2R-07

# Neural Network Approach to English Pronunciation Evaluation through Error Detection using Phonetic Alignment

Jovan Dalhouse[*], Katunobu Itou[†]

## 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 independent of native assistance. Research aimed at improving pronunciation quality through Automatic Speech Recognition (ASR) techniques has grown increasingly popular in recent years. Attention Mechanisms have been shown to improve the precision of alignment in sequence-to-sequence tasks [8]. This is done by encoding an entire input sequence into a sequence of context vectors using trainable attention weights. This was developed as an improvement to the traditional encoder-decoder method where input sequences are reduced to a fixed length vector, which would lead to a degradation in performance with longer sequences. The use of Attention Mechanisms have been applied to phonetic alignment in [9] using both Spectral and Phonetic side context, noting great improvements in alignment as well as slight improvements in Phone Error Rate when combined with CTC. Such attention based training, in combination with CTC loss should lead to improvements in the classification of non-native mispronunciation using a strictly probabalistic approch, in addition to alignment, using RNNs.

## 2 Related Research

In the past few years, there have been good contributions to English-based Computer Assisted Pronunciation Training (CAPT) systems, including those aimed specifically for Japanese native speakers. In one notable example, a CAPT system was developed to identify mispronounced phonemes in English speech utterances by Japanese Native Speakers (L2 Learners) using Native and Non-Native (Japanese) English acoustic models [1]. In this research, similar to [3], error patterns were outlined based on common errors uttered by Japanese Speakers in English Pronunciation for optimal performance. The key error patterns in these research are as follows:

- Phoneme skipping
  (Ex. /k//aa/ instead of /k//aa//r/)
- Pause insertion (reflection of Japanese sokuon)
  (Ex. /k//i/-pause-/t//o/)
- Phoneme Insertion
  (Ex. an additional /o/ in /s//ch//r//i://t//o/)
- Phoneme Substitution
  (Ex. /b//e//r//i/ instead of /v//e//r//i/)

The problem with this phonetic graph is that it is narrow and completely disregards certain phonetic groups such as vowel-based errors which are very prominent in English Speech by Japanese speakers and crucial to English pronunciation improvement. For this research, in addition to the aforementioned patterns, select vowel-based mispronunciation errors will also be tackled. An anticipated problem with such a statistical approach is the classification of certain vowels which are either solely or primarily discerned through duration (such as /i/ and /i:/). For this reason, in addition to the current model architecture, a Normalized Rate of Speech Duration score will be calculated to estimate irregularities in phone duration relative to the duration of the entire utterance.

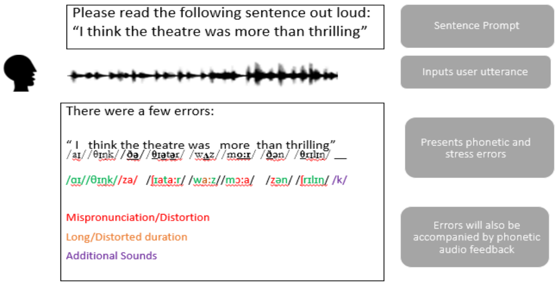
Please read the following sentence out loud:
"I think the theatre was more than thrilling"
Sentence Prompt
Inputs user utterance
There were a few errors:
" I think the theatre was more than thrilling"
/aɪ/ /θɪŋk/ /ðə/ /θɪətər/ /wʌz/ /moːr/ /ðən/ /θrɪlɪŋ/ __
/ɑɪ//θɪŋk//za/ /ʃɪata:r/ /wa:z//mɔ:a/ /zən/ /ʃrɪlɪn/ /k/
Presents phonetic and stress errors
Mispronunciation/Distortion
Long/Distorted duration
Additional Sounds
Errors will also be accompanied by phonetic audio feedback

**Figure 1.** Visual of the expected pronunciation evaluation system interface.

## 3 Model Architecture

### 3.1 LSTM

Long Short-Time Memory RNNs (LSTM) were used for this approach due to it high performance in time-sequence data classification as shown in previous speech recognition tasks such as [2, 4, 5], which makes it a suitable base for this systems model architecture. The current model architecture is built similar to that in [6] with 5 bidirectional LSTM layers, 4 of which are equipped each with a drop-out rate of 0.5 and Rectified Linear Unit (ReLu) Activation. Convolutional Layers were also added preceding these layers.

### 3.2 Connectionist-Temporal Classification (CTC)

Most traditional RNNs in Speech Recognition utilize Cross-Entropy Loss functions to train weights by mapping features to pre-segmented data. This allows for relatively accurate alignment as seen in [7], which would suggest that such a method would be suitable for this specific task. However, this would require data with accurate time-aligned labeling which in most cases is unavailable and time consuming to pre-process (see Section 4.2). For this reason, the CTC loss function was employed. CTC allows for an alignment to be modeled without a fixed input and a pre-segmented label sequence of the same length. While CTC loss is proven to model alignment with unsegmented labels and reduce Phoneme/Word Error Rate, one major drawback is the increase in time-alignment imprecision [7].

### 3.3 Attention Mechanism (Basic Concept)

This research will take on a similar approach to [9] measuring the extent to which additional phonetic

[*] Hosei University Graduate School of Computer Information Science

[†] Hosei University

context can be used to improve phonetic alignment, while also improving mispronunciation (particularly substitution error) classification accuracy. Unaligned phonetic sequences are initially encoded into 'activation matrices'.

$$W = \sigma\left(\xi_s(w\xi_c)^T\right) = \frac{\exp\left(\xi_s(w\xi_c)^T\right)}{\sum_{i=1}^{n} f(x_i)\exp\left(\xi_s(w\xi_c)^T\right)}$$

$$C = W\xi_c$$

This encoded spectral input vector $\xi_s$ and encoded phonetic side information $\xi_c$ representations are both used to estimate the attention weight $W$ and finally converted into context vector $C$.

## 4 Corpora

### 4.1 Darpa TIMIT Speech Corpus

Native data from the Darpa TIMIT Acoustic-Phonetic Speech Corpus contains phonetically labeled speech utterances from 630 native English speakers from 8 different regions of the United States. This is one of the largest phonetically time-aligned speech corpora available and is ideal for native model training [1, 4].

### 4.2 UME-ERJ Speech Corpus

For Non-Native English samples the UME-ERJ Speech Corpus, which is provided by the NII-SRC committee, was used. This English Speech corpus contains 460 phonetically balanced sentences including 32 sentences containing phonetic sequences challenging for Japanese Native speakers, as well as 100 sentences specifically designed to be utilized as a model test set. The utterances in this corpus are not phonetically labeled, therefore labeling had to be applied. To automate this process, the Penn's Forced Aligner.

## 5 Results

The baseline system to be used for the final comparison of the end system has been evaluated. This model was trained on Mel Frequency Cepstrum Coeffecients of both Natvie and Non-native english samples from the UME-ERJ corpus, however as the priority of this experiment was to determine its performance in Non-native phonetic classification, Native utterances were excluded from the test set during evaluation of the model (See Table 1). The model currently achieves an overall Phone Eror Rate of 0.24, however as anticipated a reletively high misclassification rate between phoneme pairs with inherent disaprity in duration were noticed (Ex. /eh/ and /ey/, or /i/ and /iy/ etc). In addition to this, there was also a low detection rate of mispronunciations involving post-consonantal vowel insertion (Ex. at the end of words such as "street" or "keep"). While a lack of training samples and a low occurence of such errors may contribute to this, another possible contributing factor is the lack of prominence in such utterances. This is seen in [10] where occurences of brief forms of epenthesis, both voiced and unvoiced, have been shown to occur in English speech by Japanese Speakers, similar to the occurence of devoiced vowels in the Japanese language. These occurence can often be relatively short (less than 35ms) and as a results can be difficult for detecton with limited samples. To circumvent this issue, such errors will be treated as "long-duration" consonants as opposed to "vowel" insertions. As a result, such cases will be evaluated based on a thresh-holded normalised rate-of-speech.

**Table 1.** Data used to train the baseline RNN Mispronunciation model. JPN and NAT refers to English speech by Japanese Speakers, and English Speech by American Native Speakers respectively

| Dataset | Training# | Testing# | Total Samples |
|---|---|---|---|
| JPN | 15,434 | 7,784 | 23,218 |
| NAT | 5,423 | - | 5,423 |
| Full Corpus | 20,857 | 7,784 | 28641 |

## 6 Conclusion

The Implementation of the Attention Mechanism for Attention Weight calculation is currently in progress. From this point, the improvement to the baseline will be evaluated and also, experiments will be conducted on how modifying the current attention weight can further improve the models ability to discern certain error pairs.

## References

[1] A. Ito, et al., Automatic detection of English mispronunciation using speaker adaptation and automatic assessment of English intonation and rhythm, Educational technology research, 29(1-2), 13-23, 2006.

[2] W. Li, et al., Improving Mispronunciation Detection for Non-Native Learners with Multisource Information and LSTM-Based Deep Models, In Interspeech (pp. 2759-2763), 2017.

[3] L. Neumeyer, et al., Automatic scoring of pronunciation quality, Speech communication, 30(2-3), 83-93, 2000.

[4] Y. Zhang, et al., A speech recognition acoustic model based on LSTM-CTC, In 2018 IEEE 18th International Conference on Communication Technology (ICCT) (pp. 1052-1055), IEEE, 2018.

[5] L. Yang, et al., Pronunciation Erroneous Tendency Detection with Language Adversarial Represent Learning, In INTERSPEECH (pp. 3042-3046), 2020.

[6] A. Hannun, et al., Deep speech: Scaling up end-to-end speech recognition, arXiv preprint arXiv:1412.5567, 2014.

[7] A. Graves, et al., Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, In Proceedings of the 23rd international conference on Machine learning (pp. 369-376), 2006.

[8] D. Bahdanau, et al., Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473, 2014.

[9] Y. Teytaut, et al., Phoneme-to-Audio Alignment with Recurrent Neural Networks for Speaking and Singing Voice, Proc. Interspeech 2021, 61-65, 2021.

[10] N. Sounders, Morphophonemic variation in clusters in Japanese English. Language Learning, 37(2), 247-272, 1987.