Zero-Shot Learning for Audio Classification Using Phoneme Detection

Kaylyn Clark
Genomics and Computational Biology
kaylync@pennmedicine.upenn.edu

Vera Lee
Robotics
veralee@seas.upenn.edu

Peter Galer
Bioengineering
galerp@seas.upenn.edu

Arielle Stern
Data Science
arstern@seas.upenn.edu

Abstract

The ability to identify speakers through audio is a complex challenge; emerging technologies which rely on speech processing typically suffer when audio samples contain non-Anglican accents. As such, our goal is to predict the nationality and gender of audio samples from 31 different nationalities through training only on audio samples from individuals of English-speaking nationalities, a problem which requires zero-shot learning. Phonetic attributes are used for this task. Each language has a unique set of associated phonemes, some of which overlap with those in the English language. We explore LSTM, CNN, and FF-NNs for extracting phonemes from audio samples. Predictions are based on maximal overlap in phoneme content with the unique phoneme sets for each nationality. We achieved a 92.7% test accuracy for predicting gender, and accuracies of 13.3,%, 0.97%, and 25.9% for predicting nationalities using the LSTM, CNN, and FF-NN models, respectively. Overall, our work highlights the feasibility of applying zero-shot learning approaches to audio classification tasks.

1 Introduction

The ability to parse and identify speakers through audio is a complex challenge facing emerging technologies. AI and voice-assisted devices such as Siri frequently encounter problems in processing speech from individuals with heavy, non-Anglican accents, many of whom are not native English speakers. As early as 2001, accent and gender were identified as some of the most important features contributing to the variability in performance of automatic voice recognition systems (Huang and Zhou, 2001). Similarly, text-to-speech technology often suffers a significant drop in accuracy when the speaker has a non-Anglican accent. Many speech-to-text APIs even suffer from accents such as those from the UK or Australia, as training is often focused on American samples. Consequently, these tools which have become increasingly integrated into many people's lives are less accessible to certain populations.

1.1 Relevant Work

Zero-shot learning for audio classification is not a well-studied subject, but we performed a literature review to determine key features in speech that can distinguish gender and nationality. For instance, the average female voice is within a power spectra of 175-250 Hz, while a male voice is within 75-150 Hz (Traunmüller and Eriksson, 1995). Most individuals in the dataset are speaking English regardless of nationality, so the speaker's accent is the distinguishing feature. Although the speakers' accents are slight, we can potentially identify their nationality through frequency content - when individuals speak a non-native language, they may vary their pitch range (Li, 2016). Phonemes from the speaker's native language are often still in their utterances as they speak a second language. Previous studies on classifying audio files have converted audio to spectrograms for classification (Khamparia et al., 2018), allowing for the use of image classification models. This motivated our use of spectrograms as input for our CNN model, simplifying our task from one of audio classification to image classification, which is well-studied in the literature.

Nationality	Individuals
USA	774
UK	210
Canada	52
Australia	37
India	25
Norway	19
Ireland	15
Other	79
Gender	**Individuals**
Male	665
Female	546

Table 1: Number of individuals by nationality and gender. For more details, see our provided code or VoxCeleb.

1.2 Problem Definition

We explore several machine learning techniques to identify the accent and gender of individuals with the VoxCeleb1 dataset. VoxCeleb1 is an open-source dataset of over 100,000 short audio files from 1,251 celebrities (Table 1). This data was generated by extracting videos from YouTube and carrying out speaker verification (or facial recognition) via a two-stream synchronization CNN (Nagrani et al., 2017).

1.3 Expectations

Identification via audio remains a complex problem in machine learning due to variability and noise integrated into a signal. Further, effective models typically require a significant amount of time and computing power, the latter being a frequent obstacle for us throughout our work. Finally, zero-shot learning (ZSL) presents an especially difficult case as the model must be tested on categories it has not been trained on. Consequently, we set our goal to produce a model for gender and nationality that achieved better-than-chance accuracy.

2 Approach

2.1 Phonetic Attributes for ZSL

ZSL often requires identifying attributes belonging to both the training and testing distributions for effective transfer learning (Xian et al., 2019). We used phonetic attributes of language, which classify various sounds present in speech, to bridge the gap between audio features and unseen nationalities in testing. As different languages are comprised of a unique set of phonemes, phonetic features have been shown to be useful in speech and language recognition tasks (Li et al., 2019).

To generate a mapping from phonemes to nationality, we used Phoible, a repository of phonological inventory data, which defines the phonemes present in different languages using the International Phonetic Alphabet (IPA) (Moran and McCloy, 2019). The number of phonemes represented across the nationalities present in the VoxCeleb dataset is shown in Figure 1.

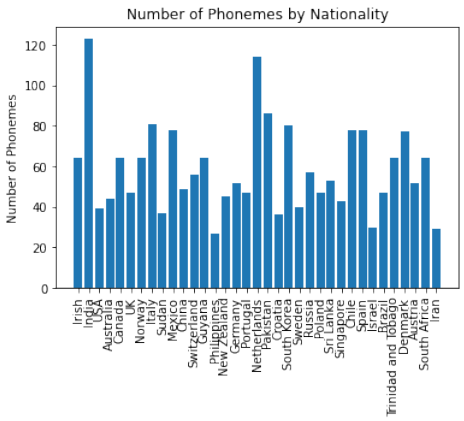


Figure 1: Number of phonemes present in VoxCeleb languages.

3 Methodology

3.1 Neural Network for Gender Prediction

To predict gender from audio, we pre-processed the data by passing each audio clip through a band-pass Butterworth filter of order 4 with cutoff frequencies 75 Hz and 5000 Hz. This process filters out most of the low-frequency noise while retaining the typical frequency range for males and females (Traunmüller and Eriksson, 1995) and any significant high-frequency features in the voice (Monson et al., 2014).

To compute features for the neural network, for each filtered audio sample we constructed the mel-spectrogram and averaged power in 128 mel frequency bins over time. The mel features serve our purpose well, as they divide the power spectra of each audio file into frequency bins evenly spaced by distance in terms of human aural perception. We used the default values of 2048 samples for each FFT window and frame shift of 512 samples - the result was a 128-element feature array for each audio clip.

The features were passed through a 4-layer feed-forward neural network (FF-NN), where the first three layers were fully connected layers with a ReLU activation function, and the last layer was a

fully connected layer with one output neuron using a sigmoid activation function. We experimentally determined the 4-layer fully connected network with 10% dropout to be best for this task. We used early stopping with a patience of 5 epochs to prevent overfitting of the model. The model was configured with a binary cross-entropy loss and Adam optimizer. Overall, the model had 74,241 trainable parameters, which were tuned using a validation set containing 30% of all English-speaking nationality audio clips.

3.2 Phoneme Extraction

To train a model to identify phonemes, we transformed raw audio files from English-speaking nationalities to text using Google's Speech Recognition and then translated the associated text to phonemes with English to IPA, which uses Carnegie-Mellon University's Pronouncing Dictionary to convert English text into IPA. Using data from Phoible, we defined a set of phonemes that appear across English-speaking countries. We added an unseen tag to account for phonemes that would not be seen during training, accounting for all phonemes present in non-English-speaking nationality languages. Therefore, our phoneme set, which contains 92 phoneme tags, is defined as:

$$A_{phoneme} = A_{EnglishPhonemes} \cap A_{unseen}$$

For non-English-speaking nationalities, we removed phonemes that do not appear in $A_{EnglishPhonemes}$ and added A_{unseen} to the set of phonemes associated with the language.

3.3 Predicting Nationality via Phonemes

All methods explored for using phonemes to predict nationality follow a similar architecture as shown in Figure 2. Raw speech is mapped to the acoustic space via extracted features, which are fed as inputs to a neural network model to predict phonemes. Phoneme predictions are made with the help of secondary models such as perceptrons and decision trees, and predicted phonemes are mapped to nationalities. Although there were 35 nationalities represented in the VoxCeleb1 dataset, we removed samples from 4 nationalities which only had one audio sample each to better balance the dataset.

3.3.1 Long Short-Term Memory Network

Our first approach at predicting nationalities through phonetic attributes used Mel-frequency

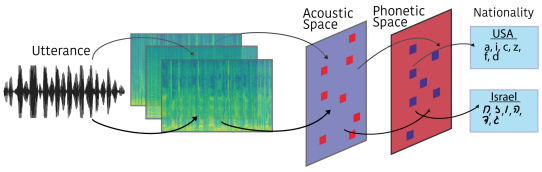


Figure 2: General architecture to identify nationality with phonemes. Raw audio is mapped to the acoustic space, which are fed as inputs into LSTM, CNN, or FF-NN for phoneme prediction, which are used to predict nationality.

cepstral coefficients (MFCCs) and a Long Short-Term Memory (LSTM) Network. MFCCs are commonly-used features in speech recognition tasks that capture phonetically important aspects of speech (Borde et al., 2015). For each recording, we extracted 13 MFCCs per 25 ms window with 20 ms overlap. We used a total of 10,000 coefficients for each recording, truncating longer vectors and zero-padding shorter ones as needed.

We fed the MFCC coefficients as inputs into a single layer bidirectional LSTM with 48 cells. We followed the LSTM layer with a dropout layer of 10% to avoid overfitting and a dense layer with 92 output cells and softmax activation to predict phonemes. We trained the network for 10 epochs using an Adam optimizer configured with a learning rate of 0.01 and binary cross-entropy loss, a commonly used loss function in multilabel prediction problems. The model contained a total of 28,126 trainable parameters.

After training the LSTM, we used it to make predictions on the validation set. Since the output predictions from the LSTM represent the probability of a phoneme being present, we defined a criterion to determine which phonemes are present in the signal. We trained 92 decision tree (DT) models (one per phoneme) to determine the presence of a particular phoneme based on the LSTM output. After extracting MFCCs from the test set in the same way, we ran the test set through both the trained LSTM and DT to predict which phonemes were present in each recording.

We created a baseline LSTM model to predict nationality directly from MFCCs as a comparison. The model had 31 cells at the output layer, representing each of the nationalities in the dataset.

3.3.2 Convolutional Neural Network

Another approach to predict nationality leveraged spectrograms and convolution neural net-

works (CNNs). Spectrograms are a visual representation of how frequencies in a signal change over time. Using SciPy's signal processing library, we computed spectrograms for each recording with a hamming window of size 25 ms with 15 ms of overlap. The resulting spectrograms were band limited to 5 kHz and were truncated in time to 200 samples to produce uniform size spectrograms for all inputs.

We fed the spectrograms as inputs to a CNN to predict phonetic attributes. While most literature in speech recognition is based in LSTM models, several studies have shown success using standard deep CNNs for audio classification (Palanisamy et al., 2020). Our CNN model relied on VGG16, a deep CNN pre-trained on the ImageNet dataset, which contains millions of images of common objects categorized into more than 20,000 categories (Simonyan and Zisserman, 2015). Since the weights in VGG16 have been extensively trained to extract meaningful features in images, we used these trained weights to extract features from our spectrograms. Using Keras, we downloaded VGG16's architecture and weights, omitting the model's top layer. We added an additional max pooling layer to reduce the complexity of the output from the VGG16's last convolutional layer, flattened the output, and passed it through a fully connected layer with 92 cells using a softmax activation to predict phonemes. The resulting model had 14,856,092 parameters, 141,404 of which were trainable and the rest of which were pre-trained. Since VGG16 expects three separate channels for RGB, we stacked our spectrograms to create a three-dimensional input. Additionally, we pre-processed the stacked spectrograms in the same way images input to VGG16 were pre-processed using Keras's pre-processing functionality. We compiled the model using an Adam optimizer with a learning rate of 0.01 and binary cross-entropy loss.

We then used the validation set output from the CNN to train a DT model to determine the presence or absence of a phoneme in the same way as was done for the output for the LSTM model. The test set was run through the trained CNN to the DT model to predict the presence of the 92 phonemes in each audio sample. For comparison, a baseline model was constructed where the output layer is 31 nodes representing the non-English speaking nationalities. In this model, no ZSL approaches were applied using the same CNN architecture.

3.3.3 Feed-Forward Neural Network

We also approached predicting phonemes from audio samples using dense FF-NNs with averaged melspectrogram values, similar to the features for gender predictions. Each audio sample was filtered with a Butterworth bandpass filter of order 4 with cutoff frequencies 75 Hz and 5000 Hz to preserve useful frequency bands (Monson et al., 2014). Melspectrograms were calculated and averaged over time so each audio sample had a 1x128 feature vector. These mel features were fed into a 12-layer dense FF-NN, where the first 11 layers were fully connected layers with a ReLU activation function, and the last layer was fully connected with a sigmoid activation function and 92 output neurons, corresponding to the 92 phonemes tags. Although the softmax activation function is more commonly used for multi-label classification, we experimented with using the sigmoid function such that the total probability summed over the 92 output nodes is not restricted to 1. We used a dropout rate of 10% between each layer to avoid overfitting, a binary cross-entropy loss function, and Adam optimizer. We experimentally determined the 12-layer network to be best for the task, and used early stopping with a patience of 5 epochs to prevent overfitting. Overall, the model had 2,068,320 trainable parameters.

The output from the neural network is a 1x92 vector of values from 0 to 1 representing confidence in the presence of each phoneme in a given audio sample. We used the validation set to train an averaged perceptron (AP) model for each phoneme to determine on a per-phoneme basis whether a given phoneme should be considered present. The test data is then passed through the neural network and the AP model to predict phonemes in each non-English nationality audio sample.

For comparison, a baseline model was constructed where the output layer is 31 nodes representing the non-English speaking nationalities. In this model, no ZSL approaches were applied using the same neural network architecture, making it an apt baseline for comparison.

3.4 Mapping From Phonemes to Nationality

After generating predictions for phonemes present in an audio sample using each of the three models as described above, we defined a map-

ping from the set of predicted phonemes to nationality using the Jaccard Index. The Jaccard Index is a measure of the similarity between sets and is defined as follows where $Pred$ is the predicted set of phonemes and $Nationality$ is the set of phonemes within $A_{phoneme}$ associated with a given nationality:

$$J(Pred, Nationality) = \frac{|Pred \cap Nationality|}{|Pred \cup Nationality|}$$

For each instance, we calculated the Jaccard Index between the predicted set of phonemes and the set of phonemes associated with each nationality. Given an instance, the predicted nationality is:

$$argmax_{Nationalities} J(Pred, Nationality)$$

4 Results

4.1 Gender Prediction

The neural network trained for 23 epochs before hitting the early stopping condition, resulting in a final training and validation loss of 0.220 and 0.256, respectively. The prediction on test data of non-English speaking nationality audio files had an accuracy of 92.67%.

4.2 Nationality Prediction

Nationality	LSTM	CNN	FF-NN
Austria	9.4	0	8.6
China	3.4	0	0
Croatia	1.7	0	0
Denmark	3.5	0	0
India	31.2	1.3	66.7
Iran	5.0	0	5.0
Israel	6.7	100	9.6
Italy	12.4	1.5	9.7
Mexico	5.0	0	4.3
Norway	1.1	0	0
Netherlands	1.0	0	6.3
Pakistan	5.3	0	0
Philippines	0.6	0	7.6
Russia	6.1	0	6.8
Singapore	1.6	0	1.6
Sri Lanka	4.3	0	0
Sudan	3.6	0	0
Sweden	0.5	0	0
Overall	13.3	0.97	25.9

Table 2: Prediction accuracy for LSTM, CNN, and FF-NN models on non-English speaking nationalities. All other non-English speaking nationalities not shown had 0.0% accuracy for all three models.

4.2.1 Long Short-Term Memory Network

The LSTM trained for 10 epochs, resulting in a final training loss 0.3253 and validation loss of 0.3260. The final output after the LSTM and DT had an overall accuracy of 13.3%. The model's performance by nationality is shown in Table 2. This model predicted 23 unique nationalities and had decently high performance across several nationalities. The baseline LSTM model achieved an accuracy of 0% as it predicted "UK" nationality for every sample in the test set.

4.2.2 Convolutional Neural Network

After training the CNN for 10 epochs, the model had a final training loss of 0.336 and a final validation loss of 0.3515. After passing through the CNN and DT, the model had an overall accuracy of 0.97%. It predicted 8 unique nationalities, only 3 (India, Israel, and Italy) of which with some success (Table 2). The baseline model, which directly predicted nationality without using phonetic attributes, had an accuracy of 0.0% on the test set, predicting "South Africa" for all samples.

4.2.3 Feed Forward Neural Network

The neural network trained for 29 epochs before hitting the early stopping condition, resulting in a final training and validation loss of 0.303 and 0.316 and accuracies of 0.0097 and 0.033, respectively. The model had an overall test accuracy of 25.9%, and a summary of accuracies by nationality are presented in Table 2. The baseline model, which directly predicted nationality without using phonetic attributes, had an accuracy of 0.0% on the test set, predicting "UK" for all samples.

5 Discussion

5.1 Gender Predictions

Overall, our gender predictions using the 4-layer dense FF-NN was very successful, with a test accuracy of 92.67%. This result is unsurprising - given that both male and female samples were present in the training and testing data, this task did not require transfer learning, and thus we expected our resulting accuracies to be quite high.

5.2 Nationality Predictions

Our models for nationality prediction had varied results. Two of our models, the LSTM and the FF-NN, achieved higher-than-chance accuracy with accuracies of 13.3% and 25.9%, respectively;

for comparison, chance accuracy in predicting 31 nationalities is 3.3%.

The poor performance of the CNN model may suggest limitations of transfer learning from image to audio processing. In using the pre-trained VGG16 model, we used weights trained to identify common objects along three color channels. Whereas the features defining objects are generally clearly pronounced, the features defining important regions of a spectrogram are small and with poorly-defined edges. This variation in important features, along with the fact that our images only contained one color channel, likely contributed to the CNN model's poor performance.

The LSTM model performed well across a range of nationalities. These high accuracies suggest that the model was likely able to pick up on some of the important phonemes defining various languages. This model may have been improved with added complexity - whereas our model utilized only one bidirectional LSTM layer with 48 cells, many LSTMs used in the literature are far more complex with multiple bidirectional LSTM layers, each of which contains hundreds of cells (Li et al., 2019). Our model's simplicity likely limited its expressivity, making it unable to fully capture variations in phonemes across languages. Unfortunately, we did not have the computational power to build a more complex model.

The FF-NN model performed well across several nationalities. The 128 averaged mel frequency bins proved to be useful features for both phoneme and gender prediction, likely because the frequency bins correspond to distinctions in aural perception by the human ear, and both tone and inflection differences between accents and genders are easily perceptible. Because we built this neural network ourselves, the number of hidden layers, nodes in each layer, and dropout rates were experimentally determined, and thus although our model predicted sufficiently well, it is by no means the optimal architecture for the task. Although the final model predicted a wide range of the 31 nationalities, it tended to predict India more often; given the high prevalence of Indian samples in the test set, this explains the FF-NN's high overall test accuracy compared to the LSTM.

The AP and DT classifiers following the neural networks proved to be key for higher accuracies - initially, we tried using a uniform threshold on the neural network outputs to determine whether a phoneme was present in the example. However, we discovered that each phoneme required a different threshold for optimal mappings, and the APs and DTs allowed us to create per-phoneme weights.

5.3 Limitations and Future Work

A major challenge we faced was in defining a mapping from phonemes to nationalities. We used a set overlap metric to map from phonemes to nationality, but in practice, this approach was not ideal as it gives equal weights to all phonemes. A more accurate approach would have been to store the phonemes present in each nationality as a distribution that weighs phonemes that appear more frequently in a given nationality's speech more heavily. Approximating such a distribution would have required another set of sample recordings or sample texts from each nationality. Further, since our approach relied on the language spoken in a given country, we were unable to distinguish between several countries with the same language. For example, Mexico, Spain, and Chile are all Spanish speaking countries, so any recording identified as Spanish was mapped to Mexico. Additional phonetic information broken up by dialects of language in various regions would have been beneficial to identify nationality across countries with the same spoken language.

6 Conclusion

Our work highlights the feasibility of applying ZSL techniques, which have largely been limited to applications in NLP and image classification, to audio classification. By using phonetic attributes of language, we were able to decompose the challenge of predicting speaker nationality to one of identifying sounds present in a recording. We were able to successfully classify audio samples from nationalities unseen during training with greater-than-chance accuracy. In addition, we demonstrated that frequency features can be used to identify a speaker's gender with high confidence. The ability to predict speaker nationality and gender effectively could have great impact in voice recognition technology and opens the door for more research in the application of ZSL to audio data.

References

Prashant Borde, Amarsinh Varpe, Ramesh Manza, and Pravin Yannawar. 2015. Recognition of isolated words using zernike and mfcc features for audio visual speech recognition. 18:167–175.

Chang Huang, Chen and J. L. Zhou. 2001. Analysis of speaker variability. volume 2, page 1377–1380.

Aditya Khamparia, Deepak Gupta, Nguyen Gia Nhu, and Ashish Khanna. 2018. Sound classification using convolutional neural network and tensor deep stacking network. pages 7717–7717.

Guo Li. 2016. Pitching in tone and non-tone second languages: Cantonese, mandarin and english produced by mandarin and cantonese speakers. pages 548–552.

Xinjian Li, Siddharth Dalmia, David R. Mortensen, Florian Metze, and Alan W Black. 2019. Zero-shot learning for speech recognition with universal phonetic model.

B.B. Monson, E.J. Hunter, A.J. Lotto, and B.H. Story. 2014. The perceptual significance of high-frequency energy in the human voice. *Frontiers in Psychology*, 5.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pages 2616–2620.

Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. 2020. Rethinking cnn models for audio classification.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

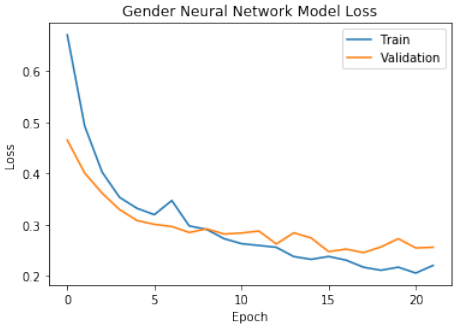
Hartmut Traunmüller and Anders Eriksson. 1995. The frequency range of the voice fundamental in the speech of male and female adults. 2.

Y. Xian, C.H. Lampert, B. Schiele, and Z. Akata. 2019. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265.

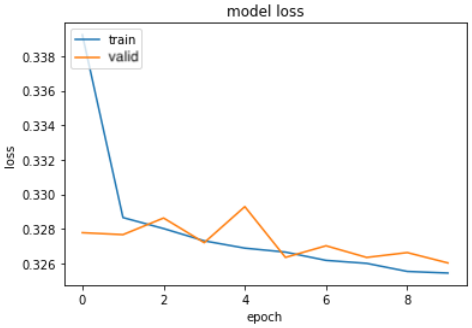
7 Video and Code

Link to Google Drive Folder with video, code, and processed data:
https://drive.google.com/drive/folders/1yysJq-upyEkm112ZDeMK7ewQ7zCkZJ8Q

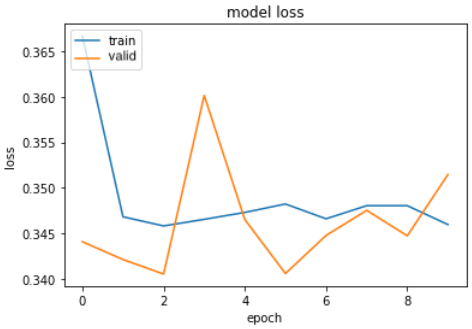
8 Appendix



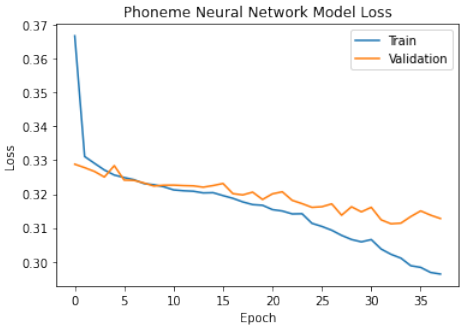
(a) FF-NN for Gender Prediction



(b) LSTM for Phoneme Prediction



(c) CNN for Phoneme Prediction



(d) FF-NN for Phoneme Prediction

Figure 3: Training and validation losses for the four models built in this project.