# RELATIVE DYNAMIC TIME WARPING COMPARISON FOR PRONUNCIATION ERRORS

*Caitlin Richter, Jón Guðnason*[*]

Language and Voice Lab
Reykjavik University, Reykjavik, Iceland
{caitlinr, jg}@ru.is

## ABSTRACT

We propose using a dynamic time warping (DTW) difference-to-sum ratio to classify speech as either matching or diverging from a linguistic standard. This measure effectively recognises non-native Norwegian speakers' mispronunciations in words and phonetic segments. The contributions of the approach include (a) using DTW comparisons from two parallel sources, which represent the linguistic standard (e.g. native speakers) and an error model, to identify pronunciation errors; (b) recognising a heterogeneous standard, in this case the highly variable range of Norwegian dialects, instead of only a specified canonical phoneme sequence; (c) handling unanticipated pronunciation variants, both acceptable and unacceptable, beyond those seen in the standard and error models; and (d) requiring minimal training or pretraining data in the target language, which helps to make pronunciation error detection accessible even in low-resource languages without functional ASR.

***Index Terms***— computer assisted pronunciation training, pronunciation error detection, speech processing, dynamic time warping, low resource languages

## 1. INTRODUCTION

We develop a method of using dynamic time warping (DTW) for binary classification of spoken words and phonetic segments, focused on the application of identifying non-nativelike mispronunciations in adult Norwegian learners' speech while accepting the full range of native Norwegian dialect variation as correct.

Previous approaches to second-language mispronunciation detection have relatively poor accuracy, and need human-annotated corpora of speech errors and/or speech processing technologies that are inaccessible in lower resource target languages [1–7]. We compare a speech sample to two corresponding sets of reference speech, one set representative of a selected linguistic standard (native speakers of a language) and one not (non-native speakers), to classify the sample based on both comparisons. This requires only a small unannotated corpus of parallel speech, a more attainable option if high-quality ASR or TTS is yet to be developed. Furthermore, with this approach phone pronunciation scoring is not influenced by the identity of phone labels, so many variants can be classified as correct. We evaluate mispronunciation detection with a proxy task of classifying speech samples as being either a native speaker (L1; by definition has no non-nativelike mispronunciations) or a non-native speaker (L2; may have non-nativelike mispronunciations). This task can be used in the vast majority of languages with no dedicated learner error corpus, and it demonstrates that the proposed $relative\ DTW$ measure is sensitive to distinctions between correct pronunciations and Norwegian learners' non-nativelike errors. This work provides a new path towards accurate mispronunciation detection.

### 1.1. Motivation

There is high demand for automatic pronunciation error detection, or classification of pronunciations as representative or not representative of some linguistic standard, in several areas of second language teaching, including computer assisted pronunciation training (CAPT) and automatic test scoring [1]. Other important applications for this task include tracking typical child language development, disorders, and speech therapy, and assessing symptoms related to memory degeneration or brain damage such as aphasias [8, 9].

However, current methods are unsatisfactory for these applications. Performance for L2 English word-level mispronunciation detection is summarised as '60% precision at 40%–80% recall' in a recent review [2], while the same paper's proposal to generate additional training data with TTS improves state of the art for area under precision/recall curve (AUPR) to a maximum of 0.75. Phone-level[1] error detection could be more valuable to end users, but performance is even lower. Earlier methods are exemplified by Goodness of Pronunciation (GoP), which uses forced alignment to segment a spoken word into a pre-specified sequence of phones, and then measures similarity to canonical acoustic models for each phone [10]. Extensions include better speech representations and acoustic modelling [6, 7, 11]. End-to-end pipelines also appear recently, for example to decode whether a sample's apparent phone sequence matches the correct sequence [3, 5]. E2E methods have ever-increasing demands for labelled training data of both native and L2 speakers, and approaches aiming to mitigate the data problem struggle to match the performance of a GoP baseline [4, 12]. More competitive results have been achieved only for strictly delimited particularly high-resource learning settings [3, 13, 14]. In any case, phone error detection still critically depends on the identity of phones, so 'correct' pronunciations are constrained by pre-defined canonical phone spellings regardless of the range of acceptable variants in the target language. Recent work acknowledging the large variety of possibly correct pronunciations performs only word-level error detection [15].

Meanwhile, there is a major disconnect between speech processing research outcomes and mispronunciation detection as deployed in commercial and non-commercial CAPT systems [16]. Second

[*] We thank our colleagues in developing pronunciation training: Ragnar Pálsson, Kolbrún Friðriksdóttir, Marc Daníel Skipstað Volhardt, Branislav Bédi, Róbert Kjaran, Júlíus Reynald Björnsson, Safa Jemai, Eydís Huld Magnúsdóttir, Luke O'Brien, and Staffan Hedström. This work was supported by The Icelandic Centre for Research (RANNIS) under the project Computer Aided Pronunciation Training in Icelandic, under Grant 200065-5301.

[1] We use the term *phone* for the smallest relevant unit because the literature in this field does not consistently distinguish between phonemes and phonetic segment variants (allophones).

language teachers may incorporate a module meant to detect pronunciation errors on the assumption that it simply does so, encouraged by commercial services that conceal how their product works and provide no formal evaluations [17]. Research results like optimal F1 score can be uninformative to teachers who strictly require high precision [11]. Additional barriers between research insights and L2 education include difficulty comparing evaluations across corpora with different base error rates and severity, and the expense of creating relevant corpora for languages without them [5]. Much of the phone-level error detection research mentioned above was evaluated on held-out test sets of only 4-6 speakers (e.g. L2-ARCTIC; exceptions include [3, 5]), but CAPT software may need to serve hundreds to millions of users, highlighting the need for scalable evaluations.

### 1.2. Related work

DTW has a long history in measuring similarity between sequences of speech [18]. It successfully quantifies phonetic differences in applications including word recognition, spoken term detection, dialect clustering, and strength of foreign accent when speaking a second language [19–21]. DTW has previously been used for phone pronunciation error classification only with MFCC (mel frequency cepstral coefficient) speech representations [22], but several adjacent pronunciation-scoring tasks have directly compared MFCCs with Wav2Vec 2.0 [23] and consistently find the latter better [14, 24, 25].

While previous efforts at DTW-based CAPT have compared learners' speech to correct native-speaker references only [20–22], other L2 pronunciation scoring methods benefit from incorporating secondary databases of non-native/incorrect speech [26, 27]. This helps to calibrate numeric scores, leading to better speaker independence as well as higher correlation to human perception of errors [27]. Non-target training data can help to detect mispronunciations beyond simply clean substitutions of one target-language sound for a different target-language sound, which are common when learners use a sound from their native language that is not part of the one they are learning [28]. Therefore, an innovation of our proposal is to perform Wav2vec2-based DTW pronunciation comparison twice separately, once to compare the learner with acceptable speech and a second time to compare with non-target speech, and combine both sources of information as described in §2.2.

## 2. METHODS

### 2.1. Data

Experiments use the NB Tale[2] corpus of 260 native (first language; L1) Norwegian speakers representing maximum possible dialect diversity, and 117 non-native (second-language; L2) speakers from numerous first language backgrounds. Norwegian is known for high dialect variation with no standard prestige dialect [29], and the native speakers were asked before recording to reflect on their dialect and speak naturally. NB Tale includes many sentences, of which 3 are recorded by all speakers, so our experimental data consists of these three parallel sentences which have a total of 50 words. Recordings were collected in the same controlled quiet environment for all speakers, from two microphones; we use the Sennheiser recordings.

Speech embeddings for NB Tale were generated with several Transformer models using Wav2Vec 2.0 architecture [23]:

**w2v2** Wav2Vec 2.0 Large, 960 hours English pretraining, no fine tuning; 24 Transformer layers and 317M parameters [23]

**XLS-R 300** 436k hours pretraining on 128 language varieties, including 130 hours of Norwegian (45 of which are specified as Nynorsk, a distinction which applies to accompanying transcripts rather than audio itself), no fine tuning; 24 Transformer layers and 317M parameters [30]

**XLS-R 1B** The same training data as XLS-R 300; 48 Transformer layers and 965M parameters [30]

**NB-1B** Based on XLS-R 1B pretraining, finetuned on two National Library of Norway (Nasjonalbiblioteket, NB) L1 Norwegian datasets; 48 Transformer layers and 965M parameters.[3]

### 2.2. Relative DTW

The relative DTW score of a spoken test sample $T$ with respect to two sets of reference samples $R_{Standard}$ and $R_{Other}$ represents how strongly $T$ resembles the speech in one reference set more than the other:

$$RelDTW(T) = \frac{Cost(R_{Other}, T) - Cost(R_{Standard}, T)}{Cost(R_{Other}, T) + Cost(R_{Standard}, T)} \quad (1)$$

$Cost(R_{Standard}, T)$ and $Cost(R_{Other}, T)$ are the averages of the length-normalised dynamic time warping alignment cost for aligning $T$ with each reference speech recording $r \in R_{Standard}$ and $r \in R_{Other}$. These are nonnegative values, with higher cost indicating greater difference between $T$ and the recordings in the given reference set; when $R_{Standard}$ is a set of native speakers and $T$ is a non-native speaker of the same language, $Cost(R_{Standard}, T)$ is identical to the quantity proposed by [19, 24] to estimate foreign accent strength in English (see [19] for complete details).

We classify $T$ according to a threshold on the relative DTW score (Equation 1), which ranges between -1 and 1. When $R_{Standard}$ contains L1 references while $R_{Other}$ contains L2 speakers, relative DTW for L1 $T$ samples is reliably positive, because $T$ will resemble the set of fellow L1 speakers more than L2 speakers. Scores when $T$ contains L2 mispronunciations, i.e. pronunciations that do not occur in the $R_{Standard}$ data, may be negative if the mispronunciation is common among a subset of L2 speakers, or around 0 for novel mispronunciations that are not characteristic of either reference set.

### 2.3. Experiment

Experiments classify both native and non-native Norwegian speakers from the NB Tale corpus, according to a threshold on DTW-based scores. We compare our proposed relative (two-way) DTW to the baseline one-way DTW measure of [19]. Classification performance is reported for words and phones, as described in §2.4. We test all Transformer layers of the pre-trained speech embedding models listed in §2.1, because phonetic information is often represented in some intermediate layers [1, 24, 31].

All evaluations use repeated k-fold cross validation, where k=2 and repeated with 5 different random splits of NB Tale. 130 L1 speakers are used as $R_{Standard}$ and 58 or 59 L2 speakers are used as $R_{Other}$, while the held-out speakers are classified by comparison to these. Results below are the average values across the 10 runs. The runtime for DTW and evaluation was slightly under 1 hour per Transformer layer on an Intel Xeon Gold 6248R CPU.

The main evaluation metrics are the area under the receiver operating characteristic curve (AUROC) and equal error rate (EER, the

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

[3] 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/

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.

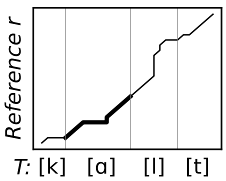

**Fig. 1**. DTW alignment path basis for phone cost of [ɑ] in test word $T$ 'kalt' and one reference word $r$, to be repeated across each $r \in R$.

### 2.4. Word and phone scoring

Aligned word pairs are the basic linguistic unit $T$ for DTW cost comparisons, following [19], but from this it is possible to derive related scores for individual phones. To replace length-normalised word alignment costs in Equation 1, phone costs come from the average of local costs along the steps of DTW alignment path assigned to a phone's location in the test word $T$, as shown in Figure 1.

Deriving phone scores from a portion of the word's alignment path allows every phone segment in $T$ to be compared to the most appropriate part of each reference word. This is preferable to simply applying Equation 1 with phones replacing words as the unit of $T$ for samples input to DTW, because it is more flexible with words whose variants have different numbers of phonemes, and with acceptable pronunciations that do not quite match any pre-specified canonical pronunciation.

Word and phone time alignments are distributed with NB Tale; this otherwise requires a pronunciation dictionary and forced alignment. Dictionaries can be made by hand or assisted by tools like wav2vec2phoneme [32], and minimally need only the words whose pronunciations will be scored. Forced alignment benefits from 20-30 hours of word-transcribed L1 speech to train acoustic models, but another language's models can be substituted if not available [33].

## 3. RESULTS

Figure 2 shows the AUROC of relative DTW and the baseline measure, for each Transformer layer of the different pretrained models. As these models have 24-48 layers each, it is impractical to print complete results, so Table 1 includes details of AUROC and EER for a subset of middle layers, which are among the overall best.

Relative DTW always outperforms the baseline, for the same pretrained model/layer and linguistic units of classification. Both methods have consistent performance across runs in cross-validation.

Finetuning in the target language (model NB-1B) is evidently superior. It is especially helpful to the baseline, reducing the performance gap from the baseline to relative DTW more than all other models. However, performance gains for relative DTW are rather modest compared to using w2v2, which has just 1/3 of the parameters and only 960 hours of English pretraining. Relative DTW shows no discernible benefit from the multilingual pretraining data of XLS-R 300, nor the larger parameter space of XLS-R 1B. The baseline is more sensitive to these factors, particularly for phone classification, and therefore has a greater dependence on language resources in the target language.

## 4. DISCUSSION

NB Tale provides a fairly challenging classification task. Norwegian L1 dialects have distinct phonetic inventories and word variants, so perceptually dissimilar pronunciations must all be classified as L1. Adding to the difficulty, NB Tale L2 speakers are advanced learners who have lived in Norway for several years, learning their local dialect; nearly all use Norwegian in their workplace and daily life, so their mispronunciations are more subtle than beginning language students who may struggle to be comprehensible. Despite this, relative DTW separates most utterances of L1 and L2 Norwegian speakers, as shown in §3. Still, it is impossible to tell how much additional L2 mispronunciation the relative DTW score fails to detect, since there is no available ground truth regarding the accuracy of NB Tale L2 speech. It is only certain that there is at least as much mispronunciation in the data as has been identified by the best classifier.

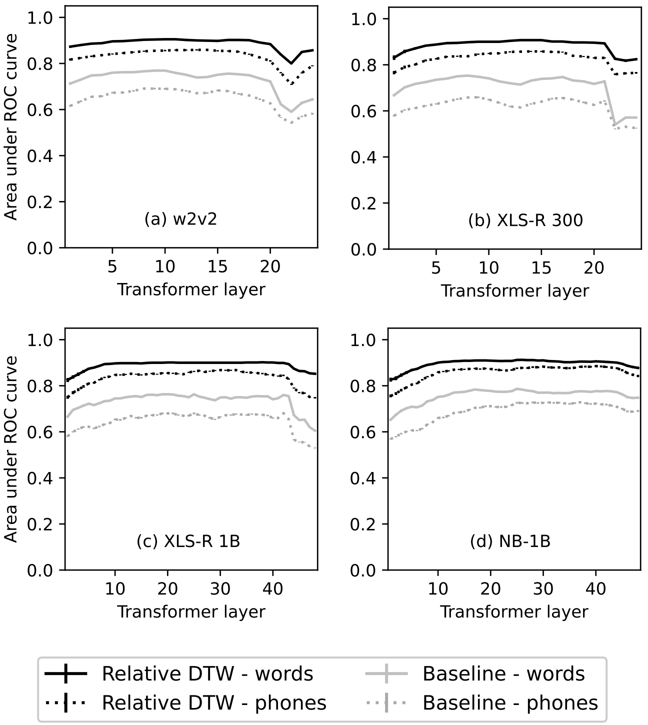

**Fig. 2**. Area under ROC curve by Transformer layer (1-24 or 1-48), for different pretrained Transformer models. Bars show standard error in AUROC across 10 runs.

The comparison between our proxy task and the CAPT error detection task, classifying L2 speech as acceptable or unacceptable, is undetermined. Another study which used the L1/L2 proxy task did not evaluate on held-out speakers, and had different recording conditions for the L1 and L2 speakers which can also affect discriminability [13]. Error detection evaluations (see §1) are popularly reported with (area under) precision/recall, which acknowledges the impact of errors being less common than acceptable pronunciations in L2 speech, but unlike ROC it is not directly interpretable about how the same method performs on other learner populations. Therefore, to help contextualise our results, Figure 3 shows ROC curves for the example of Transformer layer 28 in the NB-1B model.

Since we aim to minimise the language resources needed to classify speech, Figure 4 reduces the number of speakers in $R_{Standard}$ and $R_{Other}$ from 130 and 58/59 to 10 and 10 respectively. Due to the large number of L1 dialects and L2 backgrounds in NB Tale, this task also incorporates classifying L1 and L2 speaker types beyond the 10 present in each respective set. Performance drops slightly, as AUROC for Relative DTW is 0.88 for words and AUROC = 0.85 for phones. Table 2 further shows how performance gently degrades with major reductions in speaker set size. This pattern is consistent

|  |  | Relative DTW |  |  |  | Baseline |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| Model | Layer | AUROC |  | EER |  | AUROC |  | EER |  |
|  |  | Word | Phone | Word | Phone | Word | Phone | Word | Phone |
| w2v2 | 13 | 0.90 | 0.86 | 0.17 | 0.22 | 0.74 | 0.67 | 0.33 | 0.38 |
| w2v2 | 14 | 0.90 | 0.86 | 0.17 | 0.22 | 0.74 | 0.67 | 0.32 | 0.38 |
| w2v2 | 15 | 0.90 | 0.86 | 0.18 | 0.22 | 0.75 | 0.68 | 0.31 | 0.37 |
| w2v2 | 16 | 0.90 | 0.86 | 0.17 | 0.22 | 0.76 | 0.68 | 0.31 | 0.37 |
| XLS-R 300 | 13 | **0.91** | 0.86 | 0.17 | 0.22 | 0.72 | 0.61 | 0.34 | 0.42 |
| XLS-R 300 | 14 | **0.91** | 0.86 | **0.16** | 0.22 | 0.73 | 0.63 | 0.33 | 0.41 |
| XLS-R 300 | 15 | **0.91** | 0.86 | 0.17 | 0.22 | 0.74 | 0.64 | 0.32 | 0.40 |
| XLS-R 300 | 16 | 0.90 | 0.85 | 0.17 | 0.22 | 0.74 | 0.65 | 0.32 | 0.39 |
| XLS-R 1B | 25 | 0.90 | 0.86 | 0.17 | 0.21 | 0.76 | 0.68 | 0.30 | 0.37 |
| XLS-R 1B | 26 | 0.90 | 0.86 | 0.17 | 0.21 | 0.75 | 0.67 | 0.31 | 0.38 |
| XLS-R 1B | 27 | 0.90 | 0.86 | 0.17 | 0.22 | 0.75 | 0.67 | 0.32 | 0.38 |
| XLS-R 1B | 28 | 0.90 | 0.86 | 0.17 | 0.21 | 0.74 | 0.66 | 0.32 | 0.39 |
| NB-1B | 25 | **0.91** | **0.88** | **0.16** | **0.20** | **0.79** | **0.73** | **0.29** | **0.34** |
| NB-1B | 26 | **0.91** | **0.88** | **0.16** | **0.20** | 0.78 | **0.73** | **0.29** | **0.34** |
| NB-1B | 27 | **0.91** | **0.88** | **0.16** | **0.20** | 0.78 | 0.72 | 0.30 | **0.34** |
| NB-1B | 28 | **0.91** | **0.88** | **0.16** | **0.20** | 0.78 | **0.73** | 0.30 | **0.34** |

**Table 1**. Selected details of results: area under ROC curve (AUROC) and equal error rate (EER), averaged over 5 repetitions of 2-fold cross validation.

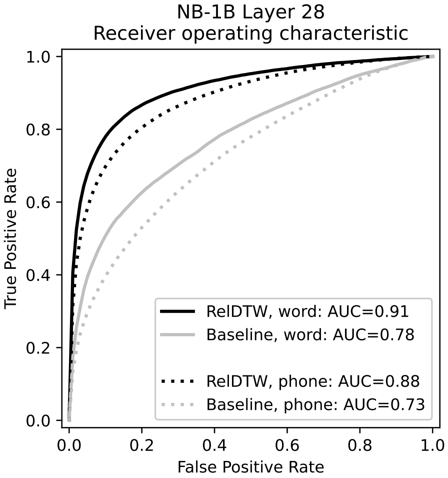

**Fig. 3**. ROC curves (average across 10 runs) for word and phone classification, using layer 28 of NB-1B.

across all models/layers, including English-trained w2v2, so the burden for target-language speech data can be significantly relieved in extreme low-resource settings without badly compromising quality.

It is notable that, even for the phoneme task, relative DTW performance in Figure 2 was almost invariant across many intermediate Transformer layers, although a marked decline towards final output layers (echoing [24]) justifies the comprehensive testing. Still, quite different linguistic information probably underlies the successful classification across layers, which implies that there must be some non-nativelike aspects to L2 speakers' productions at all of these linguistic levels [1, 24, 31]. This brings duration to mind, as less fluent speakers tend to talk more slowly; therefore, although length-normalised DTW should be insensitive to different speaking rates, we also confirmed that there is no correlation between speech duration and DTW when controlling for pronunciation quality, i.e. when restricting the sample to L1 speakers who all have nativelike pronunciation quality. A classification task targeting more specific information might constrain choices for the best Transformer layers.

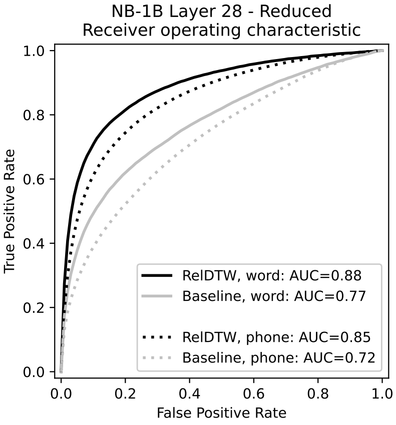

**Fig. 4**. ROC curves (average across 10 runs) for classification with $R_{Standard}$ and $R_{Other}$ reduced to 10 speakers each.

|  | RelDTW EER |  |
|---|---|---|
| Size | Word | Phone |
| 50 | 0.16 | 0.20 |
| 30 | 0.17 | 0.21 |
| 15 | 0.18 | 0.21 |
| 10 | 0.19 | 0.23 |
| 5 | 0.22 | 0.26 |

**Table 2**. NB-1B Layer 28 RelDTW EER, with decreasing number of reference speakers. Baseline EER remained 0.30 (words) or 0.34 (phones) for all reference set sizes.

## 5. CONCLUSIONS

We have proposed a relative DTW method for binary classification of pronunciations, which is able to identify non-nativelike pronunciations of L2 Norwegian speakers while also correctly classifying many different L1 variants as nativelike. Phones are classified based on acoustic similarity of the speech segment to corresponding reference speech as determined by word alignment, with no regard to identity of the phone label, which eliminates a problem inherent to previous phone error detection methods that artificially impose a single correct pronunciation [3–7, 10–12, 14, 22]. Our scalable evaluation task facilitates pronunciation scoring development for the large population of prospective CAPT users in many target languages.

Going forward, tuning on an annotated error corpus will be valuable to refine performance for classifying L2 speech into categories like *major pronunciation error* vs. *acceptable (not necessarily nativelike)*. Possibilities afforded by relative DTW include relaxing the acceptance threshold for the difference-to-sum ratio, or adding known correct L2 samples to the $R_{Standard}$ set, since as established, relative DTW performs well with a heterogeneous 'standard'. Indeed many tasks might be accomplished by varying the composition of $R_{Standard}$ and $R_{Other}$, such as an example suggested by classifying children's speech as typical or disordered with MFCC-based DTW [34]. Although a pattern qualitatively appears in the ratio of a child speech sample's distance to typical references vs. disordered references, this comparison was not formalised, and without it classification accuracy is fairly low [34]. This indicates an opportunity for further research with relative DTW in a variety of analogous applications.

Ultimately, relative DTW will be most beneficial if freed from reliance on parallel sentence recordings. One option to score novel sentences is to extract words or subwords from minimal parallel contexts, like trigrams. The recent use of speech synthesis to generate mispronounced training data points to another promising direction [2], as this could potentially synthesise reference speech for relative DTW, and facilitate progress towards accurate open-vocabulary word and phone pronunciation classification for any language.

## 6. REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

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

[8] A. N. Asad, S. C. Purdy, E. Ballard, L. Fairgray, and C. Bowen, "Phonological processes in the speech of school-age children with hearing loss: Comparisons with children with normal hearing," *J Commun Disord*, vol. 74, pp. 10–22, 2018.

[9] B. Moëll *et al.*, "Speech data augmentation for improving phoneme transcriptions of aphasic speech using wav2vec 2.0 for the psst challenge," in *LREC*, 2022, pp. 62–70.

[10] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, no. 2-3, pp. 95–108, 2000.

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

[12] 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.

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

[14] 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.

[15] D. Korzekwa, J. Lorenzo-Trueba, T. Drugman, S. Calamaro, and B. Kostek, "Weakly-supervised word-level pronunciation error detection in non-native english speech," in *Interspeech*, 2021, pp. 4408–4412.

[16] D. M. Chun and Y. Jiang, "Using technology to explore l2 pronunciation," *Second Language Pronunciation: Bridging the Gap Between Research and Teaching*, p. 129, 2022.

[17] H. Bozorgian and E. Shamsi, "Computer-assisted pronunciation training on Iranian EFL learners' use of suprasegmental features: A case study," *Comput. Assist. Lang. Learn.*, vol. 21, no. 1, pp. 93–113, 2020.

[18] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE trans.*, vol. 26, no. 1, pp. 43–49, 1978.

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

[20] T. Shi, S. Kasahara, T. Pongkittiphan, N. Minematsu, D. Saito, and K. Hirose, "A measure of phonetic similarity to quantify pronunciation variation by using ASR technology.," in *ICPhS*, 2015.

[21] J. Yue *et al.*, "Automatic scoring of shadowing speech based on DNN Posteriors and Their DTW.," in *INTERSPEECH*, 2017, pp. 1422–1426.

[22] Z. Miodonska, M. D. Bugdol, and M. Krecichwost, "Dynamic time warping in phoneme modeling for fast pronunciation error detection," *Comput. Biol. Med.*, vol. 69, pp. 277–285, 2016.

[23] 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.

[24] M. Bartelds, W. de Vries, F. Sanal, C. Richter, M. Liberman, and M. Wieling, "Neural representations for modeling variation in speech," *J. Phon.*, vol. 92, p. 101 137, 2022.

[25] Q. Zeng, D. Chong, P. Zhou, and J. Yang, "Low-resource accent classification in geographically-proximate settings: A forensic and sociophonetics perspective," in *Interspeech*, ISCA, 2022, pp. 5308–5312.

[26] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *Eurospeech*, 1999.

[27] M. Tu, A. Grabek, J. Liss, and V. Berisha, "Investigating the role of L1 in automatic pronunciation evaluation of L2 speech," in *Interspeech*, 2018, pp. 1636–1640.

[28] S. Mao, X. Li, K. Li, Z. Wu, X. Liu, and H. Meng, "Unsupervised discovery of an extended phoneme set in l2 english speech for mispronunciation detection and diagnosis," in *ICASSP*, IEEE, 2018, pp. 6244–6248.

[29] U. Røyneland, "Dialects in Norway: catching up with the rest of Europe?" *Int J Sociol. Lang.*, vol. 196/197, pp. 7–30, 2009.

[30] A. Babu *et al.*, "XLS-R: self-supervised cross-lingual speech representation learning at scale," in *Interspeech*, 2022, pp. 2278–228.

[31] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *ASRU*, IEEE, 2021, pp. 914–921.

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

[33] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner," in *Interspeech*, 2017, pp. 498–502.

[34] J. Liu *et al.*, "Speech disorders classification in phonetic exams with MFCC and DTW," in *IEEE CIC*, 2021, pp. 35–40.