



# Cross-Lingual Transfer Learning Approach to Phoneme Error Detection via Latent Phonetic Representation

Jovan Dalhouse<sup>1</sup>, Katunobu Ito<sup>2</sup>

<sup>1</sup>Institute of Integrated Science and Technology, Hosei University, Japan

<sup>2</sup>Graduate School of Computer and Information Science, Hosei University, Japan

jovan.dalhouse.3x@stu.hosei.ac.jp, itou@hosei.ac.jp

## Abstract

Extensive research has been conducted on CALL systems for Pronunciation Error detection to automate language improvement through self-evaluation. However, many of these previous approaches have relied on HMM or Neural Network Hybrid Models which, although have proven to be effective, often utilize phonetically labelled L2 speech data which is expensive and often scarce. This paper discusses a “zero-shot” transfer learning approach to detect phonetic errors in L2 English speech by Japanese native speakers using solely unaligned phonetically labelled native language speech. The proposed method introduces a simple base architecture which utilizes the XLSR-Wav2Vec2.0 model pre-trained on unlabelled multilingual speech. Phoneme mapping for each language is determined based on difference of articulation of similar phonemes. This method achieved a Phonetic Error Rate of 0.214 on erroneous L2 speech after fine-tuning on 70 hours of speech with low resource automated phonetic labelling, and proved to additionally model phonemes of the native language of the L2 speaker effectively without the need for L2 speech fine-tuning.

**Index Terms:** Speech Recognition, Computer Assisted Pronunciation Training (CAPT), Computer Assisted Language Learning (CALL), L2 speech

## 1. Introduction

In Japan, native practical assistance for English pronunciation can often be difficult to come by. It is for this reason that there has been great demand for an automated means of improving one’s pronunciation independently without native assistance. Research aimed at improving pronunciation quality through Automatic Speech Recognition (ASR) techniques has grown increasingly popular in recent years. This includes those aimed specifically for Japanese native speakers.

A notable contribution to the field of phonetic error detection in Computer Assisted Pronunciation Training systems (CAPT) is [1] which introduced the “Goodness of Pronunciation” score which determines whether or not individual phonemes in an utterance are correct based on their posterior probabilities. The use of posterior probabilities in CAPT systems would go on to be further developed in research [2, 3] where native English and non-native (L2) speech trained models were used to improve its mispronunciation detection via Hidden Markov Models (HMM) and Neural Network Hybrid Models. The use of HMMs in these works were proven beneficial in their ability to produce accurate segmentation of phonemes.

However, in recent years, Neural Network models equipped with Connectionist Temporal Classification loss have shown success in accurate phone segmentation comparable to that of HMMs without the need for time-aligned speech data [4]. In such methods CTC Loss is often used in combination with

attention-based models such as Recurrent Neural Networks or LSTMs for end-to-end approaches as seen in [5]. While such models have shown high performance, similar to HMM-based models, many often still have shortcomings which limit their performance. These include (1) Need for phonetically labelled L2 speech data, and (2) Dependence on pre-defined error patterns for classification.

The reliance on labelled L2 speech for modelling erroneous pronunciation in research such as [3, 6, 7] require large amounts of data and are therefore dependent on its availability. This can create difficulty when training such models as, depending on the target language, L2 speech corpora from a specific native language group can be scarce or non-existent. Additionally, labelling such corpora manually can pose more of a challenge due to the ambiguity of certain distortion-based mispronunciation errors [8]. As manual labelling is a subjective task, this causes inconsistency in such corpora which can throttle the performance of the resulting model [9].

The goal of this research is to develop a base Neural Network pipeline trained on only native speech data to transcribe and discern phonetic errors in L2 speech using a Cross-lingual “zero-shot” Transfer-learning approach. This was done using the XLS-R Wav2Vec2.0 pre-trained model. This model has been pre-trained on over 300 million unlabeled raw speech samples from 128 different languages and in previous studies has shown promising results in zero-shot approaches in phonetic transcription of unseen languages outside of the tuning set [10]. We take on a similar approach for error detection in L2 speech, however in this research the models were tuned using native speech from the respective native and target language pairs. The data used in this study are labelled using phone mappings based on similarity of articulation between common phonemes in both languages similar to the method used in [11].

Linguistic research has shown that errors in pronunciation are not random errors but a reflection of patterns which originate from the speaker’s mother tongue which they have adapted to [12]. Considering this, the proposed model trained on the user’s native language not only benefits through increasing availability of applicable data for training, but also increases the capability of modelling native phonemes from both languages without the need for predefined error patterns. While it has been argued that phonetic errors contribute less to the intelligibility of an L2 Learner’s pronunciation when compared to prosodic features such as stress, the improvement of such errors, particularly phonetic errors common to Japanese-native L2 learners (see Table 1), are still an essential foundation in pronunciation improvement. To show the flexibility of the proposed system, variations of the model trained on high resource (manual labelling) and low resource (automatic labelling) data were evaluated on L2 utterances from the SRC UME-ERJ corpus.

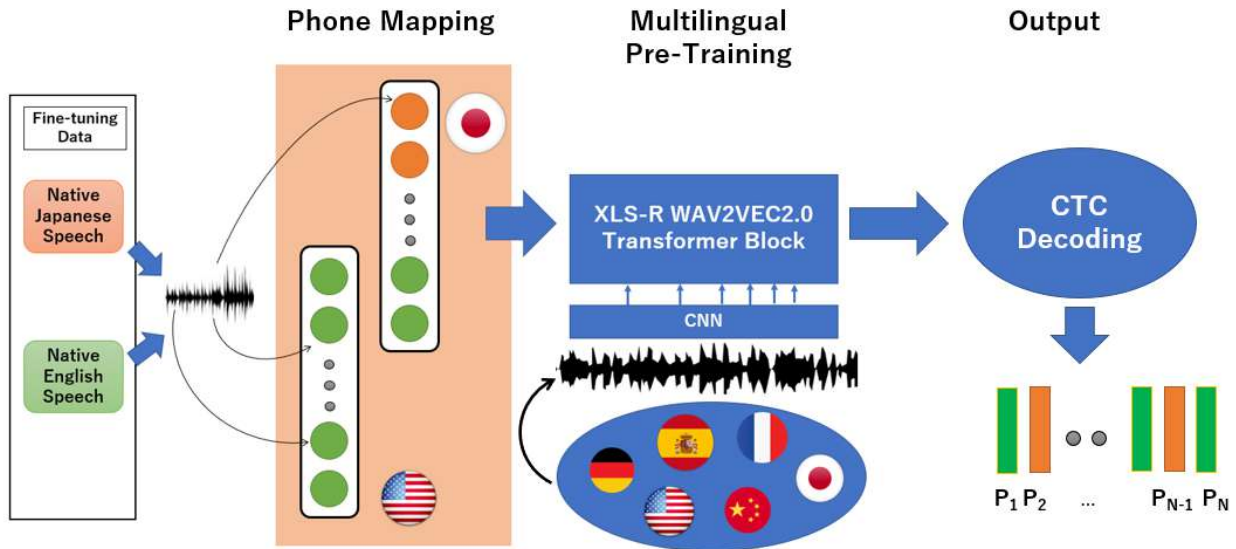


Figure 1: Illustrated overview of the proposed system using XLS-R pretrained WAV2VEC2.0 Model

## 2. Method

### 2.1. Wav2vec2.0 (XLS-R)

Wav2vec2.0 [13] performs ASR tasks through a Transformer model architecture initially pretrained on hours of unlabeled raw speech. The first stage of the pretraining process involves encoding raw speech data into a normalized latent representation using a feature encoder composed of multiple convolutional blocks and GELU activation. This latent output is then quantized into fixed representations, using mapped entries from multiple code books. Code book entry selection is learned using a differentiable Gumbel SoftMax function. The transformer learns through the contrastive task of accurately identifying the true quantized representation amongst “distractors” of randomly masked sequences of the feature encoder output. These tasks are propagated through a contrastive loss, which optimizes the model’s ability to learn representations in the latent speech, and a diversity loss which promotes the equal use of multiple code book entries. Through this process, wav2vec2.0 is able to learn representations of unlabeled data during pretraining.

In the case of the XLS-R-300M wav2vec model, over 300 million unlabeled speech samples spanning over 128 different languages were used to pretrain the model. This self-supervised approach through multilingual pre-training allows wav2vec2.0 to adapt to speech representations in different linguistic contexts and has shown to achieve high performance when fine-tuned on multiple languages to phonetically transcribe unseen languages [14]. For this research, raw feature vectors for only the native (Japanese) and target (English) language pairs will be used for inference. Contrary to [14], in respect to the scope of this task, phoneme mapping will be made based on shared articulatory similarities between phonemes in the designated language pair, rather than a global phone recognizer.

Fine tuning involves learning on labelled feature vector input on a randomly initialized output layer which is optimized using Connectionist Temporal Classification Loss (CTC) (see Figure 1). The sequence before decoding will be output as follows:

$$Out_{CTC} = P_E \cup P_J \cup \beta \quad (1)$$

Where  $P_E$ ,  $P_J$  and  $\beta$  represent English phonemes, Japanese Phonemes (see Section 2.2) and the CTC blank tokens respectively. The optimal path from the CTC output logits are decoded using a Greedy decoder algorithm.

### 2.2. Phonetic Mapping

Considering phonetic classification of our proposed method uses a single acoustic model with a shared output layer, similar to other multilingual approaches, phoneme classes of both languages used will be unified to avoid out-of-vocabulary (OOV) token errors [11, 10]. With respect to the language pairs used for inference, the mapping format for phonetic assignment is determined based on difference of articulatory features of similar phonemes which exist in both languages. In short, similar phonemes in both languages which have an almost allophonic relationship will be classified as a single phoneme class. This is done so as to not only reduce the complexity of the acoustic model, but to also reduce the possibility of a high False Rejection Rate (FRR) which is unfavourable for error identification in CAPT systems [15]. Phonetic mapping is outlined for each phonetic category as seen below.

**Vowels.** In the case of vowels, English is known to have up to 15 different vowel phonemes which include combined sounds such as /ow/ and /ay/, while the Japanese language consists of only 5 basic vowels. In addition to the number present in each language, each group of vowels differ in their degree of tenseness, where English vowels are more often lax compared to those in Japanese which require more muscle tension [16]. This difference in vowel tension contributes to difficulty in English vowel pronunciation by Japanese speakers which leads to substitution errors of /i/ and /iy/ such as in the words “bit” and “beat” respectively [17]. For this reason, all vowels in the native Japanese utterances were labelled as separate distinct phonemes from those in the native English corpora. Also, unlike Japanese, vowel duration in the English language is non-phonemic, mean-

ing that the identity of a phoneme does not change with its duration [18]. For this reason, no distinctions were made between long and short sounding Japanese vowels.

**Consonants.** Unlike vowels, consonants in Japanese share more similarities with those in English. While the existence of certain Japanese phonemes in the English language are sometimes questioned, such as the bilabial fricative in *fuji-san* and voiceless palatal fricative in *hito*, these phonemes were treated as allophones of the /f/ and /h/ in English respectively [16]. One major distinction that was made was the retro-flex liquid /r/ in Japanese which differs from the /r/ in the English language. While this /r/ is sometimes compared to the alveolar tap /ɾ/ in English (such as in *water*) [18], in this study these are treated as two separate phonemes.

### 3. Experiments

This section describes the experiments conducted using the different corpora to evaluate the error detection performance of the cross-lingual model as well as the flexibility of the model when trained on data with varying degrees of labelling quality. All corpora in this study contains read speech which, although has shown to have prosodic differences with spontaneous speech [19], is suitable for the training and evaluation of this model as it is compatible with the expected input speech for this text-dependent task.

#### 3.1. Corpora

##### 3.1.1. TIMIT

Native data from the TIMIT Acoustic-Phonetic Speech Corpus contains speech utterances from 630 native English speakers from 8 different regions of the United States. The TIMIT corpus has been used in ASR and phonetic research such as [20, 3] to train and evaluate models and is one of the few manually transcribed and time-aligned speech corpora available. This corpus is used as the native corpus to train and evaluate the phone error detection model’s performance on high resource data. For all experiments, post-consonantal closes and pauses were removed and allophones were merged to maintain consistency with other corpora [21]. Additionally, the alveolar tap /ɾ/ was merged with the phoneme /d/.

##### 3.1.2. UME-ERJ (Native English)

The UME-ERJ Native English Corpus contains approximately 6 hours of speech from 40 American native English speakers (equally split by gender) above the age of 20. This corpus consists of read speech from 400 sentences. Contrast to TIMIT, the utterances in this corpus are not phonetically labelled. For this experiment, the UME-ERJ corpus was phonetically labelled through automated forced alignment using the Penn’s Aligner HMM Toolkit software. Forced alignment was conducted using a modified phone dictionary which allowed for cases of the medial /t/ phoneme (such as the /t/ in *water*) to be classified as /d/ to represent the alveolar tap, as this is the condition where this phoneme is most likely to occur [22]. Although other variations of the /t/ phoneme exist, such as the vanishing /t/ after the alveolar nasal /n/, these changes were not included. Given the degree of accuracy of forced aligned labelling, this corpus is used to evaluate the model’s performance on lower resource labelled speech, compared to manual labelling.

Table 1: Category of error patterns which occur in the labelled utterances from the UME-ERJ L2 English corpus. **q** represents a pre-consonantal pause.

Error Pattern	Example
Substitution	let l/eh/t/- > rj/eh/t/
Phone Insertion	slash s/l/ae/sh/- > s/rj/aj/sh/uj/
Phone Skipping	bar b/aa/r/- > b/aa/
Pause Insertion	bat b/ae/t/- > b/ae/q/t/oj/

##### 3.1.3. Librispeech

The Librispeech corpus is a large speech corpus containing up to 1000 hours of spoken American native English. The data from this corpus comprises mainly of read speech from audiobooks. While some of the speech from Librispeech, similar to TIMIT, are relatively out-dated, the large quantity provided is essential in measuring the potential detection performance when trained on extended hours. For this experiment, a selected 50-hour subset of the Librispeech corpus was used. This data was transcribed automatically using word-phoneme mapping and will be used to evaluate the performance on low resource labelled speech.

##### 3.1.4. JNAS

The JNAS corpus contains over 16,000 speech samples from 306 Japanese native speakers (equally split by gender) which amounts to roughly 21 hours. This corpus contains speech read from a total of 503 phonetically balanced sentences originating from newspaper articles. For all experiments involving this corpus, phonemes were labelled automatically using phonetic mapping (similar to the Librispeech subset).

##### 3.1.5. UME-ERJ L2 English

For evaluation of the performance of each model the L2 English UME-ERJ corpus was used. This corpus contains English speech from Japanese Native University students and been used in ASR related tasks such as in [23]. For Error detection accuracy, a subset of the corpus containing 50 single word utterances was used. For this task utterances were selected based on 2 primary conditions. (1) Contain no more than 2 erroneous phonemes, and (2) Only errors with relatively low ambiguity.

This is done to lower the degree of subjectivity that is expected from labelling certain phonetic errors, such as an utterance which falls in between similar sounding phonemes (e.g., /ae/ and /aa/) [24]. Error patterns which occurred in the selected samples can be classified into four main categories (see Table 1) Additionally, the False Rejection Rate (FRR) will also be measured. For this evaluation native sounding Japanese utterances with no noticeable phonetic errors were used. These utterances were selected based on speakers that were evaluated with a perfect score via a mark sheet provided in the corpus documentation.

## 4. Results and Discussion

In this section, we show the evaluation results for the bilingually trained Wav2vec2.0-XLS-R models used in this study to measure the error detection performance, and detection sensitivity with respect to the hours of speech as well as the method of data labelling used. All models in this study were trained with a batch size of 8 and a learning rate of 0.001. The model checkpoints with the optimal average performance were selected from each experiment. As a result, the number of training steps as well as the proportion of speech samples for each language set varies.

In the primary experiment, models trained on different sets of data were compared (see Table 2). From these results we can see that the overall performance of the model increases with the total duration of speech used for inference. While the concept of performance improvement through an increase of training data is expected [25], this improvement occurs despite the decline in labelling quality of the utterances used.

Table 2: *Phone Error Rate (PER) and False Rejection Rate (FRR) for models trained on different corpora. UME - UME-ERJ Native Corpus, LS - Librispeech subset, TMT - TIMIT training set*

	TMT-JNAS (3 +3) 6hours	UME-TMT-JNAS (3+7+10) 20hours	LS-JNAS (51+21) 72hours
<b>PER</b>	0.68	0.59	<b>0.214</b>
<b>FRR</b>	0.74	0.5587	<b>0.1452</b>

This is shown as the 51 hour subset of Librispeech transcribed via word-phone mapping achieving a PER of 0.214, which is comparable to phone recognition results evaluated on the TIMIT corpus using previous approaches such as Bi-LSTM-CTC [26]. In another experiment, the evaluation performance of our bilingually trained models were compared with a wav2vec2.0-XLS-R model trained on the TIMIT training set only. In this experiment, models were evaluated on both the TIMIT test-set and the UME-ERJ L2 native-sounding subset (used to measure FRR)(see Table 3). Although the TIMIT trained model achieved the highest performance when evaluated on the TIMIT test set, the bilingually-trained models surpassed the TIMIT-trained model when evaluated on the phonetically fluent native-sounding L2 English utterances.

Table 3: *Evaluation of PER on TIMIT test set and FRR for TIMIT-trained and bilingually-trained wav2vec models.*

	TIMIT	LS-JNAS	UME-LS-JNAS
<b>TIMIT(test)</b>	<b>0.1337</b>	0.2843	0.2586
<b>FRR</b>	0.2589	0.1452	<b>0.127</b>
<b>Average</b>	0.1963	0.2147	<b>0.1928</b>

Accurate modelling of non-native utterances often utilizes speaker adaptation using accurate pronunciation speech from L2 learners to improve error detection accuracy [20]. However, these results show that by using the wav2vec2.0-XLS-R, fine-tuning models on native speech from the L2 users native language can be used as a form of language adaptation and attain sufficient results on L2 speech. Additionally, when adding speech data (UME-ERJ) to the Librispeech-JNAS

trained model, the average PER surpasses that of the TIMIT-trained model.

Table 4: *Error Detection Rate of the current system (Wav2Vec2.0-XLS-R) by category.*

	Total Errors	Detected (Wav2vec)
Substitution	71	68
Phone Insertion	18	16
Phone Skipping	5	5
Total	94	89
Detection Rate		<b>0.947</b>

Table 5: *Statistics of the detected errors of the LS-JNAS fine tuned model with respect to phoneme type (consonants & vowels)*

	Consonants		Vowels	
Substitution	44	41	27	27
Phone Insertion	0	0	18	16
Phone Skipping	4	4	1	1
Total	48	45	46	44

Table 4 shows the number of errors successfully identified with respect to the total number of errors transcribed for each error pattern. Contrast to the PER metric which measures the classification error rate with respect to the manually transcribed labels, this evaluation includes substitution error classifications that are similar to the labelled substitution errors. For instance, in the case of the word "rat" ( $r/a/t$ ), a misclassification of the label  $rj/a/t$  as  $d/a/t$  would still be accepted as an accurate identification of this substitution error due to the fact that both  $/d/$  and  $/rj/$  (Japanese  $/r/$ ) are both alveolar type phonemes with similar articulation and thereby lead to some degree of subjectivity in their discernment (see Table 5 for error detection statistics by phoneme type). From these results, it can be seen that the Librispeech-JNAS model has a high detection rate of common Japanese error patterns.

## 5. Conclusions

This research proposes a Cross-Lingual Transfer Learning approach to phonetic error detection in L2 English speech by Japanese speakers using a multilingual wav2vec2.0-XLS-R model fine-tuned on solely native English and Japanese speech. With our proposed model trained on 70 hours total speech, we achieve a low Phone Error Rate on erroneous L2 speech without any L2 fine-tuning. Although this approach can be seen as data-intensive requiring hours of speech to attain high performance, requiring only native data which is more easily obtainable greatly reduces the challenge of sufficiently fine-tuning the model. Additionally, from the results we also show that this can also be accomplished using speech labelled through low-resource automated methods which further broaden the range of accessible data for model fine-tuning.

## 6. References

- [1] 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.

- [2] L. Neumeier, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech communication*, vol. 30, no. 2-3, pp. 83–93, 2000.
- [3] O. Ronen, L. Neumeier, and H. Franco, "Automatic detection of mispronunciation for language instruction." in *EUROSPEECH*. Citeseer, 1997.
- [4] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
- [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] S.-Y. Yoon, M. Hasegawa-Johnson, and R. Sproat, "Landmark-based automated pronunciation error detection," in *Eleventh annual conference of the international speech communication association*, 2010.
- [7] Z. Zhang, Y. Wang, and J. Yang, "Text-conditioned transformer for automatic pronunciation error detection," *Speech Communication*, vol. 130, pp. 55–63, 2021.
- [8] R. Duan, J. Zhang, W. Cao, and Y. Xie, "A preliminary study on asr-based detection of chinese mispronunciation by japanese learners," in *Fifteenth annual conference of the international speech communication association*, 2014.
- [9] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 6135–6139.
- [10] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," *arXiv preprint arXiv:2109.11680*, 2021.
- [11] X. Li, S. Dalmia, D. Mortensen, J. Li, A. Black, and F. Metze, "Towards zero-shot learning for automatic phonemic transcription," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8261–8268.
- [12] R. J. DiPietro, "Learner english: A teacher's guide to interference and other problems. michael swan and bernard smith (eds.). new york: Cambridge university press, 1987. pp. xv+ 265." *Studies in Second Language Acquisition*, vol. 10, no. 3, pp. 406–407, 1988.
- [13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.
- [14] 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.
- [15] Y. Gao, Y. Xie, W. Cao, and J. Zhang, "A study on robust detection of pronunciation erroneous tendency based on deep neural network," in *Sixteenth annual conference of the international speech communication association*, 2015.
- [16] K. Ohata, "Phonological differences between japanese and english: Several potentially problematic," *Language learning*, vol. 22, pp. 29–41, 2004.
- [17] M. Jo Martens, "Kenworthy, joanne. teaching english pronunciation. london and new york: Longman, 1987," *Canadian Modern Language Review*, vol. 47, no. 4, pp. 802–804, 1991.
- [18] T. Riney and J. Anderson-Hsieh, "Japanese pronunciation of english," *JALT Journal*, vol. 15, no. 1, pp. 21–36, 1993.
- [19] P. Howell and K. Kadi-Hanifi, "Comparison of prosodic properties between read and spontaneous speech material," *Speech communication*, vol. 10, no. 2, pp. 163–169, 1991.
- [20] A. Ito, T. Nagasawa, H. Ogasawara, M. Suzuki, and S. Makino, "Automatic detection of english mispronunciation using speaker adaptation and automatic assessment of english intonation and rhythm," *Educational technology research*, vol. 29, no. 1-2, pp. 13–23, 2006.
- [21] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.
- [22] V. W. Zue and M. Laferriere, "Acoustic study of medial/t, d/in american english," *The Journal of the Acoustical Society of America*, vol. 66, no. 4, pp. 1039–1050, 1979.
- [23] K. Radzikowski, L. Wang, O. Yoshie, and R. Nowak, "Accent modification for speech recognition of non-native speakers using neural style transfer," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–10, 2021.
- [24] G. Kawai, "Detecting and correcting mispronunciation in non-native pronunciation learning using a speech recognizer incorporating bilingual phone models," *Journal of the acoustical society of Japan*, 2001.
- [25] H. Strik, K. Truong, F. De Wet, and C. Cucchiari, "Comparing different approaches for automatic pronunciation error detection," *Speech communication*, vol. 51, no. 10, pp. 845–852, 2009.
- [26] S. Fernández, A. Graves, and J. Schmidhuber, "Phoneme recognition in timit with blstm-ctc," *arXiv preprint arXiv:0804.3269*, 2008.