# An speaker age recognition system based on spectrum SIFT features

Thales A. de Lima    Márjory C. da Costa-Abreu

January 25, 2021

**Abstract**

This work gives a substantial literature review, and proposes applying the Scale-Invariant Feature Transform to identify a speaker age through the spectrogram of the speech/sound sample. We use the Portuguese subset of Common Voice in our experiments with a KNN, DNN, and SVM.

## 1 Introduction

In the present, paper we want to investigate the most recent approaches to age classification or regression through voice. Then, with the collected information we answer the following question: is age classification or regression research at a mature stage? Besides, follow-up informations are extracted from the data. This allows to find possible gaps in the literature, further applications, and predict research directions in the field.

Also, throughout the text the term classification refers to any of: identification, verification, regression, detection, and any other synonym used in this subject.

### 1.1 Related Works

This section provides a general overview of the methods used for classification of an speaker age from voice or sound.

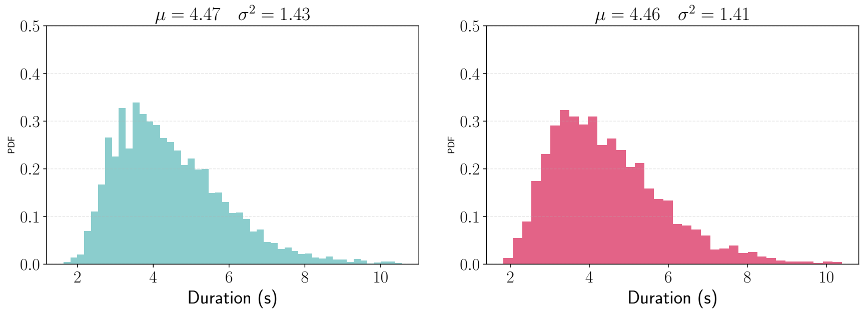
## 2 Materials and Methods

### 2.1 Speech data

In this work, we use the Portuguese subset from Common Voice 6.1 []. The dataset has a total of 63h of validated speech from 1,120 speakers. Audio are recorded using 48kHz sampling rate, and 16bit resolution. The age is divided into 7 categories, from which 6 are defined in Table 1. Besides, gender is highly unbalanced with 90.81% of male speakers, 7.49% of female, and 1.69% for unknown gender. This difference is taken into account during our analysis. Furthermore, the content of each recording differs between speakers and between the same subject.

Table 1: Age ranges in the dataset after cleaning.

| Age | Label | Cleaned (%) | Train (%) | Test (%) |
|---|---|---|---|---|
| 00 ⊣ 19 | teens | 858 (6.81) | 600 (6.81) | 258 (6.83) |
| 19 ⊢ 30 | twenties | 4016 (34.36) | 3028 (34.36) | 1298 (34.36) |
| 30 ⊢ 40 | thirties | 4136 (33.95) | 2992 (33.95) | 1282 (33.94) |
| 40 ⊢ 50 | forties | 1907 (16.88) | 1488 (16.88) | 638 (16.89) |
| 50 ⊢ 60 | fifties | 804 (7.19) | 634 (7.19) | 272 (7.2) |
| 60 ⊢ 70 | sixties | 96 (0.77) | 69 (0.78) | 29 (0.76) |
| Total | — | 12588 (100) | 8811 (100) | 3777 (100) |



(a) Training.

(b) Test.

Figure 1: Distribution of durations from training and test subsets.

The Common Voice already has a training, develop, and test subsets. The *seventies* category was discarded because had only 9 samples. While looking into the data, 3,150 samples were removed because had no age information leaving a total of 12,588 samples from originally 15,747. Furthermore, the data had to split into training, and test since samples were removed from some of these subsets and the age categories were not uniform among them. Figure 2 has a comparison between the original distribution and the results from our splits. We split data into 8811 samples for training (70%) and 3777 for test (30%) preserving the proportions for age categories and gender, as presented in Table 1. Also, this procedure reduced our unique speaker from 1,120 to 372. Finally, the duration distribution of training and test sets can be visualised in Figure 1.

## 2.2 Describing Age from Speech

Voice, as any other signal, has several representations. One of them is the spectrogram which has several variations. In general, this presents a signal as it varies through the frequency domain. In Figure 3 there are 8 spectrograms at 40dB from random Common Voice recordings.

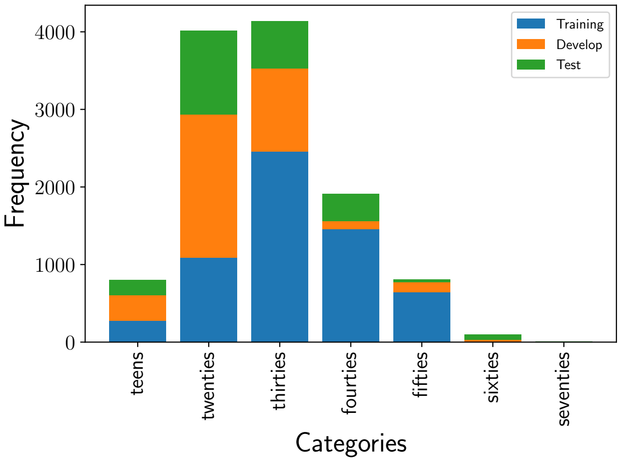


Figure 2: Data distributions before and after cleaning.

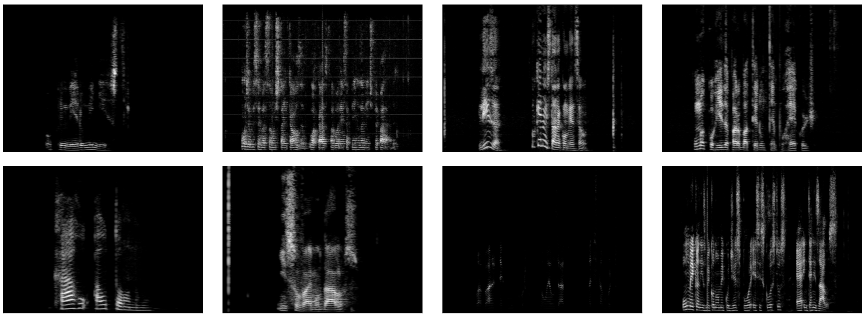


Figure 3: 40dB spectrum of several samples from Common Voice.

### 2.2.1 Scale Invariant Feature Transform

The Scale Invariant Features Transform is an image-related technique to identify points of interest from a image. Some characteristics that do not change through, scaling, rotation and translation []. Also, minimally affected by noise. These main characteristics can be very useful for speech processes. Since most of the time a voice signal can vary in length (both overall duration, speaking rate, etc.) and time (by taking the same content in different time instances). Therefore, approaching speech tasks with SIFT may provide interesting results. Mainly, in this work, we extract these features from 40dB log-spectrogram of voices attempting to identify the age of a speaker.

This technique works by generating a image pyramid where each level has a transformed version of the image. Using a Gaussian Kernel (smoothing), the first image is generated by convolving the original version with it. Both horizontally and vertically, twice, given that Equation (1) is separable

$$g(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-x^2/2\sigma^2} \tag{1}$$

This generates the first pyramid level. Next, the new level is sampled with bilinear interpolation at 1.5 spacing, on each direction.

Then, the next step is to find the stability points. This is achieved by extracting the gradients and orientation at each pyramid level.

In this work, the extraction steps are illustrated in Figure 4. Before extracting, recordings are converted from MP3 to WAV format with 16bit resolution and sampled to 16kHz. This processes was performed by the ffmpeg plug-in. To obtain the narrowband 40dB log-spectrograms signals are pre-emphasized by a 0.97 ratio. Then, we used 60ms window (960 samples) and 54ms (864 samples) overlapping. Furthermore, a Hamming Window is used as smoothing function, with a 1024 points NFFT.

Then, resulting spectrograms are converted to grey scale before extracting the SIFTs. In our experiments, the number of descriptors are limited to at most 100. Other configurations are the same as in [].

## 2.3 Bag of Visual Words

Another possible representation is the Bag of Visual Words (BOVW) []. This is an static feature representation that is automatic generated from K-Means. Since SIFT will output $N \times 128$, for $N$ descriptors, feature vectors for each spectrum, this requires a distinct approach to training some classifiers. Thus, with the descriptors, a $M$ cluster K-Means is trained with all the SIFTs. Then, each $m$-th feature vector is predicted as an cluster. Then, the $m$-th position of a vector $\boldsymbol{b}$ is incremented by 1. That is, the BOVW is the frequency of cluster occurrence in a set of feature vectors (in this work, SIFT).

$$\boldsymbol{b}_m = count(C_m, S_i) \tag{2}$$

$$\boldsymbol{b} = (b_1, \ldots, b_m, \ldots, b_M) \tag{3}$$

This approach allow us to not only reduce the horizontal dimension, but also reduces the representation of spectrograms from $N$ to a single vector. However,

Figure 4: Flowchart with feature extraction steps.

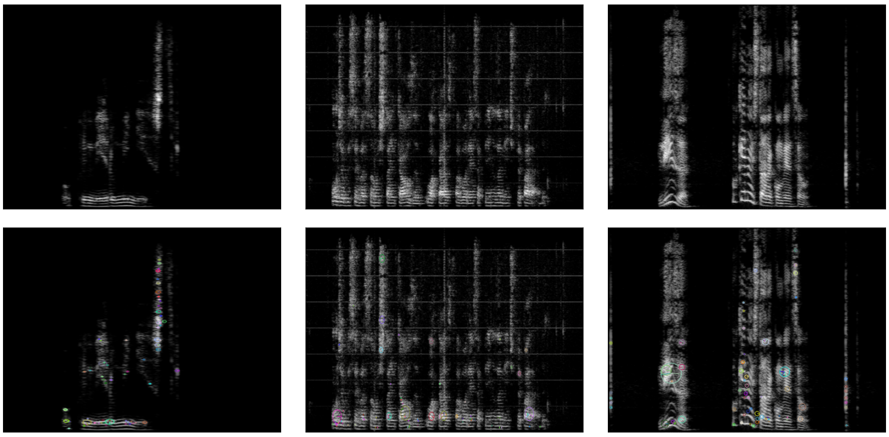


Figure 5: 40dB spectrum (upper row) and identified key points (bottom row).

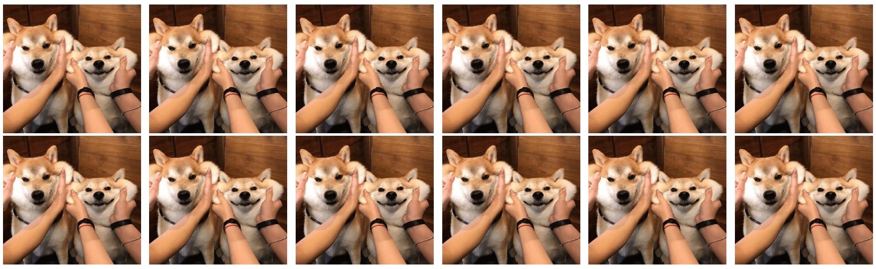


Figure 6: Spectrum (upper row) and SIFT (lower row) intra-class correlation.

Table 2: Summary of feature parameters and models for experiments

| Feature | Models | Dimension |
|---|---|---|
| SIFT | SVM | 100×128 |
| BOVW | SVM, KNN | 1×50<br>1 × 300 |

at the same time this also does not preserve time dependencies that could represent subtle variations in the signal. In our experiments, we used 50 clusters, and K-Means was initialized 10 times with random centre seeds. The best one is kept. Furthermore, the optimization stops when it reaches 300 iterations or a tolerance of 0.0001.

## 2.4 Configuring Models for Age Recognition

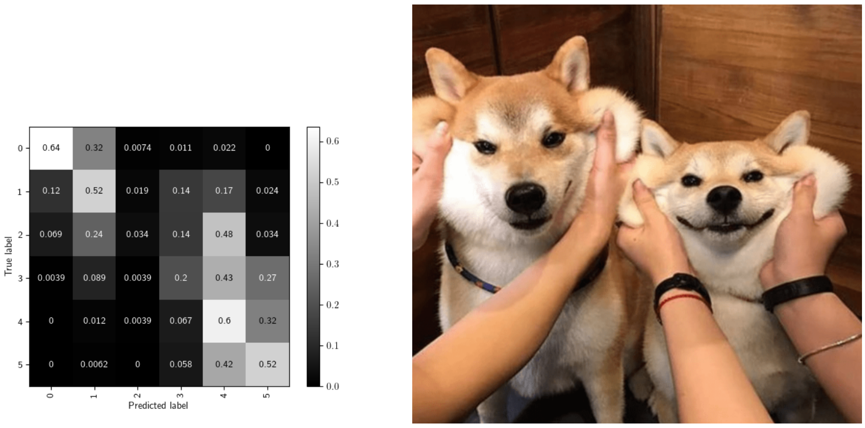
In this work, we propose to identify age through the image-related features SIFT, extracted from voice spectrograms. For purposes of investigation, after splitting the data into 7/3 ratio, both splits are normalized and the training subset is used for fine tuning. A 5-fold cross validation is applied while ensuring that every age class is present on the subsets. We choose a 5-fold because of the small amount of data for older speakers, which can make it difficult for training purposes if a higher number of folds were used.

For our models, we use a SVM and a KNN for age classification. All parameters are adjusted through a grid search with the cross validation. Following we describe the parameters ranges used for each model.

**SVM** uses Linear, RBF, and Sigmoid kernel. The penalisation parameter $C$ ranges in $\{0.001,\ 0.01,\ 0.1\}$, as well as the $\gamma$. In all experiments, the multi class strategy is fixed as One-vs-Rest. No adjustments are made on residuals. For the Polynomial kernel, we adopted the same ranges, besides the degrees 2 and 3.

**KNN** vary $k \in \{2, 4, 8, 16, 32, 64\}$, while the similarities are cosine, Euclidean, and Minkowski. The data is structured as a KD-Tree through all experiments, as well as uniform weighting.

Given the distributions of our data, using Accuracy (ACC) to evaluate our models would not be fair. Even though, we provide the ACC for matter of comparison with other works, as well as the Weighted Accuracy (WACC) for a better description of the performance. Furthermore, we also provide a confusion matrix to enable comparison with other metrics from the literature and future works. We train and evaluate both models with the BOVW, whereas SIFT is used only with SVM. The strategy used to classify the age from the sequence of descriptors is to assign the most frequent age category among the predictions of each descriptor. Table 2 summarises the features parameters and the respective models used in our experiments. Finally, experiments were executed in an AMD Ryzen 5 1600, 16 GiB memory (2×8GiB 2400MHz), and a 480GB Kingston SA400S3 as storage.



(a) KNN.

(b) Test.

Figure 7: SVM and KNN confusion matrices from test using Bag of Visual Words.

Table 3: Training and test WACC and ACC. Training values are mean ± standard deviation.

| Model | WA (%) | | ACC (%) |
|---|---|---|---|
| | Training | Test | |
| SVM | 90 ± 3 | 88 | 88 |
| KNN | 41.63 ± 1.58 | 44.39 | 52.9 |

# 3 Results

This section presents the experimental results using the above methodology. Following sections provide results on both SIFT and BOVW, in this order. Here, we mainly compare the results to a random classifier (for ACC) which would achieve $^1/_6 = 16\%$ accuracy. Note that accuracy is provided only for testing results.

## 3.1 SIFT

## 3.2 BOVW

# 4 Discussion

# 5 Conclusions and Final Remarks

(a) Training.

(b) Test.

Figure 8: Confusion matrices for KNN from training and test using Bag of Visual Words.

Figure 9: Comparison between accuracy and efficiency.