3 Age Prediction on VoxCeleb

3.1 Problem Description and Literature Survey

Speaker age prediction has practical commercial applications such as age-dependent advertisements, age-adaptive dialogue adjustment in vehicles, caller-agent pairing, as well as forensic applications such as determining the age of suspects from their voice and collecting biometric evidence. Current research shows that humans are able to make accurate inferences about people's age by listening to their speech [36], but automatic paralinguistic information recognition and analysis is challenging due to limited amount of public corpora with realistic data [65].

The Age Sub-Challenge at Interspeech 2010 [64] led to substantial work on speaker age prediction. Feld et al. [19] present a GMM-SVM supervector system for speaker age preeiction that is adopted from speaker recognition research, and an experimental study on parameter selection. One of the challenges in age prediction through voices is the accuracy in underrepresented age groups. Lingenfelser et al. [39] apply decision level fusion and ensemble methods including Mean Rule and variants of Cascading Specialists on the AGENDER corpus [64], and these methods achieve balanced classification accuracy on all classes. Li et al. [38] and Bocklet et al. [8] build the age prediction system by combining multiple models together. The former focuses on acoustic-level methods and fused multiple GMM-SVM mean-supervector systems to improve classification accuracy. The latter builds five models in different feature spaces including spectral features, prosodic features, and glottal features, and combined these models in different configurations.

More recent research work on this topic includes [30] and [31], in which the authors explore methods to infer non-linguistic traits including age based on short durations of speech. Grzybowska et al. [27] combine *i-vector* and acoustic features to perform age regression, and apply cosine distance scoring as well as regression result mapping for age classification. Their model achieves an unweighted accuracy of 62.9%. Their classification is more accurate for the children, senior male, and senior female age groups, which is similar to our results shown in section 3.3.1. Ghahremani et al. [26] apply an end-to-end *x-vector* DNN architecture for age estimation which outperforms an i-vector baseline. Datasets used in these papers include aGender, NIST SRE08, and NIST SRE10, but they are not suitable for our project because they are not longitudinal. The goal of our project is to evaluate change in speech over time, so instead of using these datasets directly, we choose to curate a longitudinal dataset in which the same speakers talk over a long period of time.

3.2 Methods

Before using speech analysis to detect aging-associated diseases like dementia, we must see how well we can predict age from a person's speech. We also want to see if we can observe the aging process by analyzing how a person's speech change over years.

3.2.1 Data Preparation

We use the VoxCeleb datasets [47][9] for age prediction tasks. VoxCeleb1 and 2 are audio-visual datasets consisting of short clips of human speech, extracted from interview videos uploaded to YouTube. They are the state-of-the-art datasets for speaker identification and verification, and they are transfromed to be a dataset suitable for age prediction.

Both VoxCeleb1 and VoxCeleb2 provides gender information of the speaker as well as the Youtube URL from which the video is extracted. However, only VoxCeleb1 provides celebrity names, while VoxCeleb2 is anonymous. Although the speakers in VoxCeleb2 can be identified by watching Youtube videos, their names are not provided as metadata. In order to make the dataset suitable for age prediction tasks, data cleaning and curation is done to add age information and remove unusable data. VoxCeleb2 does not have speaker name information available, so we had to take more steps to get the speaker name and age. First the Youtube video title is requested through Youtube API, and we then run a named entity recognition algorithm on each video title. After extracting PERSON entities from video titles, we look at all the names corresponding to the same speaker id. If count of most common name is greater than the double count of the second most common name, we select the most common name as the celebrity name. Speakers for which we are not confident about their names are removed from the dataset.

Age can be annotated given speaker names and video URLs. Video publish dates are requested through Youtube API, and birth dates of celebrities are scraped from Wikipedia. Age at the time of video is calculated as the difference between video publish date and celebrity birth date. Videos clips for which we cannot get age annotation are removed from the dataset. For example, certain Youtube URLs are no longer available. Also, multiple people may have the same name so that we cannot determine who is the speaker. There are also cases when the birth date is not public on wikipedia. Celebrities who have passed away are also removed from the dataset, because some of the videos are published after their death date.

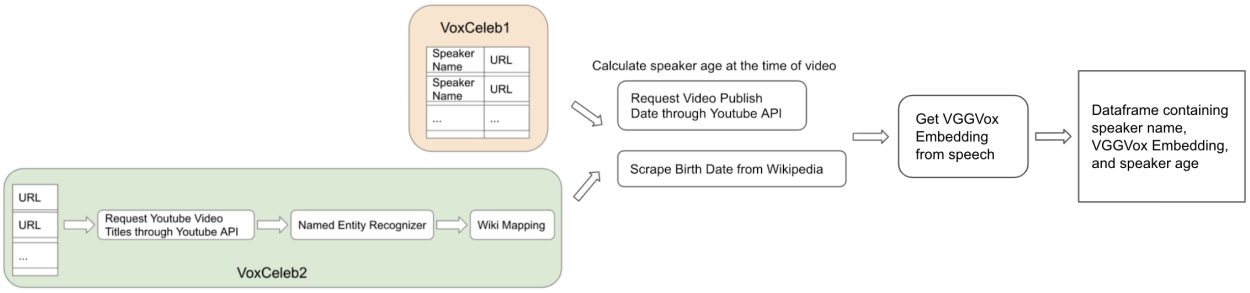
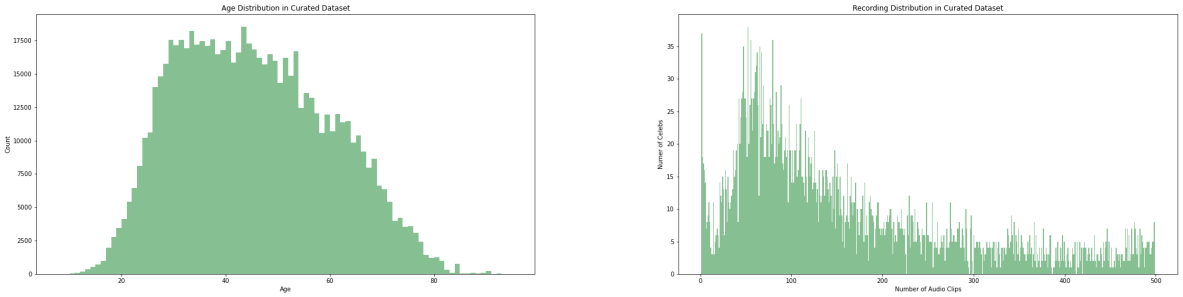


Figure 1: Steps taken to impute age.

The features we use for age prediction are the VGGVox embeddings [48]. They are 1024-

dimensional feature vectors obtained from pre-trained models that map spectrograms to a compact embedding space and get distance measurements corresponding to speaker similarity. The data pre-processing steps are shown in Figure 1. After these steps, 730,281 utterances from 4,432 speakers remain in the dataset. Of the remaining speakers, 34.7% are female, and 65.7% are male. The dataset is then partitioned into non-overlapping training set and test set.



(a) Number of clips for each age.

(b) Number of celebrities vs. number of appearances.

Figure 2: Age distribution and recording distribution in curated dataset.

Celebrity age distribution for the curated dataset is shown in figure 2a, which is a normal distribution with mean in the middle age group. There are fewer data points in the young adulthood and older adulthood. The average age in the dataset is 45. Figure 2b shows the distribution of the number of audio clips for each celebrity. Most of the celebrities appear for multiple times in the dataset, which enables us to do analysis on the process of aging. The average number of appearances is 165.

3.2.2 Models

For the age classification task, speaker ages are divided into five equal buckets so that each bucket has the same amount of data. We apply a variety of machine learning models including logistic regression, support vector classifier, decision trees, random forest, KNeighbors classifier, and AdaBoost to perform multi-class classification. The overall accuracies of these models are shown in Table 1.

Descriptions of each machine learning model are as follows:

- Logistic regression is named for the logistic function. It is an S-shaped curve that takes any real-valued number and maps it into a value between 0 and 1. Input values are combined linearly using weights or coefficient values to predict an output value. In a classification problem, logistic regression gives the probability of a data point being in each class and chooses the class with the highest probability as the prediction result. Logistic regression does not naturally support multi-class classification, but the one-vs-rest scheme used for

our experiments calculates the argmax of probabilities obtained by each model to make prediction.

- Support vector classifier finds a hyperplane that best divides a dataset into separate classes and maximizes the margin between the hyperplane and points closest to the hyperplane.
- Decision trees infer decision rules from training data and fork in tree structures until a prediction decision is made for a given record.
- Random forest is an ensemble method consisting of an uncorrelated forest of decision trees to make overall combined predictions.
- KNeighbors classifier calculates the distances between test data and all training data, selects a specified number of points closest to the test data, and votes for the most frequent label.
- AdaBoost [21] is a boosting technique that combines weak learners and adjusts error metric for each iteration to find the best split.

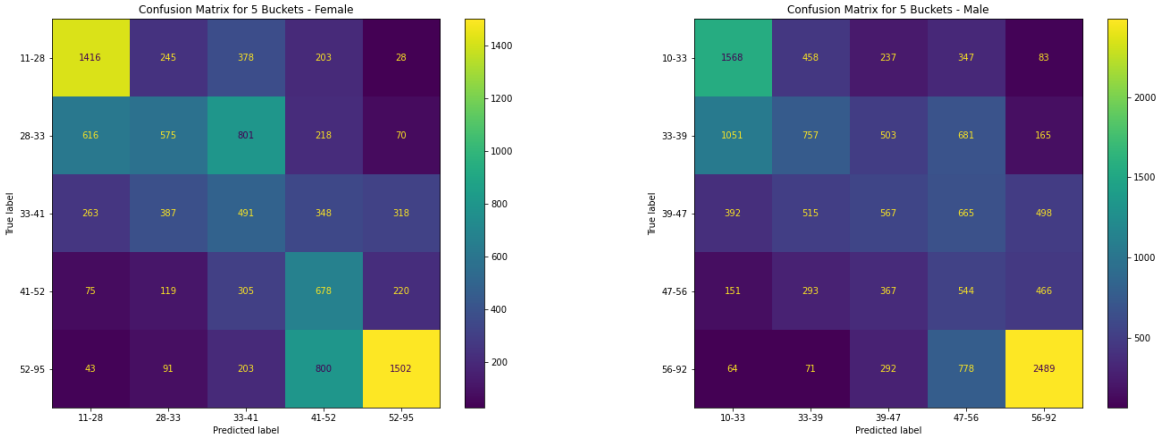
	VoxCeleb1		VoxCeleb2	
	Female	Male	Female	Male
Random Guessing	0.200	0.200	0.200	0.200
Logistic Regression	**0.473**	0.427	**0.448**	**0.374**
SVC	0.447	**0.438**	0.420	0.360
Decision Trees	0.303	0.314	0.308	0.286
Random Forest	0.444	0.399	0.412	0.348
KNeighbors	0.342	0.334	0.332	0.303
AdaBoost	0.410	0.383	0.389	0.348
Multi-layer Perceptron	0.421	0.365	0.384	0.347

Table 1: Combined classification accuracy on 5 equal buckets.

3.3 Experimental Results

3.3.1 Age Group Classification Accuracy

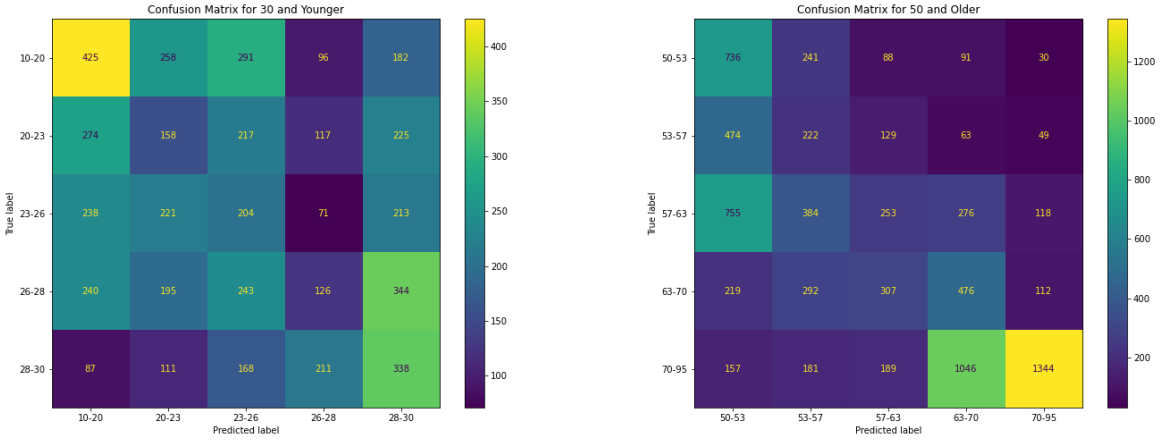
The one-vs-rest logistic regression classifier achieves the best overall accuracy. Its classification result on VoxCeleb1 is shown on confusion matrices in Figure 3. The classifier does well identifying the youngest and oldest age groups, but it confuses the rest. To examine how gender plays a role in age prediction, we separate the dataset by gender, and the classification result shows that the youngest bucket and the oldest are still the most accurate ones.



(a) 5 equal buckets - female (b) 5 equal buckets - male

Figure 3: Confusion matrices for classification result on 5 equal buckets.

We further divide the youngest and the oldest age groups into five equal buckets, and plot the classification results as confusion matrices shown in Figure 4. It is easy for the classifier to distinguish the youngest speakers (age group 10-20) and the oldest speakers (age group 70-95), but it has worse performance for age groups in between. This is similar to a human listener's performance since we may find it easier to distinguish childrens' and older people's voice compared to other age groups.

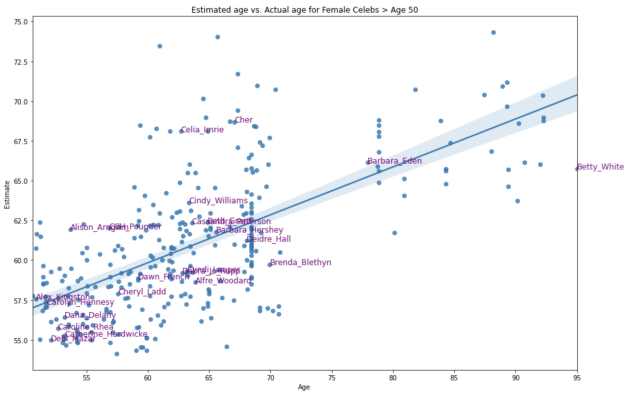


(a) 5 equal buckets - 30 and younger (b) 5 equal buckets - 50 and older

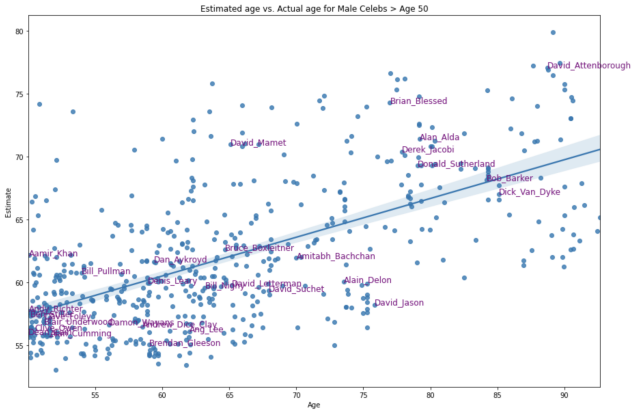
Figure 4: Confusion matrices for classification result on the youngest and oldest buckets.

3.3.2 Age Regression Accuracy

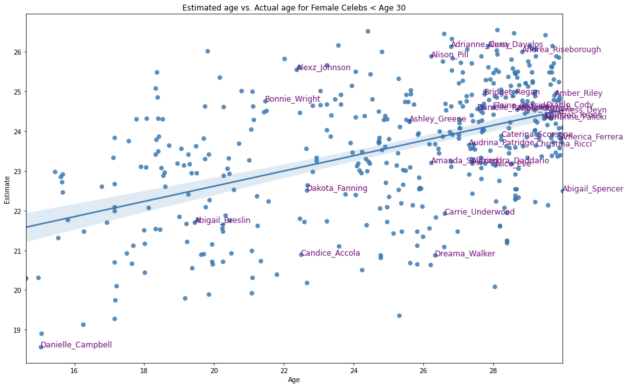
For the age regression task, we estimate the predicted age to be the product of the midpoint of age group and the probability of that age group obtained by logistic regression. Dot plots of estimated age vs. actual age are shown in Figure 5 for the most famous celebrities younger than 30 and older than 50, separated by gender. From the figures, we can see that positive correlation exists between estimated age and actual age, and the correlation is more impressive for the older age group.



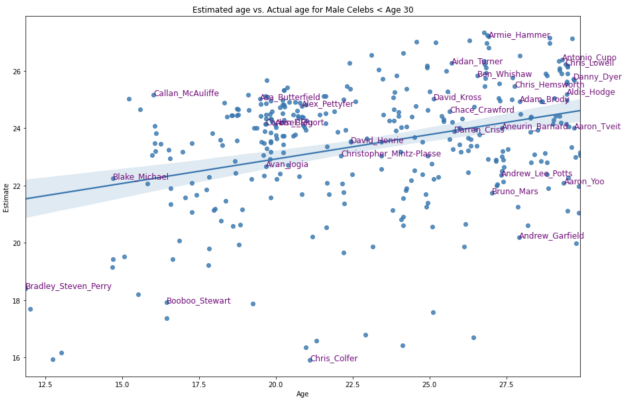
(a) Female>50, Pearson correlation=0.6487



(b) Male > 50, Pearson correlation=0.6160



(c) Female < 30, Pearson correlation=0.4928



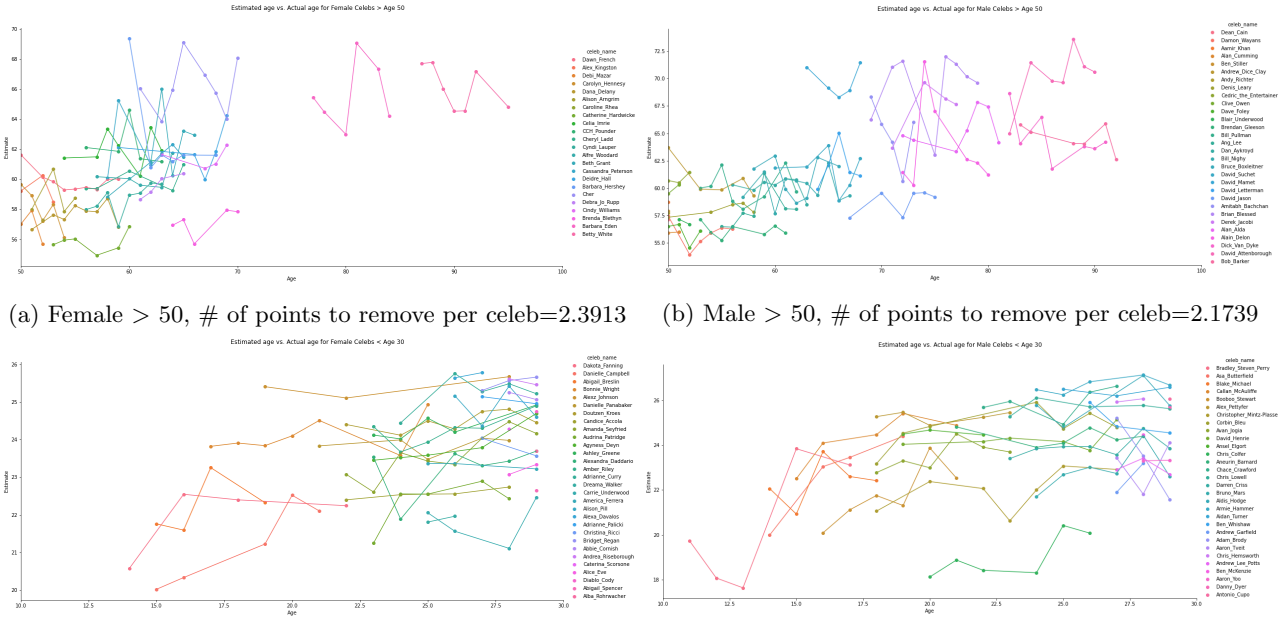
(d) Male < 30, Pearson correlation=0.3339

Figure 5: Estimated age vs. actual age with regression line.

For celebrities older than 50 (subfigures 5a and 5b), more points are clustered in the bottom-left corner; while for celebrities younger than 30 (subfigures 5c and 5d), most of the points are clustered in the top-right corner. This shows that extremely old and extremely young celebrities comprise of only a small part of the dataset. A few ouliers can be observed within these groups. In the older female group, voice actress Betty White is an outlier where her actual age is over 90, but her estimated age is around 65.

3.3.3 Monotonicity

A good age prediction result should reflect each individual process of aging. To visualize how our estimated age changes as the real age of speaker increases, dots belong to the same speaker are connected and drawn as a polyline in Figure 6. The number of points that need to be deleted to make these polylines monotonically increasing are calculated as a longest increasing subsequence problem. The average number of points to be removed per speaker is calculated for comparison. Based on the result, aging in the younger age group is more captured by the regression compared to the older age group, since fewer points need to be removed to make all the polylines monotonically increasing in the younger age group.



(a) Female > 50, # of points to remove per celeb=2.3913 (b) Male > 50, # of points to remove per celeb=2.1739

(c) Female < 30, # of points to remove per celeb=1.25 (d) Male < 30, # of points to remove per celeb=1.5667

Figure 6: Estimated age vs. actual age as polylines.

Permutation tests are performed to check if our result is statistically significant. Random permutations of estimated age are generated, and the number of points to be removed to satisfy monotonicity is calculated for each permutation. The p-value for female celebrities older than 50 and male celebrities older than 50 are 0.167 and 0.931 respectively, which means the monotonicity result for the older age group is not statistically significant. The p-value for female celebrities younger than 30 and male celebrities younger than 30 are 0.006 and 0.019 respectively, which indicates that observations in the younger age group are less likely to be obtained by chance, since they have a p-value close to zero.