: *L2 speaker identification recall, when precision for L1 speaker classification is constrained to 0.80. Bars show standard error across 10 runs.*

4. Discussion and conclusions

Experiments established clearly that L2 mispronunciation can be assessed in letter-aligned speech segments as well as phone-aligned segments. We conclude that the proposed architecture, combining label-independent relative DTW pronunciation scoring with E2E ASR to directly associate each letter with a score, advances towards a practical solution to give learners of many languages detailed pronunciation feedback in a format they can seamlessly comprehend.

Moving towards real-life practical application of this approach, the next key step is to investigate whether CTC letter alignments are indeed accurate enough to tell the learner what exactly they have mispronounced, not just that they have mispronounced something. Corpora with gold phone annotations like NB Tale can start to answer this, through comparing phone and letter alignments where it is clear what phone(s) the letter(s) should correspond to. Annotated error corpora, if available, can provide valuable information on how often particular errors are ascribed to the correct letters, in cases like the vowel quality change or consonant insertion errors described for *fljúga* in §1.1.2. Ultimately, success of the approach might vary according to orthographic and phonotactic properties of a given target language; it may be more promising for languages with relatively shallow orthographies, like Icelandic.

The general potential of E2E ASR character alignments as an alternative to phone alignments is rarely considered, but in the setting of fine-grained CAPT feedback all other solutions have their own substantial drawbacks, motivating our attempt to repurpose Wav2vec2. Evaluations showed this to be effective for the purpose, potentially deriving advantages from multilingual pretraining unavailable to many traditional phone-based approaches. The promising results can encourage similar efforts for other cases in need of innovative solutions.

5. References

[1] D. Korzekwa, J. Lorenzo-Trueba, T. Drugman, and B. Kostek, "Computer-assisted pronunciation training—speech synthesis is almost all you need," *Speech Communication*, vol. 142, pp. 22–33, 2022.

[2] R. Schmidt, "Attention, awareness, and individual differences in language learning," in *Perspectives on individual characteristics and foreign language education*, W. M. Chan, K. N. Chin, S. Bhatt, and I. Walker, Eds. Berlin, Boston: De Gruyter Mouton, 2012, ch. 2, pp. 27–50.

[3] 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.

[4] 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.

[5] M. El Tatawy, "Corrective feedback in second language acquisition," *Working papers in TESOL and Applied Linguistics*, 2002.

[6] 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*, 2019, pp. 954–958.

[7] M. Sancinetti, J. Vidal, C. Bonomi, and L. Ferrer, "A transfer learning approach for pronunciation scoring," in *ICASSP*. IEEE, 2022, pp. 6812–6816.

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

[9] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2016.

[10] W.-K. Leung, X. Liu, and H. Meng, "CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis," in *ICASSP*. IEEE, 2019, pp. 8132–8136.

[11] Y. Shen, Q. Liu, Z. Fan, J. Liu, and A. Wumaier, "Self-supervised pre-trained speech representation based end-to-end mispronunciation detection and diagnosis of Mandarin," *IEEE Access*, 2022.

[12] Y. Xiao, F. K. Soong, and W. Hu, "Paired phone-posteriors approach to ESL pronunciation quality assessment," in *Interspeech*, 2018, pp. 1631–1635.

[13] M. Wu, K. Li, W.-K. Leung, and H. Meng, "Transformer based end-to-end mispronunciation detection and diagnosis." in *Interspeech*, 2021, pp. 3954–3958.

[14] T.-H. Lo, Y.-T. Sung, and B. Chen, "Improving end-to-end modeling for mispronunciation detection with effective augmentation mechanisms," in *APSIPA*, 2021, pp. 1411–1415.

[15] H.-W. Wang, B.-C. Yan, H.-S. Chiu, Y.-C. Hsu, and B. Chen, "Exploring non-autoregressive end-to-end neural modeling for English mispronunciation detection and diagnosis," in *ICASSP*. IEEE, 2022, pp. 6817–6821.

[16] C. Richter and J. Guðnason, "Relative dynamic time warping comparison for pronunciation errors," in *ICASSP*. IEEE, 2023.

[17] S. G. Lambacher, W. L. Martens, K. Kakehi, C. A. Marasinghe, and G. Molholt, "The effects of identification training on the identification and production of American English vowels by native speakers of Japanese," *Applied Psycholinguistics*, vol. 26, no. 2, pp. 227–247, 2005.

[18] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training," *Comput Assist Lang Learn*, vol. 15, no. 5, pp. 441–467, 2002.

[19] X. Wei, C. Cucchiarini, R. van Hout, and H. Strik, "Automatic speech recognition and pronunciation error detection of Dutch Non-native speech: cumulating speech resources in a pluricentric language," *Speech Communication*, vol. 144, pp. 1–9, 2022.

[20] K. Gorman, L. F. Ashby, A. Goyzueta, A. D. McCarthy, S. Wu, and D. You, "The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion," in *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 2020, pp. 40–50.

[21] A. M. Harrison, W.-K. Lo, X.-j. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *International Workshop on Speech and Language Technology in Education*, 2009.

[22] P. Plantinga and E. Fosler-Lussier, "Towards real-time mispronunciation detection in kids' speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 690–696.

[23] K. Nishi and D. Kewley-Port, "Training Japanese listeners to perceive American English vowels: Influence of training sets," *J. Speech, Language, and Hearing Research*, 2007.

[24] A. D. Franklin and C. Stoel-Gammon, "Using multiple measures to document change in English vowels produced by Japanese, Korean, and Spanish speakers: The case for goodness and intelligibility," *American Journal of Speech-language pathology*, vol. 23, no. 4, pp. 625–640, 2014.

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[26] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "XLS-R: self-supervised cross-lingual speech representation learning at scale," in *Interspeech*, 2022, pp. 2278–228.

[27] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "CTC-segmentation of large corpora for German end-to-end speech recognition," in *SPECOM 2020*. Springer, 2020, pp. 267–278.

[28] D. E. Mollberg, Ó. H. Jónsson, S. Þorsteinsdóttir, J. V. Guðmundsdóttir, S. Steingrímsson, E. H. Magnúsdóttir, J. Y. Fong, M. Borsky, and J. Gudnason, "Samromur 21.05," 2021, CLARIN-IS.

[29] S. Hedström, J. Y. Fong, R. Þórhallsdóttir, D. E. Mollberg, S. F. Guðmundsson, Ó. H. Jónsson, S. Þorsteinsdóttir, E. H. Magnúsdóttir, and J. Gudnason, "Samromur unverified 22.07," 2022, CLARIN-IS.

[30] C. Richter, B. Bédi, R. Pálsson, and J. Guðnason, "Computer-assisted pronunciation training in Icelandic (CAPTinI): developing a method for quantifying mispronunciation in L2 speech," *Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022*, p. 334, 2022.

[31] K. Biczysko, *Automatic Annotation of Speech: Exploring Boundaries within Forced Alignment for Swedish and Norwegian*. Masters thesis, Uppsala University, 2022.

[32] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[33] A. B. Nikulásdóttir and J. Guðnason, "General pronunciation dictionary for ASR," 2017, CLARIN-IS.

[34] M. Bartelds, W. de Vries, F. Sanal, C. Richter, M. Liberman, and M. Wieling, "Neural representations for modeling variation in speech," *Journal of Phonetics*, vol. 92, p. 101137, 2022.

[35] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[36] M. Bartelds, C. Richter, M. Liberman, and M. Wieling, "A new acoustic-based pronunciation distance measure," *Front. Artif. Intell.*, vol. 3, p. 39, 2020.

[37] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," in *Interspeech 2022*. IEEE, 2022, pp. 6817–6821.

[38] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech communication*, vol. 52, no. 3, pp. 254–267, 2010.