Deep Voice Coach
Project Proposal

Caedon Hsieh, Abhinav Reddy

Northwestern University, Evanston, IL, USA

abhinavreddy2022@u.northwestern.edu, caedonhsieh2022@u.northwestern.edu

Abstract

Voice actors and non-native English speakers may seek voice coaching to learn to speak with a particular accent. However, existing accent learning platforms fail to provide real-time feedback, while personal vocal coaches are expensive and less accessible. Deep Voice Coach (DVC) aims to use deep learning to assist real-time, computer-assisted pronunciation training (CAPT). Current CAPT approaches often requires some prior knowledge about the speech. DVC consists of two deep learning parts, each including a convolutional neural network. The first part is an attention-based accent classifier that distinguishes the user spoken accent from the target accent. The second part is a phoneme classifier that trains on both target and mixed accent data and uses classification confidence to identify phoneme-level accent discrepancies. We trained and tested models with audio speech recordings with different accents, including our target accent: American English. Our results demonstrate some degree of successful learning. The accent classifier's overall confidence drops when switching from American English to not American English, and the phoneme classifier trained on only the target accent had a larger confidence gap with target and mixed accent test data compared to the model trained on mixed accents. However, the results are not conclusive enough for practical usability. Future work may include exploring different data or more data, attempting to achieve more stable results.

1. Introduction

We built a two-part deep learning system to provide feedback that helps users learn a particular accent. For this project, we focus on learning an American English accent. The two parts include accent classification and phoneme classification. The accent classifier attempts to classify the speech as a boolean American English or not by taking in an audio file prepossessed into a melspectogram and giving confidence scores for sequential audio chunks. The phoneme classifier takes in a similarly preprocessed speech audio file and attempts to classify each phoneme, giving a confidence score where a lower confidence score indicates a less American English accent on that particular phoneme.

In practice, DVC could assist in accent-learning for actors and actresses and help new or non-native language learners improve their American English accent and pronunciation. Existing accent learning systems are either impractical or insufficient. Many platforms only provide small snippets of different accents. Few tools provide recording and playback functionality, and none of them give real-time feedback on recordings. Personal voice and accent coaches are expensive and not easily accessible for most language learners. Ideally, DVC will be an accessible accent learning system that provides real-time feedback for users.

2. Prior Work

For accent classification, there is previous research with accent and language classifiers, which is similar to the problem we aim to solve. We conducted research onto a previously made classification model that uses convolutional neural networks to classify the language that a speaker is speaking [1]. Additionally, there has been research into accent classification for different regional accents of Mandarin that used bi-directional LSTMs and i-vectors [2]. This is similar to our problem, but it is limited to around 15 different accents while our problem is done in a one vs. all fashion for American English. The phoneme classification problem is similar to automatic speech recognition (ASR) problems, which have been studied in the past. ASR aims to convert speech to words, but an intermediate step of popular ASR systems such as DeepSpeech 2 is to identify phonemes and use them to generate the most likely words [3]. Research in computer-assisted pronunciation training (CAPT) considers both prosodic and phonemic pronunciation mistakes [4]. Our system primarily focuses on phonemic errors. Furthermore, there has been recent research in ASR-free CAPT approaches [5], unlike our ASR-based phoneme classifier.

3. Dataset

For our initial training, we used the Speech Accent Archive [6] provided here: https://www.kaggle.com/rtatman/speech-accent-archive. The dataset has 1GB MP3 files with 214 accents from 177 countries. All speakers speak the same sentences which are approximately 20-40 seconds long. These MP3 files are labeled through their filenames which contains the accent name and the corresponding accent number. The data has an associated CSV that gives information such as speaker id, origin, and location.

The accent classifier and phoneme classifier experiments used the Mozilla Common Voice Dataset [7], found here: https://commonvoice.mozilla.org/en/datasets. This dataset includes 1,400 hours of validated English speech and transcripts by 60,000+ speakers, with accent labels for 700+ hours. We will only use labeled data, and we simplify the classification to binary labels of American English and not American English. The data is structured in a CSV file that includes the relative path to the audio file, transcript, age, gender, and accent. The audio files are provided in the MP3 format. We will preprocess this data using the Montreal Forced Aligner (MFA) to extract ground truth phonemes with start and end time.

4. Model

4.1. Accent Classifier Architecture

The accent classifier is a feed-forward convolutional neural network that contains an attention module in between an encoder

and decoder. The encoder contains 4 convolutional blocks where each block has a 1D convolutional layer followed by batch normalization, ReLU, and dropout. The decoder is another 4 convolutional blocks with the same structure mentioned before and followed by two linear layers. These final linear layers will output the probability that the audio is American English accent using the melspectogram. We also tested an architecture for training where the attention module was placed after 8 convolutional blocks when training initially with the Speech Accent Archive.

4.2. Phoneme Classifier Architecture

The model architecture is based on the neural network used by DeepSpeech 2 [3], consisting of two one-dimensional convolutional layers, each followed by batch normalization and ReLU activation. After those layers, the model has three bidirectional GRU layers, then a fully-connected layer. All layers use 512 channels. The output of the network is a 71-element vector where each element corresponds to one of 71 possible phonemes, which includes silence and noise. We use a cross-entropy loss function between the predicted phoneme and ground truth phonemes to train the network. The cross-entropy loss is weighted to decrease the value of predicted silences, because silences are far more common than any other phoneme. The weights discourage the network from learning to always predict silence.

5. Training Results

5.1. Speech Accent Archive Training

Our initial experiments with the Speech Accent Archive were based around finding the best chunk sized for training. We tested if using attention or no attention would provide better results for training seen in Figure 2. The data for these experiments can be seen below. As stated before, we tested training on 1, 2, and 4 second chunks and their results can be seen in Figure 3. We concluded that the two second chunk size had the best performance in terms of trade-off as the 4 second chunk size took much longer for training. Additionally, the results of using the attention module were slightly better than without. The precision, accuracy, recall, and loss for our final training with a chunk size of 2 seconds and the attention module trained on the Speech Accent Archive can be seen in Figure 1. The target accent (American English) was 80 percent of the total dataset.

5.2. Mozilla Common Voice (Accent Classifier)

We initially had a training run for the Mozilla Common Voice dataset with max epoch of 50. This run used both attention and 2-second chunk size and validation occurred at the end of every training epoch. This training and validation accuracy can be seen in Figure 4 and the precision and recall in Figure 5. After noticing that the model converged at around 15 epochs, we ran another run with max epoch at 20 to confirm this. The target accent, the American English accent, was 55 percent of this dataset. However, through our evaluation we noticed the original checkpoint for the training run with 50 epochs showed more potential when looking at the probabilities given to different accents.

As part of our experiment, we also conducted additional to test our hypothesis if the model would work better with choosing a non-American accent as our target accent due to the greater diversity within American English accent.

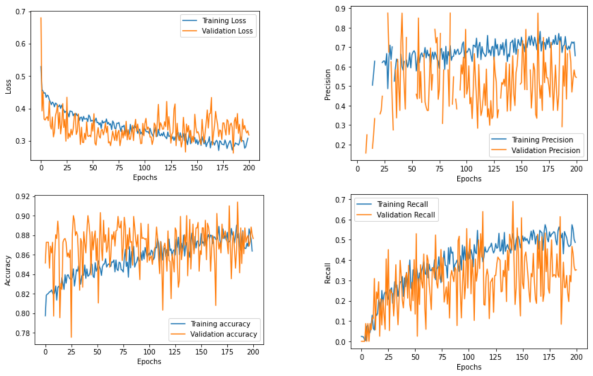

Figure 1: *Loss (top left), Precision (top right), Accuracy (bottom left), and Recall (bottom right) for the final training run using the Speech Accent Archive with a chunk size of 2. This training run had a max epoch of 200. The target accent (American English) was 80 percent of the total dataset.*

Model	Accuracy	Loss
Attention	.903	.259
Without Attention	.864	.283

Figure 2: *This table shows the best loss during training for a models with and without Attention. It also shows the corresponding accuracy at that loss*

Chunk Size used for Training	Accuracy	Loss
1 second	.867	.363
2 second	.906	.283
4 second	.917	.236

Figure 3: *This table shows the best loss during training for a models that were trained on 1, 2, and 4 second chunks of audio. It also shows the corresponding accuracy at that loss*

5.3. Mozilla Common Voice (Phoneme Classifier)

We trained two phoneme classification models, using 50 hours of training data and 10 hours of validation data from the Mozilla Common Voice Dataset. One model was trained with target accent training data, and the other was trained with mixed accent training data. We trained for 100 epochs with an Adam optimizer with a learning rate of 10^{-3}.

We also used a gradient clipping value of 0.5 due to a significant stabilization in the training loss and accuracy. Figure 6 shows the graphs for training with and without gradient clipping using the target accent data, as well as the mixed accent condition with gradient clipping. When we used gradient clipping, the learning appears to start very slowly. During this time, the network almost always predicts a small handful of phonemes, especially silence. We hypothesize that the weights are slowly growing during this time until the much faster learning begins. For both models with gradient clipping, the validation accuracy and loss becomes very unstable at this point. This may be due to having a validation dataset that is too small, and future training with a larger validation set is planned. In general, the model trained on only the target accent is able to achieve a higher best accuracy.

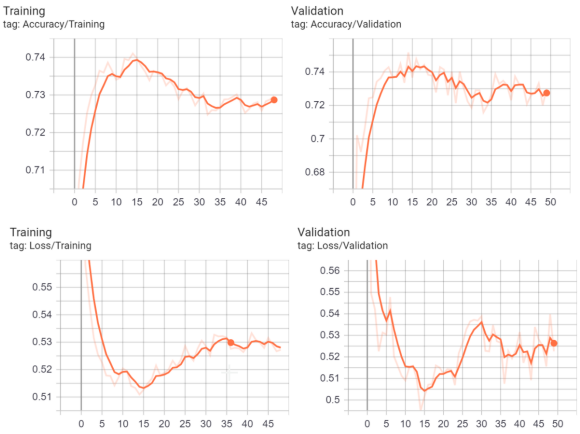

Figure 4: *These four graphs show the accuracy and loss for the max epoch of 50 run using the Mozilla Common Voice Dataset. The top left is the accuracy over epochs for training, top right is accuracy over epochs for validation, bottom left is loss over epochs for training, and bottom right is loss over epochs for validation. 55 percent of the dataset was the target accent.*

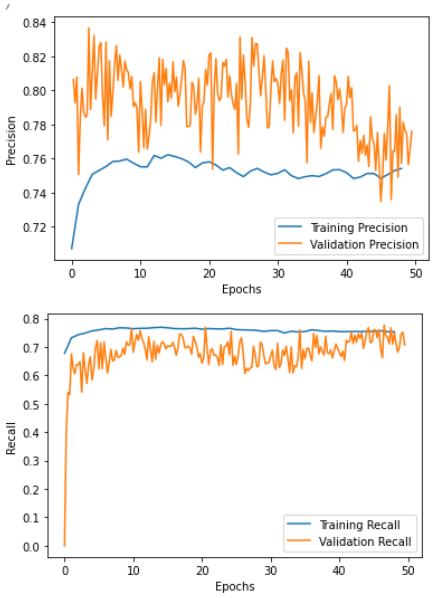

Figure 5: *These two graphs show the precision and recall or the max epoch of 50 run using the Mozilla Common Voice Dataset. The top graph is for precision for both training and validation, and the bottom graph is for recall for both training and validation. The minimum loss occurred at 15. This checkpoint was used for later evaluation. 55 percent of the dataset was the target accent.*

6. Evaluation and Results

6.1. Accent Classifier

We conducted some initial analysis with the accent classifier model after training by inputting various audio files and looking at the probability and the attention from the model. For the evaluation, we took two second chunks of each audio file with a hop size of .5 seconds that was preprocessing into a log-melspectogram. We passed these chunks into the model where it gave a probability that each chunk was American English or Non-English. The chunks were passed in sequential order for the audio file and the graph generated is the probability against the time where that given audio started.

In Figure 7, we can see a graph of probabilities (using the process above) that was generated on an audio file containing an American English accent. We also generated probability graphs for examples of Non-English, non-American English Accent, and ones that switched between American to Non-American English accents or American English to another language. After analyzing these graphs, we concluded the probabilities generally lowered when speaking non-American English. We found that due to the number of examples we analyzed that the model was able to distinguished between the accents to an extent from the average probability change. Additionally, by looking at the validation graphs seen in figure Figure 4 and 5, the precision and recall highlighted that it was able to classify many samples accurately to a statistically significant margin.

6.2. Phoneme Classifier Testing

To evaluate the model, we tested both the target-trained and mixed-trained models against both the target accent and mixed accent testing data. We trained using the saved model weights at the point with the lowest validation loss from training. Quantitatively, we look at the accuracy and loss of each model on the target and mixed testing data. A larger performance gap with the target and mixed accent testing data for the target-trained model compared to the mixed-trained model might suggest that the target-trained model has learned target-specific phoneme nuances. Qualitatively, we manually inspect audio files and analyze the model's frame-by-frame confidence as it relates to accent errors.

Figure 8 shows the average testing accuracy and loss for both models with both testing data partitions. For context, always predicting the most common phoneme (silence) would give an accuracy of 0.248. The target-trained model had a slightly larger accuracy performance gap of 3% compared to the mixed-trained model's 1.3% performance gap. In general, the target-trained model performed better than the mixed-trained model, even on the mixed accent testing data.

In Figure 9, we analyze the confusion matrix for the best test run. The confusion matrix was computed by counting each predicted-actual phoneme combination, then dividing by the total number of times the actual phoneme appears. The model is very good at predicting silence, which is phoneme 0. Additionally, there is a relatively strong diagonal, which indicates correct predictions. Finally, the slightly bright 2x2 squares along the diagonal indicate confusion between related phonemes. For example, the 2x2 square at phoneme 35 and 36 corresponds to IH0 and IH1, which are related phonemes that the model commonly confuses.

Figure 10 shows confidence and phonemes for a single example: an non-American speaker saying "what is the weather in Idaho". Confidence is computed as the max of the softmax of the neural network's output vector. Both models are confident on silence, but the confidence during speech is noisy and sharply drops between phonemes. Even though this example is not American English, we still see higher confidence with the target-trained model, so it is hard to identify particular accent errors from confidence. The highest confidence spike in the target-trained model occurs at around frame 250, which is the

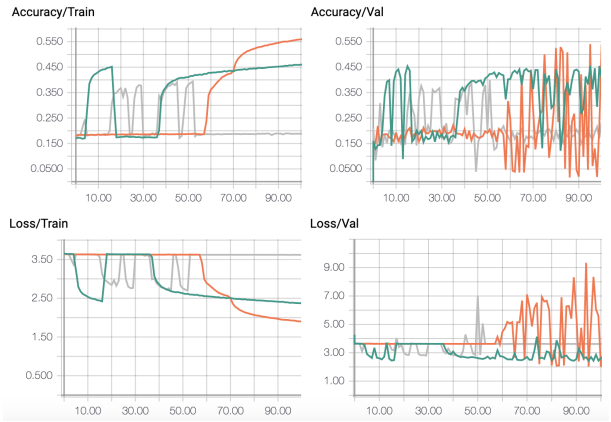

Figure 6: *Four graphs showing the training accuracy and loss for models trained using target accent with no gradient clipping (gray), target accent with gradient clipping (orange), and mixed accents with gradient clipping (green). The x axis is the epoch number.*

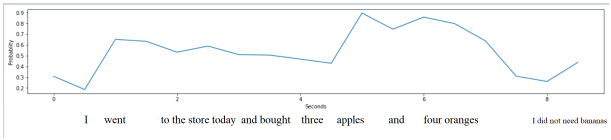

Figure 7: *This graph the show the probabilities that a given 2 second chunk with a hop size of .5 seconds is American English for an audio file that is spoken with an American English accent. The words at each time are annotated at the bottom and are approximated.*

model	accent condition	accuracy	loss
mixed-trained	mixed	0.442	2.471
mixed-trained	target	0.455	2.399
target-trained	mixed	0.504	2.230
target-trained	target	0.534	2.033

Figure 8: *This table shows the testing accuracy and loss for each combination of model and type of testing data (mixed or only target)*

"AY1" phoneme at the start of "Idaho". The speaker incorrectly pronounces an "H" sound just before, which the target-trained model classifies incorrectly.

7. Conclusion

In conclusion, Deep Voice Coach experiments showed some potential methods that could be useful for accent learning. The accent classifier was able to distinguish between American English accents and non-American English accents to an extent as seen with our manual evaluation and through the validation graphs which shows potential for the classifier to improve and distinguish. The phoneme classifier showed a small, but promising performance gap, and the model demonstrated an ability to learn 71-way phoneme classification with decent accuracy. However, the phoneme classifier does not clearly identify phoneme-level accent errors from confidence, and the training validation statistics were unstable. Overall, this DVC may contribute to accent learning research, but does not in itself provide a usable basis for accent learning with real-time, phoneme-level feedback.

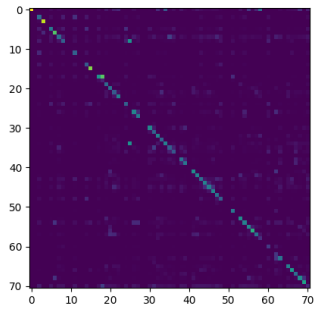

Figure 9: *The confusion matrix for the target-trained model with target accent testing data shows the relationship between actual phoneme class (horizontal) and predicted phoneme class (vertical) for each combination of model and type of testing data (mixed or only target). Each squares value represents a percentage of predicted phoneme for each actual phoneme, where yellow is a high value and purple is a low value.*

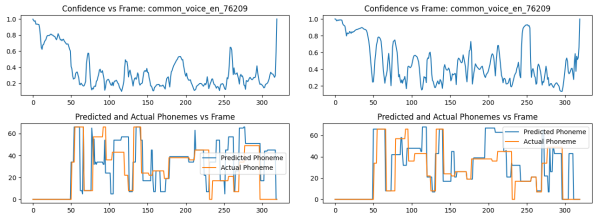

Figure 10: *Confidence, predicted phonemes (blue), and actual phonemes (orange) per frame for the mixed-trained model (bottom) and target-trained model (top).*

8. References

[1] G. Keren, J. Deng, J. Pohjalainen, and B. Schuller, "Convolutional neural networks with data augmentation for classifying speakers' native language," pp. 2393–2397, 2016. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-261

[2] J. P. D. W. P. Z. Felix Weninger1, Yang Sun, "Deep learning based mandarin accent identification for accent robust asr," pp. 510–514, 2019. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2737

[3] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," 2015.

[4] S. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," 06 2012.

[5] S. Cheng, Z. Liu, L. Li, Z. Tang, D. Wang, and T. F. Zheng, "Asr-free pronunciation assessment," 2020.

[6] S. Weinberger, "Speech accent archive," 2015.

[7] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.