Accent Classification Using a Zero-Shot Learning Approach

University of Pennsylvania
CIS 519: Applied Machine Learning

Caroline Harrison, Sally Hu, Ian MacDonald, Andrea Yoss

Abstract

This paper proposes a zero-shot learning approach to address the continued challenge of accent classification by controlling for differences attributed to a speaker's gender, such as pitch. We attempt to classify accents by nationality by extracting audio features, training on audio samples from male speakers and testing on audio samples from female speakers. Results show that our model can perform zero-shot audio classification with a relatively small dataset and achieve up to 9.35% accuracy, better than the random guessing rate of 5%.

1 Introduction

Although immense progress has been made in the development of automatic speech recognition (ASR) systems, interpreting different accents accurately still continues to pose a challenge. This challenge is underscored by the variability in voices, which is dictated by a multitude of factors that are difficult for ASR systems to sufficiently capture and interpret, of which the gender and accent of the speaker prove to be the most confounding [1]. Additionally, due to the high variability in speech, current methods used to train these systems require massive audio training corpora which is expensive and difficult to obtain [2]. Therefore, our motivation is driven by the need to come up with a method to more accurately classify accents while minimizing the number of samples required for training.

To achieve this objective, we take a zero-shot learning approach to accent classification by training on the speech of male speakers from various countries and testing on the speech of female speakers. Our aim is to develop a model that can more accurately classify accents by nationality regardless of the gender of the speaker. Our approach essentially neutralizes audio samples to remove the variability caused by the gender of speakers, which enabled us to achieve an accuracy of 9.35%, surpassing the random guessing rate of 5%.

2 Dataset and Features

2.1 Dataset

The dataset that we used came from VoxCeleb1 [3], which provided us with thousands of audio clips and the corresponding speaker names, genders and nationalities. We analyzed the dataset and decided only to keep samples from nationalities that had both male and female speakers represented within the dataset. Our final dataset consisted of 6,630 audio clips from 732 unique speakers across 20 different nationalities. Of the 732 speakers, 308 were female and 424 were male.

2.2 Preprocessing and Feature Extraction

Each audio clip was preprocessed prior to features extraction. We extracted three-second snippets from each clip and filtered out noise using Librosa [4] by taking the Fast Fourier Transform (FFT) and removing any part of the signal with a magnitude of less than twenty.

Using the filtered signal, we then extracted features of interest, again using Librosa. In general, spectrograms and Mel-frequency cepstral coefficients (MFCC) prove to be effective in capturing important aspects of speech and have been used in other audio classification works [5, 6]. MFCCs in particular take into account human perception for sensitivity at appropriate frequencies, which make these useful for the task of speech classification [1]. To add to our experiment, we also extracted chroma, spectral contrast, harmonics and percussive components. Since chromagrams profile pitch by classes, we hypothesized that these features would be particularly useful in removing the variability attributed to differences in gender.

3 Experiments

3.1 Methods

We implemented and experimented with two neural network architectures in Python using PyTorch [7]: a Multi-layer Perceptron (MLP) and a Convolutional Neural Network (CNN). In the context of classic speech classification, CNNs have proved to be quite successful [1, 8], so we believed that using neural networks was the best approach to take. The below network architectures reflect the parameters of the highest performing models.

The first neural network, MLP, consists of three fully connected dense layers. We use rectified linear units (ReLU) as activation functions in all layers, including the last one and, to prevent overfitting, we include two dropout layers with a dropout rate of 0.3. The network is trained using adaptive moment estimation, Adam, and uses cross-entropy as the loss function.

The second neural network, CNN, consisted of two convolutional layers with a max pooling layer after each, followed by three dense layers. Each layer used ReLU as the activation function. A single dropout layer with a dropout rate of 0.3 was also added in order to account for the possibility of overfitting. The network was trained in batch sizes of 32, using Stochastic Gradient Descent (SGD) as the optimization algorithm and cross-entropy as the loss function.

3.2 Results

After tuning hyperparameters and experimenting with different features, we can see that our models performed better than random guessing. Our final results from training on male speakers and testing on female speakers are displayed in Table 1.

Model	Test Accuracy %
MLP	9.35%
CNN	7.36%

Table 1: Models and Resulting Test Accuracies

3.2.1 Network Analysis: MLP

Although both models performed better than random guessing, it is interesting to see that the MLP performed better than the CNN on the test data.

After experimenting with each feature individually as well as combinations of features, we concluded that the most important combination of features that consistently elicited the highest test accuracies included the MFCC, chromagram and spectral contrast.

Using these features, we noted that the MLP achieved its best performance after 100 epochs. Due to the zero-shot learning aspect of this experiment, increasing the number of epochs above 100 severely impacted the final test accuracy negatively since the model overfit to samples that it would ultimately not see in testing. Consequently, although the training accuracy increased to over 90% with more epochs, we focused instead on improving the test accuracy. Doing so yielded a test accuracy of 9.35%.

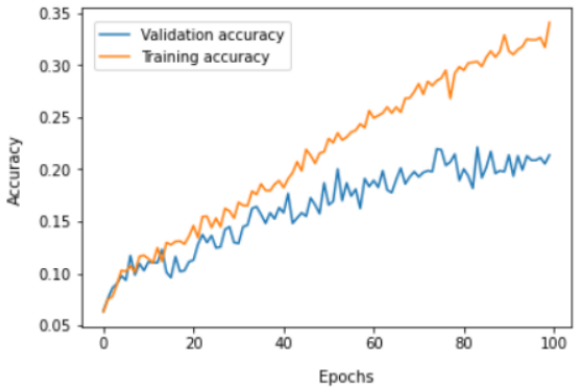

Figure 1: MLP - Training and validation accuracy

3.2.2 Network Analysis: CNN

For our CNN, we experimented with a range of different parameters as well as different input features. We implemented a CNN that took the spectrogram, chromagram, and MFCC features as inputs, and compared the accuracy over different variations. Adjustments made to the CNN included adding and removing dropout layers, trying different convolutions (1D vs 2D), changing the number of convolutional layers, pooling layers and dense layers, and changing the optimization algorithm used (SGD vs. Adam). Within each of the layers, we tested various different parameters to find the optimal input/output sizes for each layer, as well as parameters such as kernel size, strides, and padding. Beyond that, we also tried to find the optimal number of training epochs that would allow our CNN to learn from the training set without overfitting, especially given the zero-shot aspect of the project.

The final model we implemented is displayed in Figure 2 and the resulting training and validation

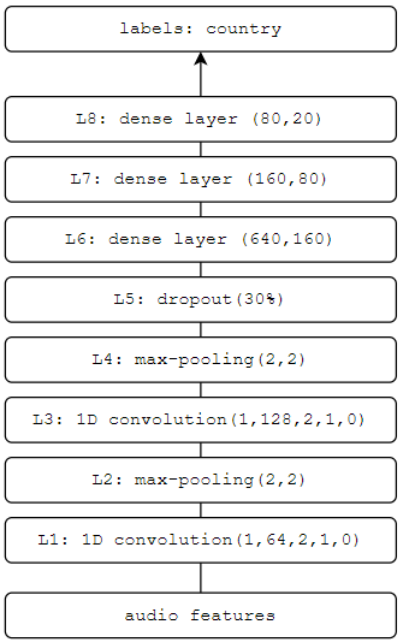

Figure 2: CNN Architecture

accuracies are displayed in Figure 3.

The field of audio recognition, especially as it pertains to zero-shot learning, is still relatively novel compared to a field such as image recognition. For image recognition, there are a number of pre-built architectures that have proven to perform extremely well in image classification competitions, such as the now famous AlexNet [9]. For audio classification in general, and specifically as it pertained to accent classification, it was extremely difficult to find any pre-built architectures. Therefore, we decided to perform extensive research on successful audio classification networks to determine what would work well for our classification problem. We found various instances of successful CNNs, specifically in the fields of speaker recognition [11], speech emotion recognition [11], and most related to our work, accent classification [1, 10]. Additionally, we found examples of 1D CNN success [11], rather than the more common 2D CNN [1,10].

After experimenting with various CNNs, we ultimately ended up with the CNN we described previously. We found that the 1D convolutions worked better with our data, which makes sense considering the sequential nature of speech [11]. We then trained the CNN on 150 epochs, again to prevent overfitting and because the validation accuracy plateaued around that point. However, we found that our model only had an accuracy of 7.36% on the testing data. The drop in accuracy from our validation set (approximately 20%) was not unexpected given the inherent challenges of using a zero-shot learning approach. Nevertheless, our model is still better than random guessing.

3.2.3 Data Analysis

After we obtained our results, we analyzed the incorrect predictions to better understand our errors. From an initial glance, we saw that 38.7% of the misclassifications involved audio samples from speakers whose nationality was recorded as Mexico. We listened to some clips from those speakers and realized that using the nationality tag from VoxCeleb1 did not necessarily imply a given speaker's accent. Some of the clips were in Spanish; some were in English but sounded American or Canadian; some were in English with what might be interpreted as a Mexican accent, but could also be misconstrued as a European accent.

We decided to run our model again without samples from Mexico and obtained a test accuracy of 10.3%. The improvement was immediate with no hyperparameter tuning, which leads us to believe that if we used an initial corpus of audio samples that was annotated by origins of accent instead of nationality, our model would perform much better. The same issue was present in samples from other nationalities as well.

4 Discussion

Given that zero-shot learning in the context of audio classification is still relatively novel, our results are significant. By controlling for a speaker's gender by pinpointing and extracting the relevant features from a signal, we saw that the task of accent classification yielded more accurate results despite inconsistencies between audio sam-

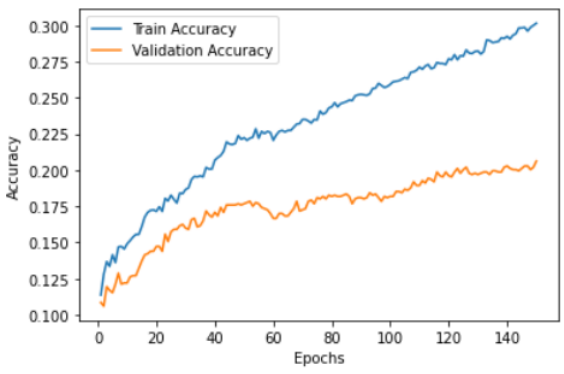

Figure 3: CNN - Training and validation accuracy

ples and labels. Had we used a dataset with samples labeled by accent origin and not speaker nationality, we believe that our accuracies would be much higher.

We also noted that CNNs are not necessarily the best models to use for audio classification. We suspect that this is because audio data cannot be treated like images in that it is not static and heavily depends on time. Compared to the CNN, the MLP not only yielded better results, but was also simpler and faster to train. Therefore, increasing the scale by adding more samples would also be computationally inexpensive.

5 Related Work

We saw very little research in the field of zero-shot audio classification for the purposes of accent and speech recognition. Work in this area primarily relates to understanding the context of the words spoken in the audio files [12, 13, 14]. However, we did find examples of success using CNNs for speaker recognition [15], speech emotion recognition [15], and most relevant for us, accent classification [1, 15]. It appears that neither of these experimented with the idea of building CNNs for zero-shot learning, but after researching, we believed that building a CNN was likely to be our best option. However, the lack of research induced us to explore other models that we felt would do well with this unique classification problem and consequently, using a CNN did not prove to be the best model in our case.

Across the different areas we researched in audio classification, we also saw a common pattern in using features extracted in the form of spectrograms and MFCCs [17]. This provided us with an initial idea of what features to extract and explore, eventually leading us to look for features that minimize the effects of frequencies spanning multiple octaves [18, 19]. Consequently, we decided to extract chromagram features as well.

6 Future Work

In this project, we used features extracted exclusively from the audio samples of male speakers from a selection of nationalities. However, to facilitate our zero-shot learning approach to accent recognition, we may also want to include additional, domain-specific features that actually exist independently of our task. While much of the related prior works have been done with respect to natural language processing, we believe there is potential for it to also be applied to audio classification tasks.

For example, a potential reason for cross-linguistic universals (i.e., similarities between different languages) is geography, as "languages that are related to one another geographically often share structures, too" [19]. Therefore, by evaluating the similarities between a country's geography (latitude/longitude of each country's centroid) and languages spoken, and then relating those to our training data, we may be able to improve the mappings between the features in our "unseen" samples to their corresponding nationalities.

Perhaps this can be achieved by combining the models in sequence. A simplified version of this process is displayed in Figure 4.

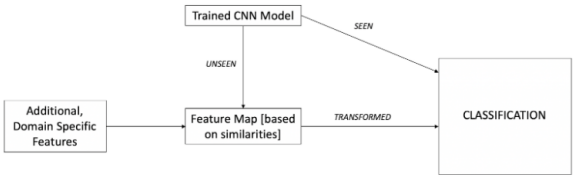

Figure 4: Proposed process to further explore

Instead of using just our trained CNN to classify all samples, we propose using some sort of outlier detection method (e.g. clustering) to decide whether or not the test data mapped onto our domain-related feature space is "seen" or "unseen." If the data is determined to be "unseen," it will then combine this unseen data that has been mapped onto our feature space with the coordinate similarities of additional, domain-specific features to help classify based on proximity to other seen or unseen labels [20]. If this proposed approach works, it would minimize the number of samples needed from every nationality, thereby lowering the cost of obtaining samples and time required to train the models.

7 Conclusion

We have shown that it is possible to more accurately classify accents with minimal samples by controlling for the gender of the speaker. Even using inaccurately labeled data, our model yielded results that are promising and are easily replicable with lower costs. Even with the application of a zero-shot learning approach, we managed to achieve an accuracy of 9.35%, but we believe that our approach and models can be improved upon

and refined by combining it with other ideas such as phonetic annotations [5] or in conjunction with our proposition for future work. Regardless, our experiment presents a solid foundation for future research in refining ASR systems to better recognize and interpret accents using a smaller corpus of audio samples.

8 Supplementary Materials

Main Submission: Google Drive
Additional Supplementary Code/Features: GitHub

References

[1] Sheng, Leon Mak An, and Mok Wei Xiong Edmund. Deep Learning Approach to Accent Classification. *Project Report, Stanford University, Stanford, CA*, 2017.

[2] Prasad, M., van Esch, D., Ritchie, S., Mortensen, J. F. Building Large Vocabulary ASR Systems for Languages Without Any Audio Training Data. *Proc. Interspeech 2019*, 271–275, 2019.

[3] A. Nagrani*, J. S. Chung*, A. Zisserman. VoxCeleb: a large-scale speaker identification dataset. INTERSPEECH, 2017.

[4] Brian McFee, Vincent Lostanlen, Alexandros Metsai, Matt McVicar, Stefan Balke, Carl Thomé, … Taewoon Kim. (2020, July 22). librosa/librosa: 0.8.0 (Version 0.8.0). Zenodo. http://doi.org/10.5281/zenodo.3955228

[5] X. Li, S. Dalmia, D. R. Mortensen, F. Metze, and A. W. Black. Zero-shot learning for speech recognition with universal phonetic model. 2019. [Online]. Available: https://openreview.net/forum?id=BkfhZnC9t7

[6] H. Xie and T. Virtanen. Zero-Shot Audio Classification Based On Class Label Embeddings. *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2019, pp. 264-267, doi: 10.1109/WASPAA.2019.8937283.

[7] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32 (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[8] Y. Lukic, C. Vogt, O. Dürr and T. Stadelmann. Speaker identification and clustering using convolutional neural networks. *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, Vietri sul Mare, 2016, pp. 1-6, doi: 10.1109/MLSP.2016.7738816.

[9] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Communications of the ACM* 60.6 (2017): 84-90.

[10] Wu, Yu, Hua Mao, and Zhang Yi. Audio classification using attention-augmented convolutional neural network. *Knowledge-Based Systems* 161 (2018): 90-100.

[11] Li, F., Liu, M., Zhao, Y. et al. Feature extraction and classification of heart sound using 1D convolutional neural networks. *EURASIP J. Adv. Signal Process.* 2019, 59 (2019). https://doi.org/10.1186/s13634-019-0651-3

[12] Lei Ba, Jimmy, Kevin Swersky, and Sanja Fidler. Predicting deep zero-shot convolutional neural networks using textual descriptions. *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[13] Mensink, Thomas, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[14] Zhang, Ziming, and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. *Proceedings of the IEEE international conference on computer vision*. 2015.

[15] Wu, Yu, Hua Mao, and Zhang Yi. Audio classification using attention-augmented convolutional neural network. *Knowledge-Based Systems* 161 (2018): 90-100.

[16] Dave, Namrata. Feature extraction methods LPC, PLP and MFCC in speech recognition. *International journal for advance research in engineering and technology* 1.6 (2013): 1-4.

[17] Eronen, Antti, and F. Tampere. Chorus detection with combined use of MFCC and chroma features and image processing filters. *Proc. of 10th International Conference on Digital Audio Effects*. 2007.

[18] Wankhammer, Alexander, Peter Sciri, and Alois Sontacchi. Chroma and MFCC based pattern recognition in audio files utilizing hidden Markov models and dynamic programming. *Proceedings of the 12th International Conference on Digital Audio Effects* (DAFx-09). 2009.

[19] University of Arizona. Why do we see similarities across languages? Human brain may be responsible. *ScienceDaily*. 1 December 2017. [Online] www.sciencedaily.com/releases/2017/12/171201135555.htm.

[20] Socher, Richard, et al. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*. 2013.