Demographic Inference From Speech Characteristics

Sean Mullins
Department of Data Science
Data Science at Fordham University
New York, New York
smullins998@gmail.com

***Abstract*—Recent advances in digital signal processing and machine learning techniques have led to exciting possibilities about the future of audio analysis. Processing large amounts of audio data has never been easier, and the applications and demand for these models has never been so widespread. This study provides a way to use speech audio for demographic inference, specifically age, ethnicity, and gender. We use a combination of several audio features that are relevant for our inference - most of them related to the frequency spectrum - then experiment with different machine learning algorithms to determine how we can best learn our binned demographic data. Results showed promising learning capabilities with individual as well as ensemble learners. Ultimately, automatic (and accurate) demographic recognition may benefit various industries and help tailor products and services to user backgrounds.**

Keywords—machine learning, digital signal processing, inference, speech recognition

I. Introduction

As aspects of our lives increasingly embrace digitization—such as healthcare (telehealth), customer care applications, advertising, and banking—the demand for automated customization and identification has grown significantly. Extracting information from speech allows automatic and immediate inference as opposed to other methods like web-activity targeting, which needs certain amounts of data to learn. This type of recognition, for example, may be able to filter and generate better search results on the web, and likely enhance the experience of a user in various domains.

This topic has been explored the past few decades, but since audio data is expensive - usually sampled ~44K times per second - exploration has expanded in the past few years as computing power increased and other learning algorithms, mainly deep learning, have emerged. As this topic of research was emerging, researchers mainly relied on extraction techniques such as MFCC (mel frequency cepstral coefficients) extraction in combination with simple linear learners, such as Support Vector Machines [1]. Later, more advanced learners, such as the "Multi-Layer Perceptron," were introduced. This model demonstrated effective gender prediction by leveraging previously researched extraction techniques and incorporated newly introduced learning methods [2]. [3] contributed to the body of work by adopting a multi-layer feature extraction approach and combining it with KNN (K-Nearest Neighbors) learning to achieve impressive classification accuracy.

This paper extends previous research by utilizing similar feature extraction techniques, but it enhances the field by exploring various learners and optimization techniques to attain high accuracy while keeping computing costs low.

II. Experiment

A. Data

To conduct our experiment, we used two open-source datasets: Speech Accent Archive collected by George Mason University, and the Meta Automatic-Speech-Recognition dataset. The Meta ASR dataset is a 27,000-record dataset used by Meta to improve command recognition for their AI products with spoken voice commands. It consists of a diverse set of speakers in nationality and age. The Speech Accent archive is a collection of ~2,000 samples of the same sentence read each by a different reader. This dataset gives us a very wide range of speakers coming from 177 different countries.

For the experiments, we merge both sets of data, resulting in a total record count of approximately 30,000. While the Speech Accent Archive does not provide many records for our dataset, it does provide more variation in accents as the Meta ASR dataset had only ~700 speakers, each speaking several times. To decrease any bias towards models learning a person's specific voice then predicting it, we use this combination.

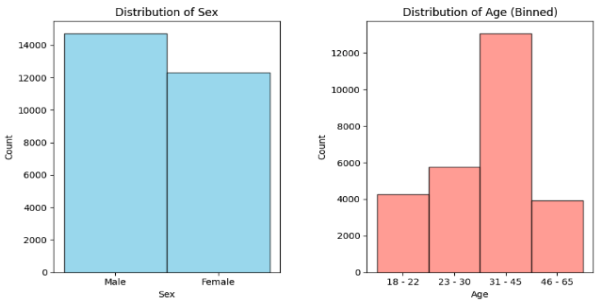

Fig. 1.2 Gender and Age Distributions

After removing missing and incomplete data, the full dataset contains ~25,000 records. The dataset has a fair gender distribution, while age and ethnicity distributions somewhat skew towards 30-45 year olds, and English speaking regions, respectively.

B. *Features*

Figure 1.3 outlines the experimentation method from our first data aggregation step to the final output learner. From our raw data we apply a few preprocessing steps to ensure the audio is ready for feature extraction.

First, we apply a trim to each audio clip, ensuring there is no silence at the beginning or the end of the clip. Additionally, we apply a "pre-emphasis" filter, which aims to raise the upper end of the frequency spectrum resulting in frequencies above ~3KHz becoming louder and more prominent. This is a vital step in the process as most of the harmonic content above 1-2KHz is attributed to the uniqueness of a voice. Finally we apply normalization of the audio followed by light dynamic compression to even out peaks and troughs. This aims to balance peak loudness in each audio file, so the feature extraction process is not affected by decibel range.

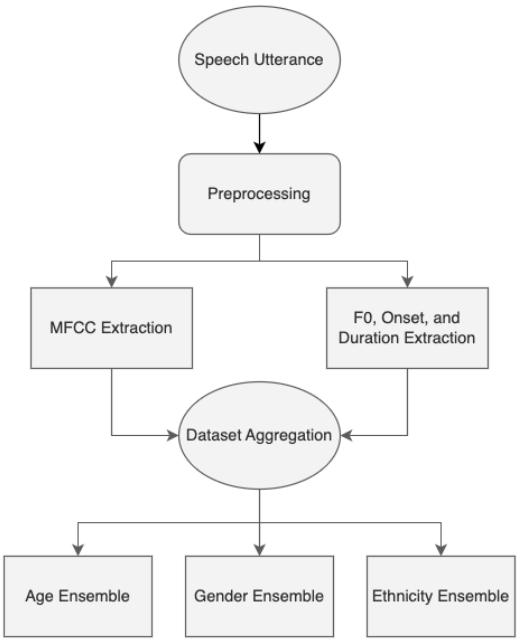

Fig. 1.3 Pipeline Description

We build off of previous work done in [1] and adopt the MFCC (Mel-Frequency Cepstrum Coefficient) extraction method for our main model features. We use 20 coefficients and, unlike related work, take the mean of each band, compressing the spectrum, and use the 20 point values as features for our model. Compressed coefficient values may depreciate the expressiveness of the MFCC method, but this will ultimately make model training faster, and expend less computing power.

Our second feature set explores the methods in [5] and extracts fundamental frequency, duration of the audio clip, number of words per second, and tempo, or cadence of how the speaker talks. The reason for this second set of features, instead of just a MFCC extraction method, is that we hypothesize that each contributes in some variation to predicting our classes. For example, fundamental frequency, or the lowest frequency in a person's voice, tends to be twice as low - ~100Hz - for men compared to women. Additionally, we believe that younger people may talk faster, so adding this second set will emphasize aspects of the audio that may be lacking in MFCC extraction.

We can visualize an example feature as it relates to one of our classes: age. Figure 1.4 shows the distribution of fundamental frequency, words per second, and two MFCC coefficients. We can see the sizable variance between male and female in the 15th MFCC coefficient, and we can also see a difference in distributions when looking at the fundamental frequency.

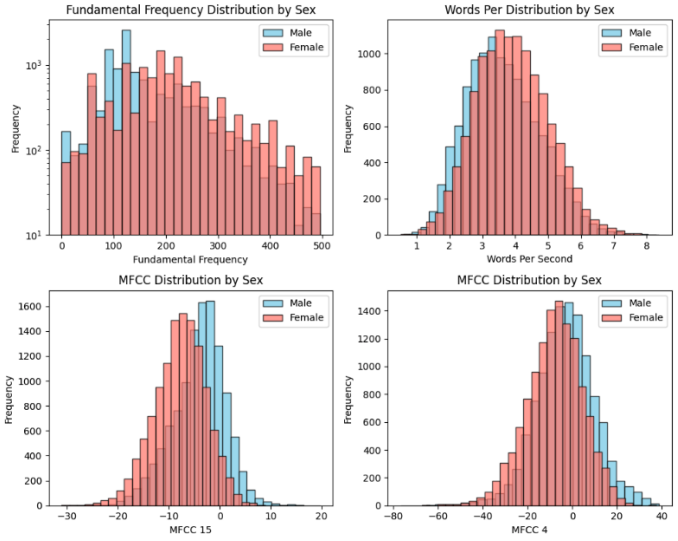

Fig. 1.4 Feature Related Age Distributions

C. *Model*

For model building we combined and adapted the processes from [3] and [4].

We are looking to predict three basic demographics from speech audio: gender, age, and ethnicity. Ultimately, we generate three separate models for each classification, all very similar. It is also possible a multi-output model could be used, but in an effort to separate workflow and experiment with parameters, this seemed like the best option. For each experiment we partition our data into a training, testing, and

validation set, each containing 80%, 10%, and 10% of the data, respectively.

In predicting gender, we did not modify the feature set before feeding it into the model as the distribution between female and male was very even. We first experimented with multiple learning models like Logistic Regression, Multi-Layer Perceptron, and KMeans clustering before landing on one model that outperformed the rest. The best model was a K-Nearest Neighbors model (KNN) with the 'neighbors' parameter varied from 1-7. Our hypothesis is that the KNN model works very well with MFCCs because in a 20 dimensional space, the nearest neighbor is likely to be something that replicates specific harmonic content very well, and is generally going to be the same gender as a point nearest to it.

We also saw that a Random Forest Ensemble classifier worked well when we set the number of estimators, or trees, to 100. The thought process for the high number of estimators is similar to that of KNN, where we think that increased granularity in a 20-25 dimensional space should yield best results as people's speech characteristics are very unique. Given both KNN and Random Forest worked well, we created a soft (probability) voting classifier as our final model for gender prediction. A soft voting classifier considers the probabilities provided by each individual classifier, averaging them before making the final prediction. Within the voting classifier we use several models listed below:

- Random Forest with the number of trees, or estimators, set to 100
- A KNN model with the number of neighbors set to 1 and the weights set to 'distance'
- A KNN model with the number of neighbors set to 3
- A KNN model with the number of neighbors set to 5

Next we look to predict the age of an individual. In related works including [6], we see regression models built for age prediction, using MAE and RMSE as loss functions. This is a useful way for profiling, but for more general demographic profiling, we bin age values and treat this as a classification problem. We use several bins and theorize which age ranges may be best grouped together, hence more useful as output: 18-22, 23-30, 31-45, 46-65.

We show the age distribution in Figure 1.2, and we can see that there is a large concentration in the bin 31-45, almost double any other bin. Before model building, we perform minority random oversampling until all of our classes have the same amount of records. This does mean the dataset will contain a fair amount of synthetic samples, but undersampling would not have been feasible given the size of the dataset, and the number of speakers in it. Before constructing the model, we hypothesized to determine the optimal features and models for age prediction. We thought that the fundamental frequency of a person's voice would play a crucial role in age prediction compared to other features. As a result, we assigned greater weight to this specific feature before incorporating it into the model. However, achieving an accurate age prediction remains challenging, even for sophisticated models such as Neural Networks. For instance, a Deep Neural Network applied to the TIMIT dataset, as described in [7], yielded age prediction errors ranging from 8 to 9 years.

For the age model, we used an identical soft voting classifier that we used for our gender prediction. This allows us to combine the features from the varied KNN models and a Random Forest model, resulting in a more robust and accurate classifier.

Finally, we predict ethnicity and combine techniques we used in the previous two models. Although some of the ethnicity data came in a more granular form - like country of origin and first language - we sought to bin these into similar categories so as to not have to have a large multi-output model. For example, the distinction between neighboring countries like Lithuania and Latvia may not provide useful information to the user. However, a group like "Eastern-European" will surely provide some value. We list the bins we used below:

- Black or African American
- White
- Native American, American Indian, Alaska Native
- Hispanic, Latino, or Spanish
- Asian, South Asian, Asian American
- Middle Eastern, or North African
- Pacific Islander

Before we feed our data into our last model, we see from Figure 1.5 that the distribution of ethnicities is uneven. To remedy this, we use minority random oversampling until all bins are equal. We also one-hot encode our bins as this category does not have an ordinal characteristic.

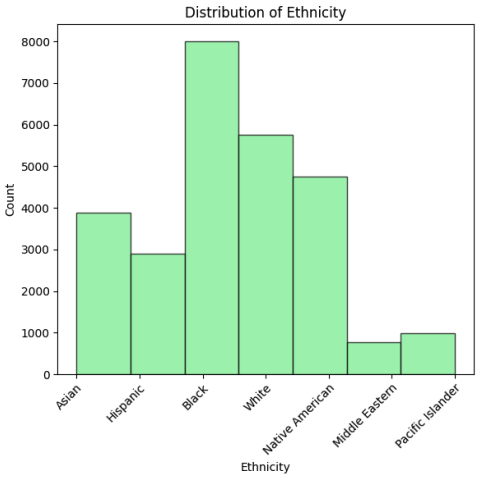

Fig. 1.5 Class-Ethnicity Distribution

We use another similar model to predict ethnicity, but this time we use a soft voting classifier on only three KNN models with 1,3, and 5 neighbors. Unlike previous predictions, the

Random Forest did not perform well, so we left this out from the ensemble. A model summary is listed below in Figure 1.6.

Model Predictions	
Gender	RandomForest(estimators=100), KNN(neighbors=1, weights="distance"), KNN(neighbors=3), KNN(neighbors=5)
Age	RandomForest(estimators=100), KNN(neighbors=1, weights="distance"), KNN(neighbors=3), KNN(neighbors=5)
Ethnicity	KNN(neighbors=1, weights="distance"), KNN(neighbors=3), KNN(neighbors=5)

Fig. 1.6 Model Specifications

III. Results

A. Model Results

We have performed three experiments using slightly different methods of data processing and model architecture. For each experiment we present the accuracy results using 5-fold cross validation in Figure 1.7.

Group	**Trained on**	**Results**		
Group	Accent Archive & Meta ASR	Accuracy	Precision	Recall
Gender		95.00%	91.70%	89.60%
Age		87.70%	88.20%	87.70%
Ethnicity		96.10%	94.30%	92.10%

Fig. 1.7 Final Group Results

We found experimentally that the use of additional features - fundamental frequency, duration, and words-per-second - did increase accuracy percentages in each experiment by 2%-3%, as opposed to just MFCC feature extraction. We also hypothesized that certain MFCCs representing certain frequency ranges may have had more impact on identifying and classifying a voice. For example, we think it is likely that the range 300Hz - 500Hz is not too important as these frequencies are just octaves of the fundamental frequency. We were wrong in this assumption, and it turns out that each coefficient contributed heavily to the final accuracy. We also varied the number of MFCC coefficients to 40 to understand if a more granular approach of smaller frequency segmentations would be better, but this ultimately decayed each model's performance by 3%-4%.

Additional augmentations to the data, such as random oversampling, improved each model's accuracy by ~1%-2%. It is difficult to justify using synthetic data in any case, especially when it relates to a pseudo-frequency spectrum like MFCCs, but ultimately some of our classes had 60%-70% number of samples below the majority class, so this was needed to balance our individual classes.

Focusing on our learners, the KNN algorithm proved to be the most robust, with accuracies that spanned 87%-95% between the three categories that we were looking to predict. The introduction of a voting ensemble added ~1% on top of the best individual classifier. Conducting a sensitivity analysis guided the selection of the constituents for the voting ensemble. Although there are three of the same algorithms in the voting ensemble, KNN, addition or subtraction of other algorithms proved to decay model performance. Additionally, a Random Forest is incorporated to effectively process the non-MFCC features - such as fundamental frequency and words per second. This results in a more diverse set of probabilities being fed into the voting ensemble, not just probabilities from three KNN models.

In the realm of age classification, related works have struggled to achieve high accuracy, especially when adopting a regressor for exact age prediction. Our binned-age method proved to help in straddling the line of high accuracy and usability of model output. However, we did experiment with a regressor to try to predict the exact age of a sample. We used both a Logistic Regression and a Perceptron and achieved a MSE (mean squared error) of 310 and 110, respectively. Converting the error into age ranges, we can see the better classifier (Perceptron) misclassified each sample by about 10 years in either direction, which is ultimately no better than using an age-binned approach of ~10-15 years. This justifies our decision to use a binned, and less specific approach.

B. Other Experimentations

As recent related works, specifically [8], have used deep learning and perceptron architectures to achieve exceptional accuracies, we thought to experiment with simple neural networks and a CNN (Convolutional Neural Network) to understand if the addition of this to our voting classifier would prove beneficial.

On each output class we constructed a neural network with an input layer with 64 units, two hidden dense layers of 32 and 16 units, respectively, and a dense output layer with a softmax activation function to classify our output. We then compile the model using the Adam optimizer, and categorical cross-entropy loss. We found that this architecture, although simple, underperforms more basic machine learning algorithms like KNN, Random Forest, and even logistic regression. More complex architectures with more data may prove to be better in achieving acceptable accuracy, though.

Additionally we took inspiration from [9] to experiment with a CNN. [9] details a network which generates music from text, and does it through replicating the frequency spectrum of a song as an image. From this, the idea of classifying age, gender, and ethnicity based on frequency spectrum images seemed possible, and even probable. To do this, we created images for each speech record using the mel-frequency spectrum, which is a frequency spectrum that is spaced logarithmically in Hertz to simulate the way the human

auditory system processes audio. We then feed the images into a the CNN in which we detail below:

- Convolutional (1): 32 filters (3x3), ReLU
- Convolutional (2): 64 filters (3x3), ReLU
- Max Pooling: 2x2 size
- Fully Connected: Dense (128, ReLU)
- Output: Dense (Softmax)
- Compilation: Adam, Categorical Crossentropy

We omit the thought process for these parameters as both the CNN and neural network were aimed at understanding if model performance and intuition were close to the accuracies of our final ensemble learner. For each classification the neural network and CNN achieved accuracies near 71% and 62%, respectively, which was very underwhelming. Experimentation with parameters and increases in speech data may benefit these particular models and help achieve better accuracies, but it seems for this classification task, and with this particular data size and variety, simple learners are more efficient, and are better able to generalize to new speech.

IV. Conclusion

Our experiments traversed a number of machine learning models, featuring a K-Nearest Neighbors (KNN) model, a Random Forest model, and a Voting Ensemble. We proved it is possible to generate an accurate, and computationally inexpensive model for demographic inference through speech. All models achieved acceptable accuracies and showcased the effectiveness of KNN, particularly in leveraging Mel-Frequency Cepstrum Coefficients (MFCCs) within a high dimensional space for demographic recognition.

Future research may vary processes discussed in this paper to improve results. Embracing more sophisticated deep learning architectures, such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), offers the potential to capture intricate patterns within the data. The exploration of transfer learning, harnessing pre-trained models on expansive speech datasets and fine-tuning them for specific demographic tasks, could enhance model performance when faced with limited data. Further advancements might arise from the investigation of data augmentation techniques tailored to speech, including pitch shifting, time warping, and noise introduction, to bolster the models' robustness. Additionally, considering alternative feature extraction methods beyond Mel-Frequency Cepstrum Coefficients (MFCCs) and the explored feature set could offer a better representation of speech characteristics. Lastly, scaling up datasets in terms of size and diversity remains a pivotal avenue for advancing the generalization and performance of these models.

References

[1] Mahmoodi D., Marvi H., Taghizadeh M., Soleimani A., Razzazi F., Mahmoodi M. Age Estimation Based on Speech Features and Support Vector Machine; Proceedings of the 2011 3rd Computer Science and Electronic Engineering Conference (CEEC); Colchester, UK. 13–14 July 2011; pp. 60–64. [paper]

[2] Jasuja L., Rasool A., Hajela G. Voice Gender Recognizer Recognition of Gender from Voice using Deep Neural Networks; Proceedings of the 2020 International Conference on Smart Electronics and Communication (ICOSEC); Trichy, India. 10–12 September 2020; pp. 319–324. [paper]

[3] Uddin M.A., Hossain M.S., Pathan R.K., Biswas M. Gender Recognition from Human Voice using Multi-Layer Architecture; Proceedings of the 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA); Novi Sad, Serbia. 24–26 August 2020; pp. 1–7. [paper]

[4] Livieris I.E., Pintelas E., Pintelas P. Gender recognition by voice using an improved self-labeled algorithm. Mach. Learn. Knowl. Extr. 2019;1:492–503. doi: 10.3390/make1010030. [CrossRef] [paper]

[5] M. Notter, "Age prediction of a speaker's voice," EPFL Extension School, Feb. 18, 2022. [Online]. Available: [link].

[6] Kalluri S.B., Vijayasenan D., Ganapathy S. Automatic speaker profiling from short duration speech data. Speech Commun. 2020;121:16–28. doi: 10.1016/j.specom.2020.03.008. [paper]

[7] Kalluri S.B., Vijayasenan D., Ganapathy S. A Deep Neural Network Based End to End Model for Joint Height and Age Estimation from Short Duration Speech; Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Brighton, UK. 12–17 May 2019; pp. 6580–6584. [paper]

[8] Buyukyilmaz M., Cibikdiken A.O. Voice gender recognition using deep learning; Proceedings of the 2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications; Xiamen, China. 18–19 December 2016; pp. 409–411. [paper]

[9] S. Forsgren and H. Martiros, "Riffusion: a neural network for generating music using images of sound," December 15, 2022. [Online]. Available: [link].