MISPRONUNCIATION DETECTION IN NON-NATIVE (L2) ENGLISH WITH UNCERTAINTY MODELING

Daniel Korzekwa[⋆†], *Jaime Lorenzo-Trueba*[⋆], *Szymon Zaporowski*[†],
Shira Calamaro[⋆], *Thomas Drugman*[⋆], *Bozena Kostek*[†]

[⋆] Amazon Speech [†] Gdansk University of Technology, Faculty of ETI, Poland

ABSTRACT

A common approach to the automatic detection of mispronunciation in language learning is to recognize the phonemes produced by a student and compare it to the expected pronunciation of a native speaker. This approach makes two simplifying assumptions: a) phonemes can be recognized from speech with high accuracy, b) there is a single correct way for a sentence to be pronounced. These assumptions do not always hold, which can result in a significant amount of false mispronunciation alarms. We propose a novel approach to overcome this problem based on two principles: a) taking into account uncertainty in the automatic phoneme recognition step, b) accounting for the fact that there may be multiple valid pronunciations. We evaluate the model on non-native (L2) English speech of German, Italian and Polish speakers, where it is shown to increase the precision of detecting mispronunciations by up to 18% (relative) compared to the common approach.

Index Terms— Pronunciation Assessment, Second Language Learning, Uncertainty Modeling, Deep Learning

1. INTRODUCTION

In Computer Assisted Pronunciation Training (CAPT), students are presented with a text and asked to read it aloud. A computer informs students on mispronunciations in their speech, so that they can repeat it and improve. CAPT has been found to be an effective tool that helps non-native (L2) speakers of English to improve their pronunciation skills [1, 2].

A common approach to CAPT is based on recognizing the phonemes produced by a student and comparing them with the expected (canonical) phonemes that a native speaker would pronounce [3, 4, 5, 6]. It makes two simplifying assumptions. First, it assumes that phonemes can be automatically recognized from speech with high accuracy. However, even in native (L1) speech, it is difficult to get the Phoneme Error Rate (PER) below 15% [7]. Second, this approach assumes that this is the only 'correct' way for a sentence to be pronounced, but due to phonetic variability this is not always true. For example, the word 'enough' can be pronounced by native speakers in multiple correct ways: /ih n ah f/ or /ax n ah f/ (short 'i' or 'schwa' phoneme at the beginning). These assumptions do not always hold which can result in a significant amount of false mispronunciation alarms and making students confused when it happens.

We propose a novel approach that results in fewer false mispronunciation alarms, by formalizing the intuition that we will not be able to recognize exactly what a student has pronounced or say precisely how a native speaker would pronounce it. First, the model estimates a belief over the phonemes produced by the student, intuitively representing the uncertainty in the student's pronunciation. Then, the model converts this belief into the probabilities that a native speaker would pronounce it, accounting for phonetic variability. Finally, the model makes a decision on which words were mispronounced in the sentence by processing three pieces of information: a) what the student pronounced, b) how likely a native speaker would pronounce it that way, and c) what the student was expected to pronounce.

In Section 2, we review the related work. In Section 3, we describe the proposed model. In Section 4, we present the experiments, and we conclude in Section 5.

2. RELATED WORK

In 2000, Witt et al. coined the term Goodness of Pronunciation (GoP) [3]. GoP starts by aligning the canonical phonemes with the speech signal using a forced-alignment technique. This technique aims to find the most likely mapping between phonemes and the regions of a corresponding speech signal. In the next step, GoP computes the ratio between the likelihoods of the canonical and the most likely pronounced phonemes. Finally, it detects a mispronunciation if the ratio falls below a given threshold. GoP was further extended with Deep Neural Networks (DNNs), replacing Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) techniques for acoustic modeling [4, 5]. Cheng et al. [8] improved the performance of GoP with the latent representation of speech extracted in an unsupervised way.

As opposed to GoP, we do not use forced-alignment that requires both speech and phoneme inputs. Following the work of Leung et al. [6], we use a phoneme recognizer, which recognizes phonemes from only the speech signal. The phoneme recognizer is based on a Convolutional Neural Network (CNN), a Gated Recurrent Unit (GRU), and Connectionist Temporal Classification (CTC) loss. Leung et al. report that it outperforms other forced-alignment [4] and forced-alignment-free [9] techniques on the task of detecting phoneme-level mispronunciations in L2 English. Contrary to Leung et al., who rely only on a single recognized sequence of phonemes, we obtain top N decoded sequences of phonemes, along with the phoneme-level posterior probabilities.

It is common in pronunciation assessment to employ the speech signal of a reference speaker. Xiao et al. use a pair of speech signals from a student and a native speaker to classify native and non-native speech [10]. Mauro et al. incorporate the speech of a reference speaker to detect mispronunciations at the phoneme level [11]. Wang et al. use siamese networks for modeling discrepancy between normal and distorted children's speech [12]. We take a similar approach but we do not need a database of reference speech. Instead, we train a statistical model to estimate the probability of pronouncing a sentence by a native speaker. Qian et al. propose a statistical pronunciation model as well [13]. Unlike our work, in which we create a model of 'correct' pronunciation, they build a model that generates hypotheses of mispronounced speech.

3. PROPOSED MODEL

The design consists of three subsystems: a Phoneme Recognizer (PR), a Pronunciation Model (PM), and a Pronunciation Error Detector (PED), illustrated in Figure 1. The PR recognizes phonemes spoken by a student. The PM estimates the probabilities of having been pronounced by a native speaker. Finally, the PED computes word-level mispronunciation probabilities. In Figure 2, we present detailed architectures of the PR, PM, and PED.

For example, considering the text: 'I said alone not gone' with the canonical representation of /ay - s eh d - ax l ow n - n aa t - g aa n/. Polish L2 speakers of English often mispronounce the /eh/ phoneme in the second word as /ey/. The PM would identify the /ey/ as having a low probability of being pronounced by a native speaker in the middle of the word 'said, which the PED would translate into a high probability of mispronunciation.

3.1. Phoneme Recognizer

The PR (Figure 2a) uses beam decoding [14] to estimate N hypotheses of the most likely sequences of phonemes that are recognized in the speech signal $\mathbf{o}$. A single hypothesis is denoted as $\mathbf{r_o} \sim p(\mathbf{r_o}|\mathbf{o})$. The speech signal $\mathbf{o}$ is represented by a mel-spectrogram with f frames and 80 mel-bins. Each sequence of phonemes $\mathbf{r_o}$ is accompanied by the posterior phoneme probabilities of shape: $(l_{r_o}, l_s + 1)$. l_{r_o} is the length of the sequence and l_s is the size of the phoneme set (45 phonemes including 'pause', 'end of sentence (eos)', and a 'blank' label required by the CTC-based model).

3.2. Pronunciation Model

The PM (Figure 2b) is an encoder-decoder neural network following Sutskever et al. [15]. Instead of building a text-to-text translation system between two languages, we use it for phoneme-to-phoneme conversion. The sequence of phonemes $\mathbf{r_c}$ that a native speaker was expected to pronounce is converted into the sequence of phonemes $\mathbf{r}$ they had pronounced, denoted as $\mathbf{r} \sim p(\mathbf{r}|\mathbf{r_c})$. Once trained, the PM acts as a probability mass function, computing the likelihood sequence $\boldsymbol{\pi}$ of the phonemes $\mathbf{r_o}$ pronounced by a student conditioned on the expected (canonical) phonemes $\mathbf{r_c}$. The PM is denoted in Eq. 1, which we implemented in MxNet [16] using 'sum' and 'element-wise multiply' linear-algebra operations.

$$\boldsymbol{\pi} = \sum_{\mathbf{r_o}} p(\mathbf{r_o}|\mathbf{o}) p(\mathbf{r} = \mathbf{r_o}|\mathbf{r_c}) \tag{1}$$

The model is trained on phoneme-to-phoneme speech data created automatically by passing the speech of the native speakers through the PR. By annotating the data with the PR, we can make the PM model more resistant to possible phoneme recognition inaccuracies of the PR at testing time.

3.3. Pronunciation Error Detector

The PED (Figure 2c) computes the probabilities of mispronunciations $\mathbf{e}$ at the word level, denoted as $\mathbf{e} \sim p(\mathbf{e}|\mathbf{r_o}, \boldsymbol{\pi}, \mathbf{r_c})$. The PED is conditioned on three inputs: the phonemes $\mathbf{r_o}$ recognized by the PR, the corresponding pronunciation likelihoods $\boldsymbol{\pi}$ from the PM, and the canonical phonemes $\mathbf{r_c}$. The model starts with aligning the canonical and recognized sequences of phonemes. We adopted a dynamic programming algorithm for aligning biological sequences developed by Needleman-Wunsch [17]. Then, the probability of mispronunciation for a given word is computed with Equation 2, k denotes the word index, and j is the phoneme index in the word with the lowest probability of pronunciation.

$$p(\mathbf{e}_k) = \begin{cases} 0 & \text{if aligned phonemes match,} \\ 1 - \boldsymbol{\pi}_{k,j} & \text{otherwise.} \end{cases} \tag{2}$$

We compute the probabilities of mispronunciation for N phoneme recognition hypotheses from the PR. Mispronunciation for a given word is detected if the probability of mispronunciation falls below a given threshold for all hypotheses. The hyper-parameter $N = 4$ was manually tuned on a single L2 speaker from the testing set to optimize the PED in the precision metric.

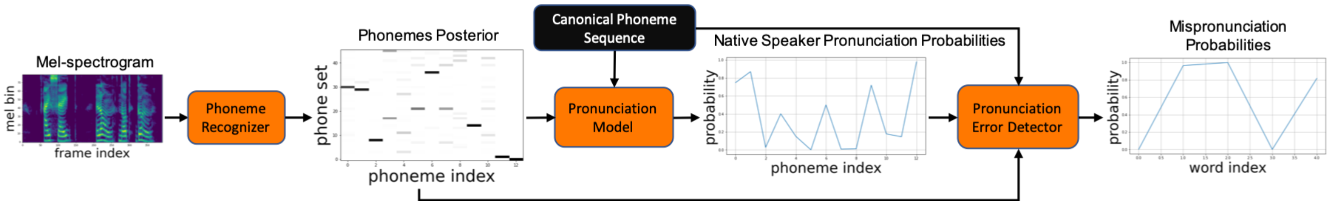

Fig. 1: Architecture of the system for detecting mispronounced words in a spoken sentence.

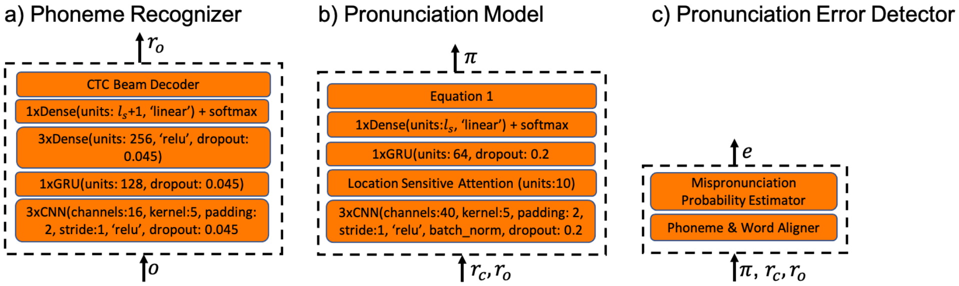
a) Phoneme Recognizer
r_o
CTC Beam Decoder
1xDense(units: l_s+1, 'linear') + softmax
3xDense(units: 256, 'relu', dropout: 0.045)
1xGRU(units: 128, dropout: 0.045)
3xCNN(channels:16, kernel:5, padding: 2, stride:1, 'relu', dropout: 0.045
o
b) Pronunciation Model
π
Equation 1
1xDense(units:l_s, 'linear') + softmax
1xGRU(units: 64, dropout: 0.2
Location Sensitive Attention (units:10)
3xCNN(channels:40, kernel:5, padding: 2, stride:1, 'relu', batch_norm, dropout: 0.2
r_c, r_o
c) Pronunciation Error Detector
e
Mispronunciation Probability Estimator
Phoneme & Word Aligner
π, r_c, r_o

Fig. 2: Architecture of the PR, PM, and PED subsystems. l_s - the size of the phoneme set.

4. EXPERIMENTS AND DISCUSSION

We want to understand the effect of accounting for uncertainty in the PR-PM system presented in Section 3. To do this, we compare it with two other variants, PR-LIK and PR-NOLIK, and analyze precision and recall metrics. The PR-LIK system helps us understand how important is it to account for the phonetic variability in the PM. To switch the PM off, we modify it so that it considers only a single way for a sentence to be pronounced correctly.

The PR-NOLIK variant corresponds to the CTC-based mispronunciation detection model proposed by Leung et al. [6]. To reflect this, we make two modifications compared to the PR-PM system. First, we switch the PM off in the same way we did it in the PR-LIK system. Second, we set the posterior probabilities of recognized phonemes in the PR to 100%, which means that the PR is always certain about the phonemes produced by a speaker. There are some slight implementation differences between Leung's model and PR-NOLIK, for example, regarding the number of units in the neural network layers. We use our configuration to make a consistent comparison with PR-PM and PR-LIK systems. One can hence consider PR-NOLIK as a fair state-of-the-art baseline [6].

4.1. Model Details

For extracting mel-spectrograms, we used a time step of 10 ms and a window size of 40 ms. The PR was trained with CTC Loss and Adam Optimizer (batch size: 32, learning rate: 0.001, gradient clipping: 5). We tuned the following hyper-parameters of the PR with Bayesian Optimization: dropout, CNN channels, GRU, and dense units. The PM was trained with the cross-entropy loss and AdaDelta optimizer (batch size: 20, learning rate: 0.01, gradient clipping: 5). The location-sensitive attention in the PM follows the work by Chorowski et al. [7]. The PR and PM models were implemented in MxNet Deep Learning framework.

4.2. Speech Corpora

For training and testing the PR and PM, we used 125.28 hours of L1 and L2 English speech from 983 speakers segmented into 102812 sentences, sourced from multiple speech corpora: TIMIT [18], LibriTTS [19], Isle [20] and GUT Isle [21]. We summarize it in Table 1. All speech data were downsampled to 16 kHz. Both L1 and L2 speech were phonetically transcribed using Amazon proprietary grapheme-to-phoneme model and used by the PR. Automatic transcriptions of L2 speech do not capture pronunciation errors, but we found it is still worth including automatically transcribed L2 speech in the PR. L2 corpora were also annotated by 5 native speakers of American English for word-level pronunciation errors. There are 3624 mispronounced words out of 13191 in the Isle Corpus and 1046 mispronounced words out of 5064 in the GUT Isle Corpus.

From the collected speech, we held out 28 L2 speakers and used them only to assess the performance of the systems in the mispronunciation detection task. It includes 11 Italian and 11 German speakers from the Isle corpus [20], and 6 Polish speakers from the GUT Isle corpus [21].

4.3. Experimental Results

The PR-NOLIK detects mispronounced words based on the difference between the canonical and recognized phonemes.

Table 1: The summary of speech corpora used by the PR.

Native Language	Hours	Speakers
English	90.47	640
Unknown	19.91	285
German and Italian	13.41	46
Polish	1.49	12

Therefore, this system does not offer any flexibility in optimizing the model for higher precision.

The PR-LIK system incorporates posterior probabilities of recognized phonemes. It means that we can tune this system towards higher precision, as illustrated in Figure 3. Accounting for uncertainty in the PR helps when there is more than one likely sequence of phonemes that could have been uttered by a user, and the PR model is uncertain which one it is. For example, the PR reports two likely pronunciations for the text 'I said' /ay s eh d/. The first one, /s eh d/ with /ay/ phoneme missing at the beginning and the alternative one /ay s eh d/ with the /ay/ phoneme present. If the PR considered only the mostly likely sequence of phonemes, like PR-NOLIK does, it would incorrectly raise a pronunciation error. In the second example, a student read the text 'six' /s ih k s/ mispronouncing the first phoneme /s/ as /t/. The likelihood of the recognized phoneme is only 34%. It suggests that the PR model is quite uncertain on what phoneme was pronounced. However, sometimes even in such cases, we can be confident that the word was mispronounced. It is because the PM computes the probability of pronunciation based on the posterior probability from the PR model. In this particular case, other phoneme candidates that account for the remaining 66% of uncertainty are also unlikely to be pronounced by a native speaker. The PM can take it into account and correctly detect a mispronunciation.

However, we found that the effect of accounting for uncertainty in the PR is quite limited. Compared to the PR-NOLIK system, the PR-LIK raises precision on the GUT Isle corpus only by 6% (55% divided by 52%), at the cost of dropping recall by about 23%. We can observe a much stronger effect when we account for uncertainty in the PM model. Compared to the PR-LIK system, the PR-PM system further increases precision between 11% and 18%, depending on the decrease in recall between 20% to 40%. One example where the PM helps is illustrated by the word 'enough' that can be pronounced in two similar ways: /ih n ah f/ or /ax n ah f/ (short 'i' or 'schwa' phoneme at the beginning.) The PM can account for phonetic variability and recognize both versions as pronounced correctly. Another example is word linking [22]. Native speakers tend to merge phonemes of neighboring words. For example, in the text 'her arrange' /hh er - er ey n jh/, two neighboring phonemes /er/ can be pronounced as a single phoneme: /hh er ey n jh/. The PM model can correctly recognize multiple variations of such pronunciations.

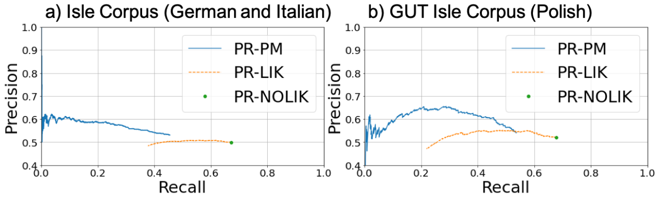

Fig. 3: Precision-recall curves for the evaluated systems.

Complementary to precision-recall curve showed in Figure 3, we present in Table 2 one configuration of the precision and recall scores for the PR-LIK and PR-PM systems. This configuration is selected in such a way that: a) recall for both systems is close to the same value, b) to illustrate that the PR-PM model has a much bigger potential of increasing precision than the PR-LIK system. A similar conclusion can be made by inspecting multiple different precision and recall configurations in the precision and recall plots for both Isle and GUT Isle corpora.

Table 2: Precision and recall of detecting word-level mispronunciations. CI - Confidence Interval.

Model	Precision [%,95%CI]	Recall [%,95%CI]
	Isle corpus (German and Italian)	
PR-LIK	49.39 (47.59-51.19)	40.20 (38.62-41.81)
PR-PM	54.20 (52.32-56.08)	40.20 (38.62-41.81)
	GUT Isle corpus (Polish)	
PR-LIK	54.91 (50.53-59.24)	40.29 (36.66-44.02)
PR-PM	61.21 (56.63-65.65)	40.15 (36.51-43.87)

5. CONCLUSION AND FUTURE WORK

To report fewer false pronunciation alarms, it is important to move away from the two simplifying assumptions that are usually made by common methods for pronunciation assessment: a) phonemes can be recognized with high accuracy, b) a sentence can be read in a single correct way. We acknowledged that these assumptions do not always hold. Instead, we designed a model that: a) accounts for the uncertainty in phoneme recognition and b) accounts for multiple ways a sentence can be pronounced correctly due to phonetic variability. We found that to optimize precision, it is more important to account for the phonetic variability of speech than accounting for uncertainty in phoneme recognition. We showed that the proposed model can raise the precision of detecting mispronounced words by up to 18% compared to the common methods.

In the future, we plan to adapt the PM model to correctly pronounced L2 speech to account for phonetic variability of non-native speakers. We plan to combine the PR, PM, and PED modules and train the model jointly to eliminate accumulation of statistical errors coming from disjoint training of the system.

6. REFERENCES

[1] A. Neri, O. Mich, M. Gerosa, and D. Giuliani, "The effectiveness of computer assisted pronunciation training for foreign language learning by children," *Computer Assisted Language Learning*, vol. 21, no. 5, pp. 393–408, 2008.

[2] C. Tejedor-García, D. Escudero, E. Cámara-Arenas, C. González-Ferreras, and V. Cardeñoso-Payo, "Assessing pronunciation improvement in students of english using a controlled computer-assisted pronunciation tool," *IEEE Transactions on Learning Technologies*, 2020.

[3] 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.

[4] 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.

[5] 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.

[6] W. 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.

[7] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[8] S. Cheng et al., "Asr-free pronunciation assessment," *arXiv preprint arXiv:2005.11902*, 2020.

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

[10] Y. Xiao and W. Soong, F. K .and Hu, "Paired phone-posteriors approach to esl pronunciation quality assessment," in *bdl*, vol. 1, p. 3. 2018.

[11] M. Nicolao, A. V. Beeston, and T. Hain, "Automatic assessment of english learner pronunciation using discriminative classifiers," in *2015 IEEE Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5351–5355.

[12] J. Wang, Y. Qin, Z. Peng, and T. Lee, "Child speech disorder detection with siamese recurrent network using speech attribute features.," in *INTERSPEECH*, 2019, pp. 3885–3889.

[13] X. Qian, H. Meng, and F. Soong, "Capturing l2 segmental mispronunciations with joint-sequence models in computer-aided pronunciation training (capt)," in *2010 7th Intl. Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 84–88.

[14] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE Intl. conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[15] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[16] T. et al. Chen, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv preprint arXiv:1512.01274*, 2015.

[17] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and David S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *STIN*, vol. 93, pp. 27403, 1993.

[19] H. Zen et al., "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Proc. Interspeech 2019*, 2019, pp. 1526–1530.

[20] E. S. Atwell, P. A. Howarth, and D. C. Souter, "The isle corpus: Italian and german spoken learner's english," *ICAME Journal: Intl. Computer Archive of Modern and Medieval English Journal*, vol. 27, pp. 5–18, 2003.

[21] D. Weber, S. Zaporowski, and D. Korzekwa, "Constructing a dataset of speech recordings with lombard effect," in *24th IEEE SPA*, 2020.

[22] A. E. Hieke, "Linking as a marker of fluent speech," *Language and Speech*, vol. 27, no. 4, pp. 343–354, 1984.