Speaker Profiling Using Machine Learning Methods

Ransford Antwi **Divya Bade** **Christopher Foley** **Celine Lee**

Abstract

The goal of our project was to determine the best method to profile specific characteristics of a speaker given an audio sample. We focused on classifying the gender and nationality of the speaker. Previous studies in the area of speaker identification and profiling have focused on the use of artificial neural networks or support vector machines independently for classification of either gender or nationality. We used an experimental method to determine if a semi-supervised approach wherein training on nationality and gender together would result in better accuracy than training independently. Ultimately, we found that training independently on nationality and gender with SVMs performed the best.

1 Introduction

1.1 Problem Definition and Background

Can we use neural networks to classify a speaker's raw audio data by the speaker's metadata, specifically nationality and gender? Can we further take a semi-supervised-type learning approach to improve classification, by using "learned" metadata as features to assist in learning other metadata? The ability to determine who the speaker is, based on their voice profile, opens opportunity for innovation within personal assistance devices. For example, Google's home tech devices Google Nest and Google Home use Voice Match to provide a customized user experience for each person that uses one of their devices. Having the ability to robustly determine characteristics of the speaker lends to securely being able to identify the speaker, which can also be largely beneficial for security technology in personal devices. A bulk of machine learning research today is focused on image recognition, for autonomous vehicles and other artificial intelligence that can take advantage of cameras in order to interact with humans in a human world. We believe that this project will help make strides for artificial intelligence to also take advantage of microphones and audio data, to even more cohesively interact with humans.

1.2 Our General Approach

Our first step was to find a consistent source of audio data, that has speaker voices and labels by metadata. We found that VoxCeleb had loads of data that was sufficiently noisy, varied, and reliably labeled. Once the data was imported to our drive, a bulk of our time was spent on preprocessing the data, to feed to our models. Preprocessing involved cutting all audio clips to the same length (4 seconds), sub-sampling the data to get a more fair breakdown by nationality and gender, and normalizing the data so that abnormalities would not skew our model. Once this was complete, the next step was to featurize the data. We elected to featurize our data to focus on pitch and cadence. To capture pitch, we computed fourier transformations, which transforms the audio from time domain to frequency domain and captures the relative magnitude of each frequency within the audio file. This fourier transformation as well as a calculation of f0 fundamental frequency score are used as features to represent pitch. To capture cadence/rhythm, we removed outliers and normalized the audio vector (before fourier transformation) to find relative peaks, and placed markers to see how quickly the "sharp" sounds occured in the audio. With all base featurization complete, we then fed the data into our different models, to examine which model performed the best. The models will be described in section 2.

2 Models

2.1 Model 1 - Independent CNNs

Our base model is to assign independent convolutional neural networks to attempt to classify gender and nationality, separately. For the nationality CNN, the inputs are the appended vectors of featurization defined in section 2.0, and the output is a 36-length vector that represents the vector of the nationalities: [Ireland India USA ...]. The order in which we process our data in the convolutional neural net is as follows:

Layer/Function	Hyperparameters
Conv1d	In channels = 1; out channels = 1; kernel size = 1; stride = 1; padding = 0
MaxPool1d	Kernel size = 2; stride = 2
Conv1d	In channels = 1; out channels = 16; kernel size = 1; stride = 1; padding = 0
Torch.relu	n/a
Reshape	n/a
Linear	Input size = 2704; output size = 130
Torch.relu	n/a
Linear	Input size = 130; output size = 72
Torch.relu	n/a
Linear	Input size = 72; output size = 10
torch.sigmoid	n/a

The output will be a 36-length vector in which each entry is a value [0, 1] that indicates the model's confidence that the input is of that nationality. Likewise, an identical (except for the size of the last output) CNN will produce an output of length 2, in which each entry is a value [0, 1] that indicates the model's confidence that the input is of that gender.

2.2 Model 2 - One bulk CNN

To examine whether a semi-supervised type of approach would improve prediction accuracy, we also test a bulk CNN model that predicts both pieces of metadata together. The inputs are the appended vectors of featurization defined in section 2.0, and the output is a 72-length vector that represents the cross product of the nationalities and gender: two genders for each nationality: [Ireland-male Ireland-female India-male India-female USA-male USA-female ...]. The order in which we process our data in the convolutional neural net is as follows:

Layer/Function	Hyperparameters
Conv1d	In channels = 1; out channels = 1; kernel size = 1; stride = 1; padding = 0
MaxPool1d	Kernel size = 2; stride = 2
Conv1d	In channels = 1; out channels = 16; kernel size = 1; stride = 1; padding = 0
Torch.relu	n/a
Reshape	n/a
Linear	Input size = 2704; output size = 130
Torch.relu	n/a
Linear	Input size = 130; output size = 72
Torch.relu	n/a
Linear	Input size = 72; output size = 10
torch.sigmoid	n/a

The output will be a 72-length vector in which each entry is a value [0, 1] that indicates the model's confidence that the input is that label.

2.3 Model 3 - Independent SVMs

Our third model was a basic Support Vector Machine, run with the same features of appended frequency magnitude vectors and cadence vectors. These SVMs were run independently, so like in model 1, the output for the nationality model was a 36-length vector predicting the country. The output for the gender model was a 2-length vector predicting gender. We then read the vector to predict the gender/nationality based on which index had the greatest confidence.

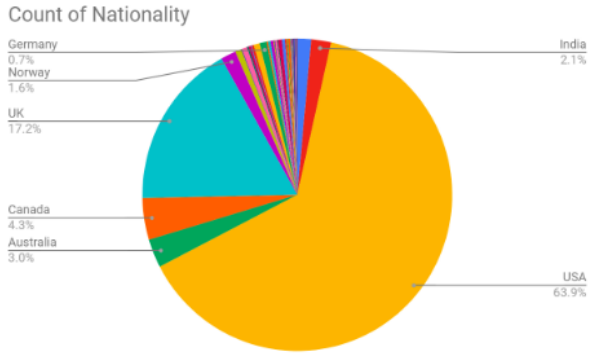
3 Hypothesis and Expectations

Our project is to experiment with different machine learning models, to determine which model will best classify a speaker's gender and nationality. Across these models, we are using the same featurization techniques. All of our models are at its core a CNN(except for the third one), which we believe to be a good training model to use for the audio data because it is particularly robust with regard to the type of data we are generating with audio. For example, it is shift invariant, which is important to handling a speaker speaking up or down an octave, because fluctuations in a speaker's tone does not change that speaker's gender or nationality. Since our categories to label are gender and nationality, we hypothesize that knowing one of these pieces of metadata will play a generous role in predicting the other. Therefore, we have designed our experiment to examine a work flow of models: we hypothesize that the model that takes on a semi-supervised learning approach will yield the best predictions.

4 Experimental Evaluation

4.1 Data Set

We chose to work with a publicly-available dataset from VoxCeleb[1]. This data set provides 4-20 second long audio clips of celebrity interviews extracted from YouTube. We used a subset of 1252 data points labeled with the nationality and gender of the speaker for our training data, with a gender distribution of 45% female / 55% male and a distribution of nationalities as depicted below.



4.2 Methodology

The first step was to sample the data set in order to get a relatively even distribution of the data. The Vox dataset had over 70GB of data and we smapled 10% of it. Initially, we sampled randomly and our nationality data was heavily skewed towards the United States and North America. So we re-sampled and set a target threshold for the number of samples we want from each country.We elected to featurize our data to focus on pitch and cadence. To capture pitch, we computed fourier transformations, which transforms the audio from time domain to frequency domain and captures the relative magnitude of each frequency within the audio file. This fourier transformation as well as a calculation of f0 fundamental frequency score are used as features to represent pitch. To capture cadence/rhythm, we removed outliers and normalized the audio vector (before fourier transformation) to find relative peaks, and placed markers to see how quickly the "sharp" sounds occured in the audio. With all base featurization complete, we then fed the data into our different models, to examine which model performed the best. For the models described earlier, each Neural Network was composed of 3 layers, any more layers could have increased the chances of overfitting the data.

[1] http://www.robots.ox.ac.uk/~vgg/data/voxceleb/

The structure of the networks was that we take in Raw Sound, convert the sound into the frequency domain via a fourier transform and we 1) predict gender using the convolutional network and 2) predict nationality using models 2 and 3 as inputs

4.3 Results

Our results were as follows. We found that independent CNNs run on our featurized vectors yielded an accuracy of about 66% for gender, and 4.7% for nationality, which are both slightly above random at 50% and 2.8% respectively. We can state that the independent CNN is capable of predicting to an extent, but not enough to be a practical learning model as it stands.

We had initially hypothesized that a joint CNN would produce better results, that predicting gender alongside predicting nationality would provide more meaningful layers within the CNN to predict nationality and gender better. However, our results showed otherwise. The accuracies were even less than random, at 34% and 0% for gender and nationality respectively. This is particularly odd, as if the model completely gave up on predicting anything at all. This would be something to look more into in a future experiment.

Interestingly enough, the model that performed best was the independent SVMs. On gender, it performed at 81% training and 88% testing accuracy, well above the 50% random. And on nationality, it performed at 44.5% training and 89% testing accuracy, also well above the 2.8% random.

	Gender Accuracy
Gender CNN	66%
Nationality CNN	n/a
Joint CNN	34%
Gender SVM n/a	Training: 81% Testing: 88% 50%
Nationality SVM Training: 44.5% Testing: 89%	n/a 2.8%

	Nationality Accuracy
Gender CNN	n/a
Nationality CNN	4.7
Joint CNN	34%
Gender SVM n/a	Training: 81% Testing: 88% 50%
Nationality SVM Training: 44.5% Testing: 89%	n/a 2.8%

	Prob | Random Dist
Gender CNN	50%
Nationality CNN	2.8
Joint CNN	34%
Gender SVM n/a	Training: 81% Testing: 88% 50%
Nationality SVM Training: 44.5% Testing: 89%	n/a 2.8%

4.4 Discussion

5 Related Work

The problem space in related work is more along the lines of building a model to determine the gender of a speaker or building a model to determine the nationality of a speaker independently of each other. Our approach is different in the sense that we're trying to predict gender and nationality simultaneously, i.e. can our predictions from gender be useful in our nationality predictions or vice versa. The current literature in predicting gender is based on signal identification using the f0 score (fundamental frequency) as the source of truth[2]. There is also extensive use of various artificial neural networks for voice-based gender classification[3]. We take a slightly different approach by using as input into our CNN the FFT representation of a speaker's sample as we believe this to be a shift invariant feature representation.

6 Future Work

One major shortcoming of our experiment design is that it assumes that our featurization technique is appropriate for the task at hand. Extraction of frequencies should theoretically work, as it examines pitch and harmonics, and the neural net should be able to subliminally generate more interesting features from the frequency information, but there are many other pieces of information about a speaking profile that are not obtainable from frequency. In future experiments, we recommend performing some more extraction that reflections on speaker audio insight such as harmonics and intonation.

Another avenue to explore, that we found throughout this project, is how predicting independent pieces of metadata can affect a model. We had initially hypothesized that predicting them together would allow us to use one as a feature to help predict the other, but our models showed otherwise. Future work should involve examining what happens when predictions and data are muddled, making for more complex models in order to predict the same thing as multiple smaller models.

6.1 Conclusion

From our project, we can conclude that with featurization via frequency bins obtained from Fourier transforms and cadence captured by marking volume peaks in an audio file, independent support vector machines ran better than independent convolutional neural networks, which itself ran better than a joint convolutional neural network. In fact, in our run of the experiment, the joint CNN seems to have completely failed, performing worse than even a random predictor. In a future experiment, it is worth exploring how predicting cross-product vector results affects the accuracy of predictors, as opposed to independent models that predict each cross-product of a model independently. We hypothesize that in fact, combining predictors might confuse the model by adding unnecessary complexity, rather than provide clear meaningful data. Another avenue to explore, that we found throughout this project, is how predicting independent pieces of metadata can affect a model. We had initially hypothesized that predicting them together would allow us to use one as a feature to help predict the other, but our models showed otherwise. Future work should involve examining what happens when predictions and data are muddled, making for more complex models in order to predict the same thing as multiple smaller models.

6.2 Video

https://www.youtube.com/watch?v=j6mal9O68lQ

[2] A. Raahul et al. 2017.

[3] Marc Gual. 2016.

References

Gual, Marc Palet. Apr 2016. *Voice gender indentification using deep nueral networks running on FPGA.* https://upcommons.upc.edu/bitstream/handle/2117/86673/113166.pdf

Raahul, A Sapthagiri, R Pankaj, K Vijayarajan, Vijayan. Nov 2017. Alternation. *Voice based gender classification using machine learning.* IOP Conference Series: Materials Science and Engineering. 263. 042083. 10.1088/1757-899X/263/4/042083.