L2-GEN: A Neural Phoneme Paraphrasing Approach to L2 Speech Synthesis for Mispronunciation Diagnosis

Daniel Yue Zhang, Ashwinkumar Ganesan*, Sarah Campbell, Daniel Korzekwa*

Amazon Alexa AI

{dyz,gashwink,srh,korzekwa}@amazon.com

Abstract

In this paper, we study the problem of generating mispronounced speech mimicking non-native (L2) speakers learning English as a Second Language (ESL) for the mispronunciation detection and diagnosis (MDD) task. The paper is motivated by the widely observed yet not well addressed data sparsity issue in MDD research where both L2 speech audio and its fine-grained phonetic annotations are difficult to obtain, leading to unsatisfactory mispronunciation feedback accuracy. We propose L2-GEN, a new data augmentation framework to generate L2 phoneme sequences that capture realistic mispronunciation patterns by devising an unique machine translation-based sequence paraphrasing model. A novel diversified and preference-aware decoding algorithm is proposed to generalize L2-GEN to handle both unseen words and new learner population with very limited L2 training data. A contrastive augmentation technique is further designed to optimize MDD performance improvements with the generated synthetic L2 data. We evaluate L2-GEN on public L2-ARCTIC and SpeechOcean762 datasets. The results have shown that L2-GEN leads to up to 3.9%, and 5.0% MDD F1-score improvements in in-domain and out-of-domain scenarios respectively.

Index Terms: mispronunciation diagnosis, speech synthesis, sequence-to-sequence model, machine translation

1. Introduction

Mispronunciation detection and diagnosis (MDD) is a core component in computer-assisted pronunciation training (CAPT) system. MDD identifies pronunciation errors and provides corrective feedback to guide non-native (L2) language learners [1]. In the past decade, significant progress has been made in addressing the MDD task and various models have been proposed [2, 3, 4, 5, 6, 7, 8]. A key technical challenge for these models is "data sparsity". In specific, MDD requires fine-grained diagnosis of pronunciation quality, which means manual annotation at the phoneme level for L2 speech is often required for training these models [7]. Such annotation efforts are extremely time-consuming and labor intensive [9, 10]. In existing literature, there are two orthogonal directions to address this L2 data sparsity issue. One is to eliminate MDD model's dependency on L2 training data via weakly supervised [11] or unsupervised [12] learning techniques. The other direction is to perform data augmentation by generating new L2 training examples [13, 14, 15, 16, 17, 18], which is the focus of this work.

A naive approach for L2 data augmentation is to automatically transcribe existing L2 speech with forced-alignment models to generate corresponding phoneme annotations [13]. However, this approach does not capture pronunciation errors and leads to poor annotation accuracy for non-native speech. Moreover, obtaining additional L2 speech itself is difficult, resulting in a very limited set of L2 data to be annotated [19]. This paper focuses on the alternative direction where one can first generate L2 mispronounced phoneme sequences by replacing phonemes in the reference L1 speech, and then synthesize audio from the generated L2 phonemes. L2 phoneme replacements may be sampled randomly [14] or based on the probability distribution of phonemes to gain knowledge from the actual mispronunciations patterns [15]. Alternatively, phoneme-to-phoneme (P2P) models can generate mispronounced speech by perturbing the phoneme sequence of the corresponding L1 speech with decision trees [16] or deep learning techniques [17]. More recently, Korzekwa *et al.* further proposed two new L2 data synthesis techniques including Text-to-Speech (T2S), and Speech-to-Speech (S2S) models [18].

In this paper, we introduce L2-GEN, a new data augmentation framework to generate L2 phoneme sequences and corresponding speech that captures mispronunciation patterns from real-world non-native ESL learners. L2-GEN takes a unique perspective of mapping the L2 phoneme generation problem to a paraphrasing problem in natural language generation (NLG) domain and proposes a machine learning-based L1-to-L2 phoneme paraphrasing model to produce a set of L2 phoneme sequences given any L1 reference phonemes.

L2-GEN jointly addresses two critical challenges that have not been well studied in existing work above. First, L2-GEN generates *realistic* L2 phoneme sequences that can capture mispronunciation patterns from real-world ESL learners, given limited L2 training data. This is in sharp contrast to phoneme replacement approaches mentioned above [14, 15] that randomly replace L1 phonemes, and therefore result in unrealistic or even unpronounceable phoneme sequences. L2-GEN addresses this challenge by jointly learning "where" and "what" phonemes to replace via a principled sequence-to-sequence (Seq2Seq) paraphrasing framework. Second, L2-GEN generates L2 phoneme sequences that can be *generalizable* to infer mispronounced phoneme sequences for unseen words as well as to produce mispronunciation patterns that are completely absent or rare in the training data. This challenge is crucial and practical in the L2 data sparsity scenario. To address generalizability challenge, a novel diversified and preference-aware decoding algorithm is proposed to jointly leverage limited L2 training data and external knowledge base to generate diversified L2 phonemes sequences with locale-specific mispronunciation patterns. L2-GEN further includes a novel contrastive data augmentation technique to improve the MDD model performance with generated L2 synthetic data.

We evaluate the proposed L2-GEN framework on two public datasets - L2-ARCTIC [13] and SpeechOcean762 [19]. The results have shown that the proposed framework increases the F1-score of detecting mispronunciations by up to 3.9%, and achieves 5.0% F1-score increase when adapting to a new locale.

*Equal contribution

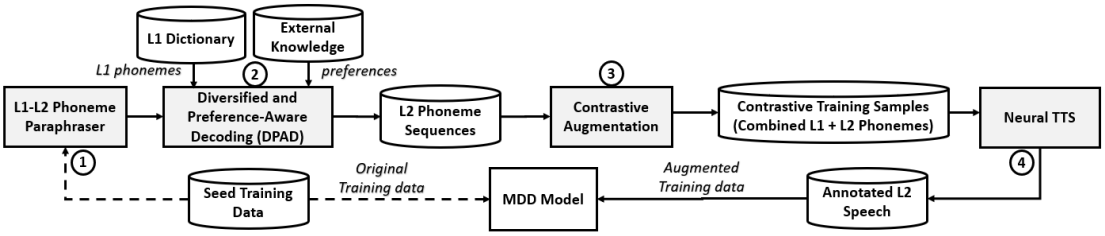

Figure 1: *System Architecture of The L2-GEN Framework*

2. The L2-GEN Framework

2.1. Overview

The L2-GEN framework is composed of four major components as illustrated in Figure 1: ① a Seq2Seq-based L1-L2 phoneme paraphrasing model that generates realistic L2 mispronounced phonemes of a word given its L1 reference phonemes; ② a diversified and preference-aware decoding (DPAD) module that incorporates external knowledge to produce diverse and generalizable L2 phoneme sequences; ③ a contrastive data augmentation module that samples and combines L1 and L2 phonemes into contrastive training pairs to improve MDD's sensitivity towards confusing phonemes; and ④ a neural Text-to-Speech (TTS) module that generates the audio for the generated mispronounced phoneme sequences. We present the details of each module below.

2.2. Seq2Seq Model for L1-L2 Phoneme Paraphrasing

The first and most critical stage of L2-GEN is to generate mispronounced L2 phoneme sequences that mimic L2 speech. We found the L2 phoneme generation problem can be mapped to the paraphrasing problem in natural language generation (NLG) domain. In specific, we can treat a L1 reference phoneme sequence of a word as the source text and its corresponding mispronounced L2 phoneme sequences as "paraphrased" target texts. In this work, we adapt the Transformer-based Seq2Seq machine translation model [20], which is the state-of-the-art model for generating paraphrases. To improve the generalizability to handle unseen words, we utilize the idea of character tokenization [21], and tokenize L1-L2 phoneme sequence pairs at the monophone level to avoid out-of-vocabulary cases. We show the example tokenized L1-L2 pairs for training and inference in Table 1 and how it handles an unseen word during inference. A small corpus of L1-L2 phoneme sequence pairs (referred to as "seed training data" in Figure 1) is needed to train the Seq2Seq model.

	Text	Source (L1)	Target (L2)
Training	Apple	/æ p ə l/	/æ b ə l/
	Rabbit	/r æ b ə t/	/l̲ æ b ə t/
Inference	Rampant (unseen)	/r æ m p ə n t/	/l̲ æ m b̲ ə n t/

Table 1: *Example L1-L2 Data for Training and Inference.*

2.3. Diversified and Preference-Aware Decoding

Next, we present the diversified and preference-aware decoding (DPAD) module to further improve generalizability for the generated L2 phoneme sequences. A fundamental way to improve generalizability of a sequence generation model is by diversifying the generated sequences [22]. Following this idea, instead of having a single L2 phoneme output, we produce the top L2 phoneme candidates from the Seq2Seq model by extending the beam search (BS) decoding algorithm.

The BS algorithm stores the top-K phoneme candidates $P_{[t-1]} = \{p_{1,[t-1]}, p_{2,[t-1]}, ..., p_{K,[t-1]}\}$ before each decoding time step t; where K is known as the beam width. Then, BS will enumerate all possible monophone extensions (denoted as $P_{[t-1]} \times \mathcal{N}$) of each of these K partial phoneme candidates, and pick the top-K candidate phonemes for the next step:

$$P_{[t]} = \underset{p_{1,[t]}, p_{2,[t]}, \cdots, p_{K,[t]} \in P_{[t]} \times \mathcal{N}}{\operatorname{argmax}} \sum_{k \in K} \Theta(p_{k,[t]}) \quad (1)$$

where $\mathcal{N}$ is the phoneme set. $\Theta(p_{k,[t]})$ is the log-likelihood of generating $p_{k,[t]}$ given a L1 reference phoneme sequence ($\mathcal{R}$) and previously generated phonemes $P_{[t-1]}$:

$$\Theta(p_{k,[t]}) = log\ Pr(p_{k,[t]} | P_{[t-1]}, \mathcal{R}) \quad (2)$$

The key limitation for this BS formulation is that the top-K phoneme candidates are observed to result in a lack of diversity (i.e., consists of similar mispronunciation patterns that are dominant in the training data) [23], making the L2 sequences hard to generalize to under-represented pronunciation patterns.

To introduce diversity, we adopt the diversified beam search technique [24]. We first define G diversified beam groups, each of which has beam width of $K' = |K/G|$. The intuition is to ensure that a phoneme will have a higher chance to be selected if it is different from phonemes in all other groups, consequently creating more diversity. Specifically, during the decoding step for group $g \in G$, for each candidate monophone p, we define a *diversity term* (Δ) as the dissimilarity between this monophone and the predicted phonemes in all other diversity groups ($\bar{g}$):

$$\Delta(p_{[t]}, \bar{g}) = \lambda \cdot \sum_{h \in \bar{g}} \sum_{p \in T^h_{[t]}} dist(p, p_{[t]}) \quad (3)$$

where λ is a normalization scalar that maps $\Delta(p_{[t]}, \bar{g})$ to [0,1]. $dist(\cdot)$ is the edit distance between two phoneme sequences.

We then introduce a novel *preference term* to regulate BS to be biased towards specific mispronunciation patterns. The intuition for introducing the preference term is two-fold: 1) it allows L2-GEN to produce mispronunciation patterns that are locale-specific (e.g., common mispronunciation patterns for native Mandarin speakers learning English); and 2) it enables generalizability towards unseen mispronunciation patterns if L2 data for certain locale is under-represented or absent in the training data. We first define "preference" as a set of M mispronunciation patterns $C =< r_1, e_1 >, < r_2, e_2 >, ... < r_m, e_m >$, where r and e refers to reference phoneme and mispronounced phoneme respectively. These mispronunciation patterns are collected from external ESL knowledge base that summarizes locale-specific mispronunciation patterns (e.g., common En-

glish mispronunciation patterns for L1 Mandarin speakers [25]). We then define the preference term $\Phi(C, p_{[t]})$ as whether a candidate phoneme $p_{[t]}$ satisfies preference C. Formally we have:

$$\Phi(C, p_{[t]}) = \begin{cases} 1 & \text{if} < p'_{[t]}, p_{[t]} > \in C \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $p'_{[t]}$ is the source L1 phoneme(s) corresponding to $p_{[t]}$ at decoding step t, which can be derived via source-target alignment based on attention weights [26].

Combining the beam search formulation in Eq. 1, the diversity term in Eq. 3, and the preference term in Eq. 4, the objective of the decoding process can be formulated as:

$$P^g_{[t]} = \underset{p^g_{1,[t]}, p^g_{2,[t]}, \ldots, p^g_{K',[t]} \in P^g_{[t]} \times \mathcal{N}}{\text{argmax}} \sum_{k \in K'} \Theta(p^g_{k,[t]}) + \Delta(p^g_{k,t}\bar{g}) + \beta \cdot \Phi(C, p^g_{k,t}),\ s.t.\ p^g_{i,t} \neq p^g_{j,t} \quad (5)$$

where $P^g_{[t]}$ is the decoded phoneme sequence at step t in diversity group g. β is a tunable weight that models the trade-off between diversity and preference. Jointly optimizing Eq. 5 for all K candidates and for all groups is intractable as the search space grows with $|\mathcal{N}|^K$. Following [24], DPAD greedily optimizes each diversity group detailed in Alg. 1.

Algorithm 1 Diversified and Preference-Aware Decoding

1: Perform a DPAD decoding with G groups using a beam width of K.
2: **for** $1 \leq t \leq T$ **do**
3: //Perform normal BS for group 1
4: $P^1_{[t]} \leftarrow \text{argmax}_{p^1_{1,[t]}, \ldots, p^1_{K',[t]} \in P^1_{[t]} \times \mathcal{V}} \Theta(p^1_{k,[t]})$
5: **for** $2 \leq g \leq G$ **do**
6: //Update log-probabilities with diversity and preference terms
7: $\Theta(p^g_{k,[t]}) \leftarrow \Theta(p^g_{k,[t]}) + \Delta(p^g_{k,t}\bar{g}) + \beta \cdot \Phi(C, p^g_{k,t})$
8: //Run normal BS within the group
9: $P^g_{[t]} \leftarrow \text{argmax}_{p^g_{1,[t]}, \ldots, p^1_{K',[t]} \in P^g_{[t]} \times \mathcal{V}} \Theta(p^g_{k,[t]})$
10: **end for**
11: **end for**
12: Return K solutions $\cup_{g=1}^{G} P^g_{[T]}$

2.4. Data Augmentation with Contrastive Samples

Next, we present a simple and effective contrastive data augmentation approach to improve MDD performance by using the L2 phoneme sequences generated by the proposed Seq2Seq model above. The idea is to concatenate L1 reference phoneme sequence of a word and its synthetically generated L2 (mispronounced) phoneme counterparts together as a single sentence and serve as one training sample for MDD. The idea follows adversarial sampling in contrastive learning [27] which improves model's sensitivity in distinguish pronunciation differences by combining confusing samples together during training. For example, assuming L2-GEN produces generates L2 samples for input "/maθ/" (math) as "/mas/" and "/mæs/" as illustrated in Figure 2. Then the concatenated phoneme sequence **/maθ. mæs. mæs/** will be used as the training data to help the MDD model to distinguish the sound difference for /s/ vs. /θ/, and /a/ vs. /æ/. We use all L1 reference phoneme sequences from CMU Pronouncing Dictionary (cmudict) [28] as input for DPAD to generate L2 mispronunciations.

2.5. L2 Speech Synthesis

Finally, L2-GEN leverages TTS system to generate the corresponding audio to the contrastive phoneme sequences generated above for MDD training. We utilizes Amazon-proprietary Polly TTS system [29, 30] that can take a custom phoneme sequence as input and generate naturally sounding speech using a neural vocoder [31]. We select 7 difference speaker profiles including 4 female, 1 male adult, and 2 male kids.

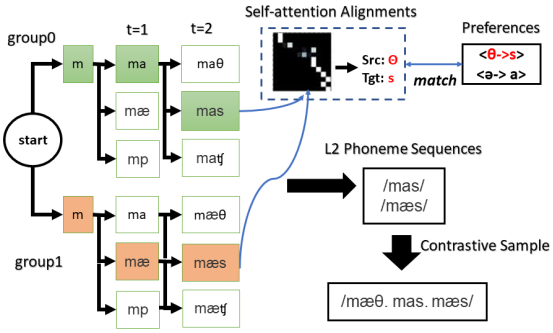

Figure 2: *DPAD and Contrastive Augmentation (K = 2, |G| = 2, L1 = /mæθ/). In step t=1, the diversity term allows DPAD to generate diverse L2 pronunciations /a/ vs. /æ/ for each group. In step t=2, both groups choose /s/ due to the preference term.*

3. Experiments

3.1. Dataset and Metrics

We consider two evaluation scenarios: *in-domain* scenario where training and test sets have L2 data from the same locales of ESL learners. *Out-of-domain* scenario mimics a locale expansion situation where no L2 training data is provided for a new locale. We use L2-ARCTIC [13] as our in-domain L2 dataset. The L2-ARCTIC training set is used to both train MDD model as well as serving as the seed training data for L2-GEN. To emulate a data sparsity scenario, we randomly selected 20% of L2-ARCTIC training data. We added extra TIMIT corpus as training data [32] which is commonly used in MDD training. We use the test split of the SpeechOcean762 [19] as out-of-domain test set, representing a new learning population whose native language is Mandarin[1]. All speech data was downsampled to 16 kHz. The dataset statistics are summarized in Table 2. Following previous work [14, 6], we chose False Rejection Rate (FRR), False Acceptance Rate (FAR), Precision, Recall, and F1-score as evaluation metrics for MDD performance.

	TIMIT	L2-ARCTIC			SpeechOcean
	Train	Train	Dev	Test	Test
Speakers	630	12	6	6	125
Hours	4.5	1.84	0.94	0.88	2.28
Utterances	6300	1800	897	900	2500
L1 Language	English	Spanish, Hindi, Korean, Mandarin[1], Arabic,Vietnamese			Mandarin

Table 2: *Dataset Summary.*

3.2. Model Training and Setup Details

We adopt the state-of-the-art end-to-end MDD model proposed by [33]. The model consists of a Wave2Vec2.0-base encoder and decoder with the CTC loss function. The model was trained for 100 epochs. We use a Transformer-based Seq2Seq model by [20] that has 6 layers for both encoder and decoder. Each of the self-attention mechanism contains 8 heads. The model is trained with a dropout of 0.15 and cross-entropy loss function. We set the DPAD's default preference weight β = 1 for

[1] Mandarin training samples removed for out-of-domain scenario.

out-of-domain scenario and $\beta = 0$ for in-domain, beam width $K = 6$, and groups $G = 2$. The parameters were heuristically selected that optimized the F1-score on L2-ARCTIC dev set. To incorporate preference, we use external knowledge of common mispronunciation patterns for native Mandarin speakers [25]. Both Seq2Seq and MDD models were trained on a single Nvidia Tesla V100 GPU.

Candidate	L2-ARCTIC (In-Domain)				
	FAR	FRR	Pre	Rec	F1
None	0.440	0.091	0.523	0.541	0.532
SIMP-AUG	0.422	0.103	0.538	0.543	0.541
P2P-SMT	**0.418**	0.120	0.528	0.543	0.535
P2P-Replace	0.426	0.097	0.529	0.543	0.536
L2-GEN	0.423	**0.094**	**0.557**	**0.548**	**0.553**

Table 3: *Data Augmentation Performance for L2-ARCTIC.*

Candidate	SpeechOcean762 (New Domain)				
	FAR	FRR	Pre	Rec	F1
None	0.571	0.158	0.490	0.463	0.476
SIMP-AUG	0.565	0.155	**0.499**	0.470	0.484
P2P-SMT	0.582	0.161	0.485	0.460	0.472
P2P-Replace	0.563	0.161	0.490	0.465	0.477
L2-GEN	**0.558**	**0.152**	0.498	**0.502**	**0.500**

Table 4: *Data Augmentation Performance for SpeechOcean.*

3.3. MDD Accuracy w.r.t Data Augmentation Algorithms

We first evaluate the effect of L2-GEN by comparing it with existing state-of-the-art data augmentation algorithms. We chose the following baselines: *None* that has no data augmentation applied; *SIMP_AUG* in [14] that combines a set of heuristic phoneme replacement strategies including replacing consonants and confusing pairs; *P2P-SMT* in [34] that applies a statistical machine translation (SMT) model originally designed for dialect conversion; and *P2P-Replace* that randomly replaces phonemes. To have a fair comparison for all candidates, we generate 20,000 synthetic L2 data using the same contrastive augmentation strategy proposed in this paper. For the out-of-domain scenario, we removed Mandarin speakers in the L2-ARCTIC training data to emulate the scenario where no L2 data is available for Mandarin ESL learners.

The results are listed in Tables 3 and 4. For in-domain scenario, L2-GEN outperforms all baselines in terms of FRR, Precision, Recall, and F1-scores. In particular, it has achieved 3.93% F1-score improvement compared to no data augmentation. For the domain transfer scenario, the L2-GEN performance improves by a wider margin over the other baselines and results in 5.0% gain over the no data augmentation candidate. The performance gaps between L2-GEN and other baselines also increase. We attribute this finding to the fact that L2-GEN is able to generate mispronunciation patterns that are specific to Mandarin ESL learners thanks to the preference term in the DAPD process. In contrast, the heuristic algorithms proposed in baselines are not designed to generalize to new domain. The above findings highlight that L2-GEN can produce generalizable L2 training data to better improve MDD in both data sparse in-domain scenario as well as when adapting to a new locale.

3.4. Ablation Study

We now investigate which elements of L2-GEN contribute the most to its performance. Along with the L2-GEN framework we trained additional variants each with a certain feature removed - including the diversity term (w/o Diversity), the preference term (w/o Preference). We also remove the contrastive augmentation by randomly concatenating L2 phoneme sequences as augmented data (w/o Contrastive). The results are summarized in Table 5. Note that, for in-domain scenario, we discard the preference term because the mispronunciation patterns already exist in training data, whereas in out-of-domain scenario, we need the preference term to inject mispronunciation patterns for Mandarin speakers which are absent in training. We can observe that all components contribute to the performance improvements whereas the diversity term and preference term contribute the most in in-domain and out-of-domain scenario respectively. The contrastive augmentation has the least improvement contribution in both cases.

Model	FAR	FRR	F1
L2-ARCTIC (In-Domain)			
w/o Diversity	0.436	0.101	0.540(-2.24%)
w/o Contrastive	0.429	0.099	0.543(-1.77%)
L2-GEN Full	0.423	0.094	0.553
SpeechOcean762 (New Domain)			
w/o Diversity	0.563	0.157	0.499(-0.91%)
w/o Preference	0.569	0.165	0.484(-3.81%)
w/o Contrastive	0.557	0.155	0.501(-0.54%)
L2-GEN Full	0.558	0.152	0.503

Table 5: *Ablation Study.*

3.5. Robustness w.r.t. MDD Model Choice

Finally, we study L2-GEN's robustness against different MDD model choices. Here, we chose variations of Wav2vec2.0 based models including Wav2Vec2.0$_{Large}$, and Wav2Vec2.0$_{XLSR}$ [35], as well as CTC-ATT that has an alternative MDD model architecture using bi-LSMT + Attention mechanism [14]. We found, the FRR rate for CTC-ATT became worse. We attribute to the fact that contrastive sampling causes the CTC-ATT model to be too sensitive towards pronunciation differences, thus causing higher false rejects. However in general, the synthetic L2 data generated by L2-GEN can consistently improve MDD F1-score for all candidate models. The results suggests that L2-GEN can be generalizable to improve different MDD models.

MDD Model Type		FAR	FRR	F1
Wav2Vec2.0$_{Large}$	no aug.	0.440	0.102	0.530
	L2-GEN	**0.439**	**0.098**	**0.543**
Wav2Vec2.0$_{XLSR}$	no aug.	0.435	0.089	0.546
	L2-GEN	**0.421**	0.089	**0.558**
CTC-ATT	no aug.	0.524	0.087	0.476
	L2-GEN	**0.493**	0.094	**0.490**

Table 6: *Robustness w.r.t. MDD Models.*

4. Conclusions

In this paper, we present L2-GEN, a novel framework for generating annotated synthetic L2 Speech generation framework. Compared to existing work, L2-GEN can synthesize realistic L2 phonemes sequences by building a novel Seq2Seq phoneme paraphrasing model. A diversified and preference-aware decoding scheme was proposed to improve the generalizability of L2-GEN by jointly diversifying the beam search process as well as leveraging known mispronunciation patterns from external knowledge. Empirical experiments have demonstrated the effectiveness of L2-GEN in improving MDD accuracy with only limited L2 data and when transferring to a new locale.

5. References

[1] C. Agarwal and P. Chakraborty, "A review of tools and techniques for computer aided pronunciation training (capt) in english," *Education and Information Technologies*, vol. 24, no. 6, pp. 3731–3743, 2019.

[2] 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.

[3] W.-K. Lo, S. Zhang, and H. Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system," in *Eleventh annual conference of the international speech communication association*, 2010.

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

[5] W.-K. Leung, X. Liu, and H. Meng, "Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8132–8136.

[6] 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," *arXiv preprint arXiv:2005.11950*, 2020.

[7] B.-C. Yan and B. Chen, "End-to-end mispronunciation detection and diagnosis from raw waveforms," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 61–65.

[8] 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.

[9] P. Bonaventura, P. Howarth, and W. Menzel, "Phonetic annotation of a non-native speech corpus," in *Proceedings International Workshop on Integrating Speech Technology in the (Language) Learning and Assistive Interface, InStil*, 2000, pp. 10–17.

[10] S. Loewen, *Introduction to instructed second language acquisition*. Routledge, 2014.

[11] D. Korzekwa, J. Lorenzo-Trueba, T. Drugman, S. Calamaro, and B. Kostek, "Weakly-supervised word-level pronunciation error detection in non-native english speech," *arXiv preprint arXiv:2106.03494*, 2021.

[12] A. Lee, N. F. Chen, and J. Glass, "Personalized mispronunciation detection and diagnosis based on unsupervised error pattern discovery," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6145–6149.

[13] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-arctic: A non-native english speech corpus." in *INTERSPEECH*, 2018, pp. 2783–2787.

[14] K. Fu, J. Lin, D. Ke, Y. Xie, J. Zhang, and B. Lin, "A full text-dependent end to end mispronunciation detection and diagnosis with easy data augmentation techniques," *arXiv preprint arXiv:2104.08428*, 2021.

[15] A. Lee *et al.*, "Language-independent methods for computer-assisted pronunciation training," Ph.D. dissertation, Massachusetts Institute of Technology, 2016.

[16] L. Loots and T. Niesler, "Automatic conversion between pronunciations of different english accents," *Speech Communication*, vol. 53, no. 1, pp. 75–84, 2011.

[17] D. Korzekwa, J. Lorenzo-Trueba, S. Zaporowski, S. Calamaro, T. Drugman, and B. Kostek, "Mispronunciation detection in non-native (l2) english with uncertainty modeling," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7738–7742.

[18] 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.

[19] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "Speechocean762: an open-source non-native english speech corpus for pronunciation assessment," 2021.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] Y. Tay, V. Q. Tran, S. Ruder, J. Gupta, H. W. Chung, D. Bahri, Z. Qin, S. Baumgartner, C. Yu, and D. Metzler, "Charformer: Fast character transformers via gradient-based subword tokenization," 2021.

[22] O. Rozen, V. Shwartz, R. Aharoni, and I. Dagan, "Diversify your datasets: Analyzing generalization via controlled variance in adversarial datasets," *arXiv preprint arXiv:1910.09302*, 2019.

[23] S. Jiang and M. de Rijke, "Why are sequence-to-sequence models so dull? understanding the low-diversity problem of chatbots," *arXiv preprint arXiv:1809.01941*, 2018.

[24] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse beam search: Decoding diverse solutions from neural sequence models," *arXiv preprint arXiv:1610.02424*, 2016.

[25] F. Han, "Pronunciation problems of chinese learners of english." *ORTESOL Journal*, vol. 30, pp. 26–30, 2013.

[26] G. Chen, Y. Chen, and V. O. Li, "Lexically constrained neural machine translation with explicit alignment guidance," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, 2021, pp. 12 630–12 638.

[27] C.-H. Ho and N. Nvasconcelos, "Contrastive learning with adversarial examples," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 081–17 093, 2020.

[28] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[29] A. Ezzerg, A. Gabrys, B. Putrycz, D. Korzekwa, D. Saez-Trigueros, D. McHardy, K. Pokora, J. Lachowicz, J. Lorenzo-Trueba, and V. Klimkov, "Enhancing audio quality for expressive neural text-to-speech," *arXiv preprint arXiv:2108.06270*, 2021.

[30] R. Shah, K. Pokora, A. Ezzerg, V. Klimkov, G. Huybrechts, B. Putrycz, D. Korzekwa, and T. Merritt, "Non-Autoregressive TTS with Explicit Duration Modelling for Low-Resource Highly Expressive Speech," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 96–101.

[31] Y. Jiao, A. Gabryś, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, "Universal neural vocoding with parallel wavenet," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6044–6048.

[32] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

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

[34] P. Karanasou and L. Lamel, "Comparing smt methods for automatic generation of pronunciation variants," in *International Conference on Natural Language Processing*. Springer, 2010, pp. 167–178.

[35] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.