Speaker Gender Identification Using Pitch-based Features Extraction

ROTEM BAR, URIA LEVKOVICH
School of Software Engineering: Intelligent Systems, Afeka College of Engineering, Tel Aviv-Yafo, Israel.
Rotem Bar: (rotem.bar@s.afeka.ac.il)
Uria Levkovich (uria.levkovich@s.afeka.ac.il)

ABSTRACT Gender recognition is an essential component of automatic speech recognition and interactive voice response systems. Typical methods for gender recognition from speech greatly depend on feature extraction and classification processes. In this study, we will evaluate the performance of various machine learning classification models employed on instances that have been performed mel-frequency cepstrum coefficients (MFCC) feature extraction for gender recognition tasks. Five classification models, including k-nearest neighbor (KNN), naïve Bayes, multi-layer perceptron, random forest, and support vector machine (SVM), are examined in this study. Addiontlty, dimension reduction on the MFCC features extracted using principal component analysis (PCA) was examined. Results showed random forest superior ability with an accuracy of 88.42% to successfully identify the speaker's gender

I. INTRODUCTION

The human voice provides the semantics of the spoken words and contains information about speaker-dependent characteristics. Such speaker-dependent information may include speaker identity[1], gender[2], age[3], and emotional state[4]. Common approaches for gender recognition are based on the analysis of the pitch of the speech. To capture differences in both the time domain and frequency domain, a set of features known as MFCC are used [5]. These are widely used features for automatic speech and speaker recognition. MFCC features are extracted from speech signals over a small window. These features are also known to work efficiently in noisy environments[6], and due to their robust nature, they are widely used in speaker recognition tasks such as gender identefication[7]. MFCC features from the training data are used to train the supervised classifiers. The classification models are then used to predict the gender of the test instances.

In this work, we have used MFCC features performed on the well-known LibriSpeech dataset and evaluated several classification methods comprehensively to determine the best setup of the classification model for gender recognition. In addition, we will examine the effect of dimensionality reduction performance by the PCA[9] algorithm on the model performance.

II. MFCC

The MFCC feature extraction technique includes windowing the signal, applying the DFT[10], taking the Log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT. A detailed description of various steps involved in the MFCC feature extraction is explained below in figure 1.

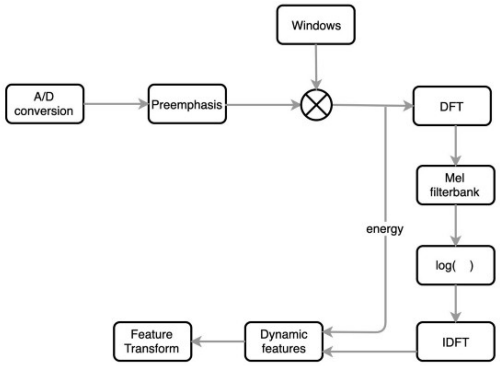

FIGURE 1. MFCC feature extraction pipeline

We will look into each step-by-step.

1. A/D Conversion: converting the audio signal from analog to digital format with a sampling frequency which is 16kHz, in our study.

2. Preemphasis: increasing the magnitude of energy in the higher frequency.

3. Windowing: breaking the audio signal into different windows. From each window, we extract 39 features.

4. DFT(Discrete Fourier Transform): convert the signal from the time domain to the frequency domain by applying the DFT transform. Analyzing in the frequency domain is easier than in the time domain for audio signals.

5. Mel-Filter Bank: applying triangular filter bank at lower frequencies and less discriminative at higher frequencies.

6. Applying Log: applying a log function on the previous model-filter bank output.

7. IDFT: applying inverse transform of the output from the previous step. The MFCC model takes the first 12 coefficients of the signal after applying the IDFT operations. Along with

the 12 coefficients, it will take the energy of the signal sample as the feature.

8. Dynamic Features: Along with these 13 features, the MFCC technique will consider the first-order derivative and second-order derivatives of the features, which constitute another 26 features. Derivatives are calculated by taking the difference of these coefficients between the audio signal samples. So overall MFCC technique will generate 39 features from each audio signal sample used as input for our gender speaker identification.

III. EXPERIMENTAL RESULTS

A. DATA DESCRIPTION

LibriSpeech[8] is a common and popular dataset for speech recognition system experiments. It includes transcribed ~960 hours of public domain audiobooks which many speakers dictate. Totally, for train purposes, the male sample of a LibriSpeech provides about 496 hours, and the female is a bit above 465 hours. The validation part is about 8 hours for males and females. Due to computation power limitations, we used only a small part of the complete data set in our study. The following Table 1 describes the data we used in our experiments.

TABEL 1. LibriSpeech dataset description used in the study

	Male	Female	Total
Speakers	20	20	40
Train Samples	1124	1052	2,176
Test Samples	250	277	527

B. EVALUATION METHOD

Three metrics are commonly used to evaluate balanced binary classification: accuracy (ACC), true positive rate (TPR), and false-positive rate (FPR). Their definitions are as follows.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$TPR = \frac{TP}{TP + FN} \quad (2)$$

$$FPR = \frac{FP}{TP + TN} \quad (3)$$

True positive (TP) is the number of instances classified correctly as Male. True negative (TN) is the number of instances classified correctly as female. False-positive (FP) is the number of instances classified incorrectly as Male, and false-negative (FN) is the number of instances classified incorrectly as female. As such, the accuracy measures the efficiency of the model, the TPR measures the portion of instances that are correctly classified, and the FPR indicates the false-alarm rate.

C. DATA EXPLORATION AND PREPROCESSING

As part of the experiment phase, we would like to perform the following preprocessing steps:

1. Maintain the balance between males and females by using stratified sampling. Since our target of the gender identification task is to determine whether the speaker is a male or female, we want to preserve a balanced amount of samples from both classes also at the training and testing steps. Maintaining this balance will reduce biased results for one of the two classes, and in addition, balanced data is more straightforward for analysis with the three described matrics Accuracy, TPR, and FPR.

2. Independent samples – In our data, each speaker has several audio records. If we do not keep those records separate between the training and the testing steps, it will create a data leakage problem. Data leakage occurs when the model is trained, and the samples in the test group are seen. A model that has a data leakage problem will yield higher results than the actual one.

3. Feature extraction with MFCC- As described in the previous section, MFCC will extract 39 features. Those transformed samples will be fed to the different machine learning classifiers.

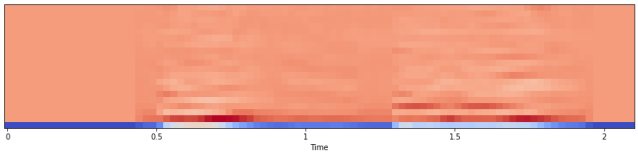

FIGURE 2. MFCC feature extraction performed on one sample from the LibriSpeech dataset

D. EXPERIMENT RESULTS

To evaluate the model performance, we will perform two experiments with a different purposes. Each experiment was performed on the data after the preprocessing steps described in the previous section.

Experiment I: in this experiment, we will train and test five different machine learning models: random forest, SVM, naïve base,multi-layer perceptron (MLP), and k-nearest neighbors (KNN). This experiment will indicate which of the five models will suit the best for the task of gender detection where the input is the features vector produced by the MFCC.

Experiment II: In this experiment, we want to examine the effect of dimensionality reduction on the model performance. As described, the MFCC yields 39 features. Using PCA, we will reduce the feature dimensionality from 38 to 2. We will use the best-performed model (random forest) from the experiment I for the classifier. This experiment will indicate if

reducing the dimensionality from the 39 features MFCC will also reduce the model performance, and if so, how much.

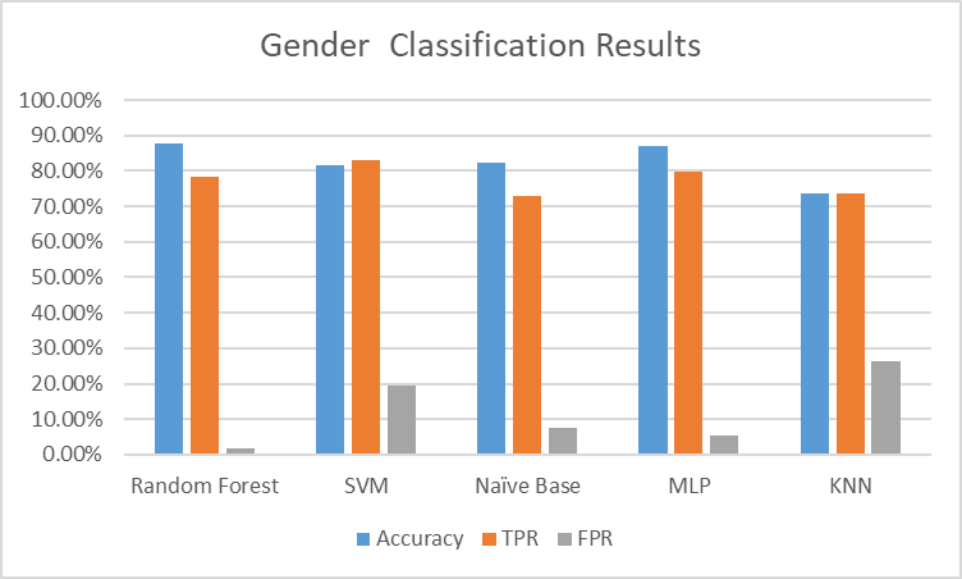

FIGURE 3. Gender classification results from the five machine learning classifiers

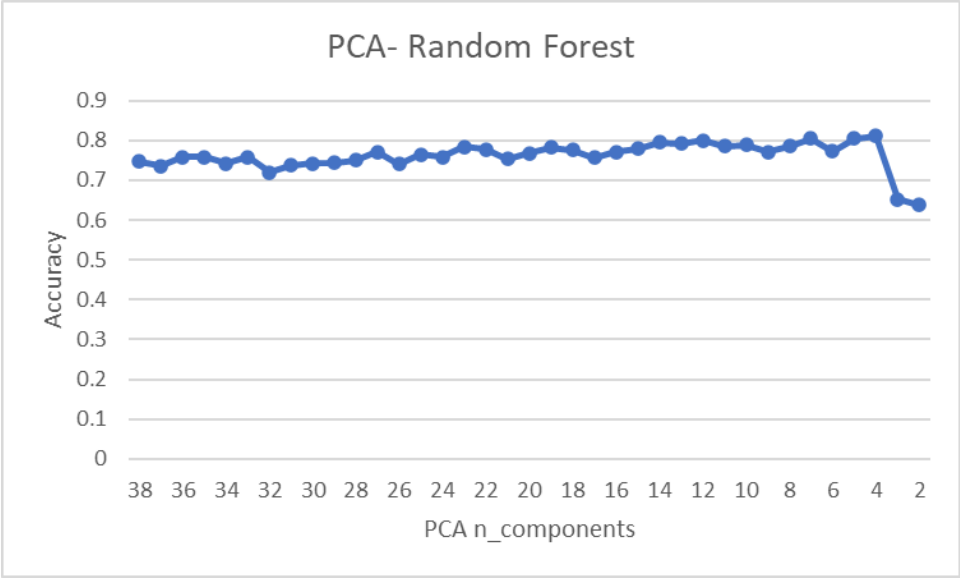

FIGURE 3. Gender classification accuracy results with a random forest classifier, where the x-axis represents the number of components reduced by PCA

From figure 3, we see that random forest performed the best among the five classifiers with an accuracy of 88.42% and with the lowest FPR of 1.21%. FPR indicates cases where the model classified incorrectly as positive, which is males in our case. The model that performed the poorest is KNN, with an accuracy of 73.62% and an FPR of 26.41%.

From figure 4, which describes the random forest performance while the dimensional input reduced from 38 to 2, we see that up to 5 components of the model maintain high results of over 80% accuracy; even more, we did not see performance decreasing up to 5 component. The lowest performance occurred with two components with an accuracy of 61.1%.

VI. CONCLUSIONS

This study examined the gender identification task by using the MFCC feature extraction. We performed two experiments conducted on the well-known dataset LibriSpeech. For the first experiment, we examined five state-of-the-art machine learning models; this experiment showed the superior ability of the random forest model, which successfully classified 88.42% of the samples. The second experiment was performed to examine how dimension reduction will affect the model performance; we chose the random forest model for this experiment. The dimension reduction was performed by the commonly used PCA algorithm that reduced from 39 to 2 features. The second experiment showed that the reduction of up to five features did not significantly decrease the accuracy results. Certain applications that prefer low competition processes required over model performance can use this dimension reduction method.

REFERENCES

[1] Reynolds, D.A., 1995. Automatic speaker recognition using Gaussian mixture speaker models. In The Lincoln Laboratory Journal.

[2] Doukhan, D., Carrive, J., Vallet, F., Larcher, A. and Meignier, S., 2018, April. An open-source speaker gender detection framework for monitoring gender equality. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5214-5218). IEEE.

[3] Safavi, S., Russell, M. and Jančovič, P., 2018. Automatic speaker, age-group and gender identification from children's speech. Computer Speech & Language, 50, pp.141-156.

[4] Swain, M., Routray, A. and Kabisatpathy, P., 2018. Databases, features and classifiers for speech emotion recognition: a review. International Journal of Speech Technology, 21(1), pp.93-120.

[5] Hossan, M.A., Memon, S. and Gregory, M.A., 2010, December. A novel approach for MFCC feature extraction. In 2010 4th International Conference on Signal Processing and Communication Systems (pp. 1-5). IEEE.

[6] Kotnik, B., Vlaj, D., Kacic, Z. and Horvat, B., 2002, September. Robust MFCC feature extraction algorithm using efficient additive and convolutional noise reduction procedures. In ICSLP (Vol. 2, pp. 445-448).

[7] Ahmad, J., Fiaz, M., Kwon, S.I., Sodanil, M., Vo, B. and Baik, S.W., 2016. Gender identification using mfcc for telephone applications-a comparative study. arXiv preprint arXiv:1601.01577

[8] Panayotov, V., Chen, G., Povey, D. and Khudanpur, S., 2015, April. Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5206-5210). IEEE.

[9] Abdi, H. and Williams, L.J., 2010. Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), pp.433-459.

[10] Jacobsen, E. and Lyons, R., 2003. The sliding DFT. IEEE Signal Processing Magazine, 20(2), pp.74-80.