# Gender Identification from Arabic Speech Using Machine Learning

Skander Hamdi[1], Abdelouahab Moussaoui, Mourad Oussalah[2], and Mohamed Saidi[1]

[1] Department of Computer Science, University of Ferhat Abbes Setif I, Setif, Algeria
{skander.hamdi,abdelouahab.moussaoui,mohamed.saidi}@univ-setif.dz
[2] Department of Computer Science and Engineering, University of Oulu, Oulu, Finland
mourad.oussalah@oulu.fi

**Abstract.** Speech recognition is becoming increasingly used in real-world applications. One of the interesting applications is automatic gender recognition which aims to recognize male and female voices from short speech samples. This can be useful in applications such as automatic dialogue systems, system verification, prediction of demographic attributes (e.g., age, location) and estimating person's emotional state. This paper focuses on gender identification from the publicly available dataset Arabic Natural Audio Dataset (ANAD) using an ensemble-classifier based approach. More specifically, initially we extended the original ANAD to include a gender label information through a manual annotation task. Next, in order to optimize the feature engineering process, a three stage machine learning approach is devised. In the first phase, re restricted to features to the two widely used ones; namely, MFCC and fundamental frequency coefficients. In the second phase, six distinct acoustic features were employed. Finally, in the third phase, the features were selected according to their associated weights in Random Forest Classifier, and the best features are thereby selected. The latter approach enabled us to achieve a classification rate of 96.02% on the test set generated with linear SVM classifier.

**Keywords:** Speech features · Feature selection · Arabic speech · Gender identification · Machine learning · Deep learning

## 1 Introduction

Speech recognition is the process by which a computer system maps an acoustic speech signal to some form of abstract meaning of speech using a set of predefined acoustic features. Speech recognition is becoming increasingly used in real-world applications such as speaker identification for securing access to confidential information or virtual spaces, automatic reading of text or dictionaries, recognizing and understanding of speech for control applications. Therefore, speech

recognition has become a growing field of research in data science and machine learning. Gender identification is another application of voice recognition. The differences in a human speech concerning gender are basically due to physiological characteristics: vocal fold thickness, vocal tract length. We can observe these differences in the speech signal and also use them to make a gender-based classification model that distinguishes between male and female from a short voice recording.

In this paper, we will build a gender classification speech-based model using a modified version of an arabic speech database where utterances are pronounced by only Arabic native speakers. Arabic language was chosen for its special characteristics compared to the Latin languages used in all related works [25]. Low levels descriptors have been extracted during the pre-processing phase. Our methodology uses a set of machine learning classifiers, including KNN, Naives Bayes, Decision Tree, Logistic Regression, Random Forest, Linear SVM, Kernel SVM, ANN, CNN, which enable comprehensive comparison. Furthermore, in order to investigate the feature engineering process, a three step-strategy has been developed where the feature set is selected according to the two most widely used features, all potential acoustic features, selected according to their weights in Random Forest classifier. We achieved 96.02% of test accuracy using Linear SVM. The second section of this paper will give a brief introduction to the used speech features while the third section will give a review of related works in gender identification. The fourth section will present the used database. The Methodology, experimental results are highlighted in the fifth and sixth section, respectively. Finally, discussion and conclusions are drawn in the emphasizes the methodology, the sixth will present and discuss the experiments results and the last one conclude the work with some perspectives.

## 2 Acoustic Features

Speech recognition tasks involve extracting a set of speech features from the speech dataset. One distinguishes three groups of features: spectral, excitation and acoustic features [2]. Spectral and acoustic features are the most relevant to our work.

### 2.1 Intensity

Sound Intensity models the loudness of the sound signal [1,2]. It is known as the power of the sound waves in an area unit in a perpendicular direction. Measured by watt per square meter ($W/m^2$). Sound intensity can be calculated with the following equation [15]:

$$I = \frac{P}{A} \tag{1}$$

where $P$ is the power of sound and $A$ is a distance.

In order to measure the sound intensity relatively to a reference, we can calculate the **acoustic intensity level** with the following formula [16]:

$$L = 10 \times \log_{10} \frac{I}{I_0}(dB) \tag{2}$$

where $I$ is the sound intensity and $I_0$ is a reference value. Sound intensity level is expressed in decibels (dB).

### 2.2 Zero-Crossing Rate

Zero-crossing rate is the rate at which a signal changes its sign during the frame (measure the duration) [1,2]. Formally defined as following [17]:

$$zcr = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{I}\{s_t s_{t-1} < 0\} \tag{3}$$

### 2.3 Fundamental Frequency

Fundamental frequency (F0) measures the pitch. It is defined as the vibration frequency of vocal chords. Defined as the lowest frequency of a periodic waveform [18]. F0 can be extracted using different methods such as Pitch Detection Algorithms [19,20].

### 2.4 Mel Frequency Cepstral Coefficients (MFCC)

MFCC takes into account human perception for sensitivity at appropriate frequencies by converting the conventional frequency to Mel Scale (4). To calculate MFCCs, some steps should be followed [21]:

1. Calculating the Fourier transform of a signal.
2. Mapping the result of the previous step to the mel scale (4) using triangular overlapping windows.
3. Calculating the logarithm of the powers at each of the calculated mel frequencies.
4. Calculating the discrete cosine transform of the mel logarithm powers.
5. MFCCs are the amplitudes of the resulting spectrum.

$$mel = 2595 \times \log_{10} \frac{1 + hertz}{700} \tag{4}$$

### 2.5 Probability of Voicing

It evaluates the probability that the voice is *voiced* or *unvoiced*. The voiced type is characterized by the periodic structure when the unvoiced presents basically a noise. Some methods to estimate the probability of voicing were discussed in [22].

### 2.6 Line Spectral Frequency (LSP)

Line Spectral Pairs (LSP) are popular alternative representation of Linear Prediction Coefficients (LPC). LSPs are useful for speech coding as they have some properties that make them superior to direct quantization of LPCs [23].

## 3 Related Works

Several works have been published in order to identify gender from speech using machine learning and deep learning approaches. Most of works we surveyed used MFCC and F0 as features and different classification methods. For instance Rami and Alkhawaldeh [3] used *Mel-spectrogram*, *MFCCs*, *Chroma-STFT*, *Spectral_Contrast* and *Tonnetz* features for a gender recognition task through a neural network based approach. For the learning and classification methods, they provide several experiments using machine learning techniques family: Bayesian Network, Naive Bayes, MLP (Multi-Layer Perceptron), Logistic Regression, SMO (sequential minimal optimization for SVM), Linear SVM, Kernel SVM (Polynomial, Radial), Latent Dirichlet allocation (LDA), Lazy learners (IBk and KStar), AdaBoost, Decision trees, random forest and three rule-based methods (OneR, Ridor and Rough Set). For deep learning approaches, they used 1D Convolutional neural network architecture composed of three convolution layers with 32, 48,120 neurons with ReLu activation function, one max-pooling layer and two fully connected layers with 128 and 64 units and an output layer with softmax activation function with two units according to class labels *male,female*. They used also some feature selection techniques to choose the relevant ones, *Evolutionary search*, *PSO search* and *Wolf search* were used. The ROC curve (AUC) was used to verify if a classifier can separate the two classes. Using a dataset of artificial voices consisting of 20 languages, each one composed of 16 voice samples (8 for male, 8 for female), they obtained the following results: as precision/recall metric, Kernel SVM (Polynomial) yields 100% while Logistic Regression yields 99.7% and 99.4% for Random Forest method.

Kabil et al. [5] proposed a deep learning approach using two datasets called AVspoof and ASVspoof 2015, different convolutional neural network (CNN) architectures have been proposed by giving the raw speech signal to the network's input to make automatically the feature learning phase by the convolution layers. To make comparisons, two artificial neural network (ANN1 + ANN2) architectures have been proposed after acoustic features extraction. The first one with MFCC feature 342 as input dimension and the second with MFCC+F0 and 351 as input dimension. Three CNN architectures with different hyper-parameters have been tested, the last one (cnn3) outperforms cnn1,cnn2, and the first baseline proposed system MFCC+F0 based. After CNN analysis, it has been shown that CNN learn formant and fundamental frequency.

Doukhan et al. [6] focused on the French corpus called REPERE where raw audio streams were used. The authors compared Gaussian Mixture Models (GMM), i-vector, and CNNs models for automatic gender identification task. In [7], a CNN based speech segmentation was used to eliminate music and empty

speech segments, low energy frames using fixed thresholding-based approach. Next, for the feature extraction step, the authors used SIDEKIT. As result, the CNN model gave the best Recall measure, 98.04% for male, 95.05% for female, and 96.52% for F1 measure.

Levitan et al. [9] used pitch and spectral features by comparing each one to others. To add more information to the features; minimum, maximum, median, mean, and standard deviation have been calculated for each value of f0 trajectory, and 21 MFCCs used as features. A database called HMIHY composed of 5002 from 1654 speakers is used with four classifiers: Logistic Regression, linear regression, random forest, AdaBoost. Logistic regression model gave the high classification rate of 95.2% using the fundamental frequency and MFCCs. In another experiment, a benchmark called aGender for german speech, has three class labels: *male, female* and *children* and to compare the previous results with german speech data classification, f0 statistics have been supplemented with another statistics (f0+). Using all features (f0, f0+, and MFCCs), random forest classifier gave the best classification rate of 85.0% close to the best achieved results in the challenge [1]. Compared to the previous results, this degradation is due to the existence of a third class and the challenge *children vs female* because of the high pitched speech in both classes which makes the problem more complicated. To validate and ensure the model performance, a cross-lingual gender detection has been presented by training different learners on the english corpus HMIHY and test the model with non-english data (aGender by eliminating children samples). The cross-lingual model with Logistic Regression learner gave 92.1% while german-only model gave 93.0% using random forest classifier.

Ioannis et al. [4] proposed a new ensemble semi-supervised self-labeled algorithm to build a more accurate gender classifier by exploiting the problem of the non-existence of sufficient data to build an efficient model. They proposed an algorithm called *iCST-Voting* which combines the predictions of *Co-training*, *Self-training*, and *Tri-training* using an ensemble as base learning. The main methodology of the proposed framework is composed of two steps: **Training** step where three classifiers are trained using *Co-training* $C_{co}$, *Self-training* $C_{self}$ and *Tri-training* $C_{Tri}$ algorithms. **Voting** step where each trained classifier ($C_{co}$,$C_{self}$ and $C_{Tri}$) is applied to every unlabeled sample on the test set, the final label will be the majority voting. Two datasets were used, Voice gender dataset and Deterding dataset, pre-processed by acoustic analysis using **seewave** and **tuneR** libraries in **R**. As classifiers: SMO, decision trees C4.5, and kNearest Neighbor were used. Comparing to the state-of-the-art, iCST-Voting outperforms all previous work in self-labeled algorithms for both datasets.

Kaushik et al. [13] used a database of 200 voice samples for celebrities 50% for males and 50% for females. They proposed a combination of two methods, called combo-classifier which use the Modified autocorrelation method and average magnitude difference function by weighting each of them. Combo-classifier achieved 99%.

Ali [14] proposed a system with front-end and back-end, the first one extracts useful feature using First Fourier Transform (FTT) algorithm and the second as

a classifier which classifies each speech signal ("A" or "B") as male/female. Classification was based on the frequency at maximum power which was extracted from estimated power spectrum and gave an average recognition accuracy of 80%.

In [8], a feed-forward multi-layer perceptron FF-MLP has been proposed after feature extraction and speech analysis pipeline. Also, different order of Linear Prediction Coefficients (LPC) filter is applied using autocorrelation function (ACF) to produce LPC coefficients where they are used as neural network input. The training was done using Levenberg-Marquardt learning algorithm. Different LPC orders have been tested from 8 to 20 and the LPC-18 gave the best rate of 95.5%.

There is no paper yet that proposes arabic data to build a gender recognition model speech-based, in this paper, we will propose a model based on machine learning and deep learning techniques using the pitch and spectral features discussed in the state-of-the-art. We will use a modified version of an arabic speech database used for emotions recognition.

## 4 Arabic Natural Audio Dataset (ANAD)

In the following section, we will present the used database, our modification which allows us to use it for the aim of building a gender recognition speech-based model.

### 4.1 Dataset Description

Arabic Natural Audio Dataset (ANAD) is used in [1,11] for emotion recognition, the dataset is available online in Kaggle[1]. Live calls between an anchor and a human outside the studio were downloaded from online Arabic talk shows from different dialects (Egyptian, Gulf, Jordan and Lebanese). Each video was divided into turns: callers and receivers. To label each video, 18 listeners were asked to listen to each video and select an emotion: happy, angry, or surprised. Silence. Laughs and noisy chunks were removed. Every chunk was divided into 1 s speech units forming the final corpus composed of 1383 records.

### 4.2 Modified Version

To build a gender recognition model using this database, the label *male, female* of each record was manually added by giving all the voice recordings to 3 listeners, each one of them listen and note the corresponding class label to each voice recording, in case of disagreement, we took the majority vote as the correct class. The following table presents the gender label distribution statistics.

[1] https://www.kaggle.com/suso172/arabic-natural-audio-dataset.

**Table 1.** Gender class label in the new ANAD.

| Class label | Samples count |
|---|---|
| Male | 736 |
| Female | 647 |

## 5 Methodology

The dataset has been pre-processed to build an emotion recognition model in [1] by extracting pitch and spectral features or what is called low-level descriptors [12]. These features are: fundamental frequency f0, f0 envelope, intensity, zero crossing rates, 12 MFCCs, 7 LSP frequencies, probability of voicing. Some statistics are calculated for each one of these descriptors: maximum, minimum, range, absolute position of maximum, absolute position of minimum, arithmetic of mean, Linear Regression1, Linear Regression2, Linear RegressionA, Linear RegressionQ, Standard Deviation, kurtosis, skewness, quartiles 1, 2, 3 and, interquartile ranges 1–2, 2–3, 1–3. These features have been used in three stages. In the first stage, we used only MFCCs, F0 features as reported in [5,9], while all features have been employed in the second stage. In order to improve the results of the two stages, we used the feature weighted provided by the Random Forest algorithm to select the best and important features by measuring the impurity (*Gini impurity* in our case). The more the impurity decreases, the more important the feature is [24].

Different classifiers have been proposed to build a gender classification model and used for the three stages, as machine learning methods family: **k-Nearest Neighbor**, **Naive Bayesian**, **Decision Trees C4.5**, **Logistic Regression**, Ensemble learning using **Random Forest** with hyper-parameter search and 5-fold cross-validation to find the best parameter for the algorithm, **Linear SVM** and **Kernel SVM** with hyper-parameter search from python scikit-learn library.

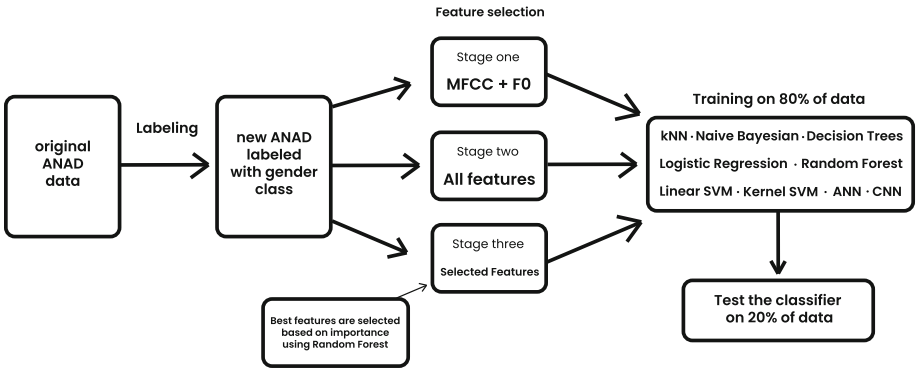


**Fig. 1.** Summary of proposed methodology

For deep learning approaches family, using Keras and Tensorflow, we proposed Artificial Neural Network (ANN) and Convolutional Neural Network (CNN) architectures based on many experiments. The dataset is divided into 80% of training and the remaining 20% for test. The Fig. 1 presents our methodology and the Table 2 shows a summary of the used classifiers for our problem.

## 6 Experiments Results

Training was done on a local PC with the following configuration: MacOS Catalina 10.15.4 16 GB of memory 2.3 GHz Quad-Core Intel Core i7 as CPU

**Table 2.** Proposed methods for building gender identification from arabic speechProposed methods for building gender identification from arabic speech

| Method | Parameters |
|---|---|
| k-Nearest Neighbor | k = 3 |
| Naive Bayesian | default scikit-learn parameters |
| Decision Trees C4.5 | default scikit-learn parameters |
| Logistic Regression | default scikit-learn parameters |
| Random Forest | n_estimators = 800 |
| | min_samples_split = 2 |
| | min_samples_leaf = 1 |
| | max_depth = 100 |
| Linear SVM | kernel = linear |
| Kernel SVM | kernel = rbf |
| | C = 1 |
| | gamma = 0.05 |
| ANN architecture | Dense(units = 32, activation=relu) + Dropout(0.2) |
| | Dense(units = 16, activation = relu) + Dropout(0.1) |
| | Dense(units = 1, activation = sigmoid) |
| | Number of epochs = 70 |
| | Batch size = 64 |
| | optimizer = adam |
| | loss = binary_crossentropy |
| CNN architecture | Conv1D(filters = 16, filter_size = 2, activation = relu) |
| | Conv1D(filters = 16, filter_size = 2, activation = relu) + Dropout(0.1) |
| | MaxPooling1D(window_size = 2) |
| | Conv1D(filters = 8, filter_size = 2, activation = relu) |
| | Conv1D(filters = 16, filter_size = 2, activation = relu) + Dropout(0.1) |
| | Dense(units = 50, activation = relu) |
| | Dense(units = 1, activation = sigmoid) |
| | Number of epochs = 70 |
| | Batch size = 64 |
| | optimizer = adam |
| | loss = binary_crossentropy |

and NVIDIA GeForce GT 750M 2 GB as GPU and after several experiments in the three stages for each one of algorithms, we obtained the following results:

### 6.1 Stage One

The result shows that CNN classifier presents the the worst test accuracy of 59.92%, while its training accuracy is 95.11%, which indicates that the proposed CNN architecture cannot accommodate perfectly the problem under consideration, while the highest accuracy 93.86% is achieved using Logistic Regression and Linear SVM with a training accuracy of 93.85% and 97.46%, respectively.

**Table 3.** Experiments results using proposed methodology for the first stage

| Algorithm | Train Acc. [%] | Test Acc. [%] |
|---|---|---|
| k-Nearest Neighbor | 92.04 | 84.83 |
| Naive Bayesian | 82.73 | 81.94 |
| Decision Trees C4.5 | 100 | 80.86 |
| Logistic Regression | 93.85 | **93.86** |
| Random Forest | 100 | 93.14 |
| Linear SVM | 97.46 | **93.86** |
| Kernel SVM | 93.58 | 92.41 |
| ANN | 92.67 | 92.05 |
| CNN | 95.11 | 59.92 |


**Fig. 2.** Summary of achieved results in the first stage which show that Logistic Regression and Linear SVM did better than other methods

The Table 3 shows the achieved results using all methods when only MFCCs and F0 features are used as input (434 features in total).

### 6.2 Stage Two

In order to improve the results of the first stage, we tried to use all remaining features with their calculated statistics: LSP frequencies, f0 envelope, intensity, zero crossing rates, probability of voicing. All classification rates have been improved in the second stage. Naive bayesian classifier achieves the worst test accuracy of 86.28% with 84.62% training accuracy, while the best test classification rate of 95.30% is achieved using Logistic Regression where, out of a total of 277 test samples, 119 were correctly classified as female and 145 as male, and only 13 misclassifications. This is followed by Kernel SVM that achieved 94.94% test accuracy, corresponding to 115 correctly classified female, 148 correctly classified male and 14 misclassifications. Although there is a significant improvement

**Table 4.** Experiments results using proposed methodology for the second stage

| Algorithm | Train Acc. [%] | Test Acc. [%] |
|---|---|---|
| k-Nearest Neighbor | 94.21 | 91.69 |
| Naive Bayesian | 84.62 | 86.28 |
| Decision Trees C4.5 | 100 | 88.08 |
| Logistic Regression | 97.46 | **95.30** |
| Random Forest | 100 | 94.58 |
| Linear SVM | 99.81 | 93.86 |
| Kernel SVM | 96.20 | 94.94 |
| ANN | 92.22 | 94.22 |
| CNN | 96.38 | 91.69 |

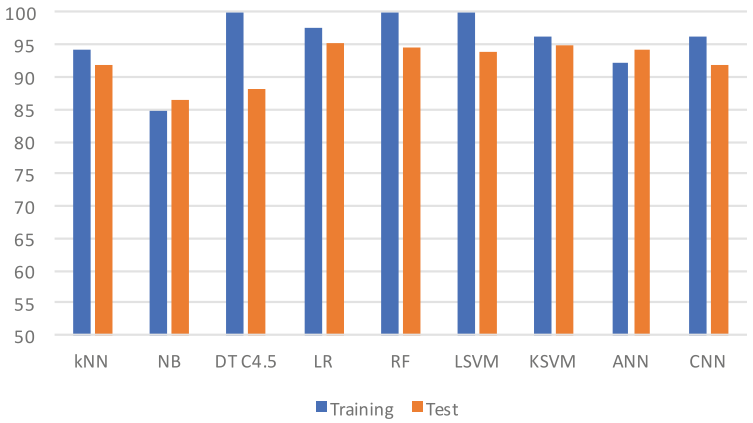


**Fig. 3.** Summary of obtained results in the second stage which show that Logistic Regression outperform the other methods followed by random forest

in the CNN result which means that the previous model is more complex than that of Stage one. However, the additional features added more informative patterns to the data. Table 4 shows the achieved results using all methods when all features are used (844 features in total).

### 6.3 Stage Three

A last experiment employs Random Forest method to extract the best and important features. Based on the same hyper-parameters found by grid search, we train again a random forest classifier. A set of 163 features has been found including MFCC, LSP frequencies, Zero Crossing Rates, F0, F0 envelope and Probability of voicing. The highest values of importance from the first to 43th

**Table 5.** Experiment results using proposed methodology for the third stage

| Algorithm | Train Acc. [%] | Test Acc. [%] |
|---|---|---|
| k-Nearest Neighbor | 95.11 | 94.22 |
| Naive Bayesian | 88.60 | 88.44 |
| Decision Trees C4.5 | 100 | 90.61 |
| Logistic Regression | 93.94 | 93.86 |
| Random Forest | 100 | 95.66 |
| Linear SVM | 96.02 | **96.02** |
| Kernel SVM | 93.67 | 94.94 |
| ANN | 94.12 | 93.14 |
| CNN | 95.56 | 88.44 |


**Fig. 4.** Summary of obtained results in the third stage which show that Linear SVM outperform the other methods followed by random forest

feature were taken by MFCC features which show the importance of MFCCs and validate the results of stage one. We used the extracted features to re-train and improve the classification rate. We achieved the best test accuracy of 96.02% using Linear SVM classifier on 277 test samples when 122 are correctly classified as female and 144 as male with 11 misclassifications followed by Random Forest with test accuracy of 95.66% where 116 are correctly classified as female and 149 as male with 12 misclassifications. Table 5 presents the experiment results for the third stage.

The following Table 6 and Fig. 5 summarize all obtained experiments results.

**Table 6.** Summary of test accuracy for each method in all stages

| Algorithm | Stage one [%] | Stage two [%] | Stage three [%] |
|---|---|---|---|
| k-Nearest Neighbor | 84.83 | 91.69 | 94.22 |
| Naive Bayesian | 81.94 | 86.28 | 88.44 |
| Decision Trees C4.5 | 80.86 | 88.08 | 90.61 |
| Logistic Regression | **93.86** | **95.30** | 93.86 |
| Random Forest | 93.14 | 94.58 | 95.66 |
| Linear SVM | **93.86** | 93.86 | **96.02** |
| Kernel SVM | 92.41 | 94.94 | 94.94 |
| ANN | 92.05 | 94.22 | 93.14 |
| CNN | 59.92 | 91.69 | 88.44 |

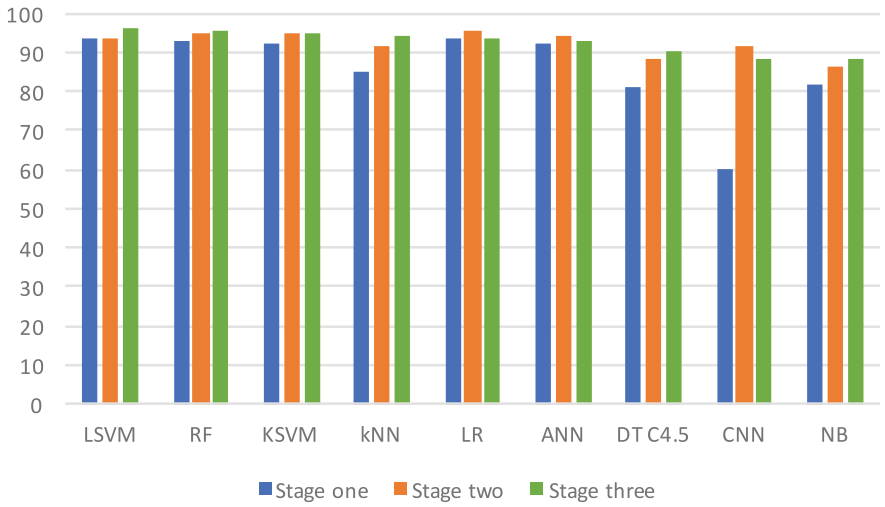


**Fig. 5.** Summary of test accuracy using all methods in the three stages which show the performance of Linear SVM after selecting important features

## 7 Conclusion and Future Work

There are a lot of researches that have used benchmarks in different languages such as English, French, German to build gender recognition models from natural speech recordings. In this paper, we used an arabic natural audio dataset with a comparison of different machine learning and deep learning approaches. In order to investigate the influence of the features on the overall approach, we devised a three stage strategy. The first stage uses only MFCCs and F0 as features. In the second stage, all features proposed in [1] were employed in each classifier. This includes acoustic intensity, which measures the power of the sound waves, Zero Crossing Rate that measures the sign signal change rate, among others. The third and last stage uses Random Forest for selecting the most important features in order to use only the discriminant ones for improving the classification rate. We have extracted 163 important features and we succeeded in constructing a model whose classification rate for the test set is 96.02% using Linear SVM which gave the best results for the first and last stages. We also comprehended the importance of speech features variety (pitch and spectral) for classification tasks or speech recognition to obtain good results. As future work, we will try to combine ANAD with another database which includes children's speech recordings and trying also to make a cross-lingual model trained on arabic data for non-arabic speech gender detection.

## References


1. Klaylat, S., Osman, Z., Hamandi, L., et al.: Emotion recognition in Arabic speech. Analog Integr. Circ. Sig. Process **96**, 337–351 (2018). https://doi.org/10.1007/s10470-018-1142-4
2. Klaylat, S., Osman, Z., Hamandi, L., et al.: Enhancement of an Arabic speech emotion recognition system. Int. J. Appl. Eng. Res. **13**(5), 2380–2389 (2018). ISSN 0973–4562
3. Rami, S., Alkhawaldeh, D.G.R.: Gender recognition of human speech using one-dimensional conventional neural network. Sci. Program. (2019). https://doi.org/10.1155/2019/7213717. ISSN: 1058–9244
4. Livieris, I., Pintelas, E., Pintelas, P.: Gender recognition by voice using an improved self-labeled algorithm. Mach. Learn. Knowl. Extr. **1**, 492–503 (2019). https://doi.org/10.3390/make1010030
5. Kabil, S., Muckenhirn, H., Magimai-Doss, M.: On learning to identify genders from raw speech signal using CNNs, pp. 287–291 (2018). https://doi.org/10.21437/Interspeech.2018-1240
6. Doukhan, D., Carrive, J., Vallet, F., Larcher, A., Meignier, S.: An open-source speaker gender detection framework for monitoring gender equality (2018). https://doi.org/10.1109/ICASSP.2018.8461471
7. Doukhan, D., Carrive, J.: Investigating the use of semi-supervised convolutional neural network models for speech/music classification and segmentation (2017)

8. Yusnita, M.A., Hafiz, A.M., Fadzilah, M.N., Zulhanip, A.Z., Idris, M.: Automatic gender recognition using linear prediction coefficients and artificial neural network on speech signal. In: 2017 7th IEEE International Conference on Control System, Computing and Engineering (ICCSCE) (2017). https://doi.org/10.1109/iccsce.2017.8284437
9. Levitan, S., Mishra, T., Bangalore, S.: Automatic identification of gender from speech, pp. 84–88 (2016). https://doi.org/10.21437/SpeechProsody.2016-18
10. Meinedo, H., Trancoso, I.: Age and gender classification using late fusion of acoustic and prosodic features. In: Proceedings of Interspeech 2010, Makuhari, Japan, pp. 2818–2821 (2010)
11. Klaylat, S., Osman, Z., Hamandi, L., et al.: Enhancement of an Arabic speech emotion recognition system. Int. J. Appl. Eng. Res. **13**(5), 2380–2389 (2018). ISSN 0973–4562
12. Low, L.A., Maddage, N.C., Lech, M., Sheeber, L., Allen, N.: Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, pp. 5154–5157 (2010). https://doi.org/10.1109/ICASSP.2010.5495018
13. Kaushik, D., Jain, N., Majumdar, A.: Gender Voice Recognition through speech analysis with higher accuracy (2014). https://doi.org/10.13140/2.1.1331.5842
14. Ali, M.: Gender recognition system using speech signal. Int. J. Comput. Sci. Eng. Inf. Technol. **2**, 1–9 (2012). https://doi.org/10.5121/ijcseit.2012.2101
15. Letter symbols to be used in electrical technology – Part 3: Logarithmic and related quantities, and their units, IEC 60027-3 Ed. 3.0, International Electrotechnical Commission, July 19, 2002
16. Fahy, F.: Sound Intensity. CRC Press, London (2017). ISBN 978-1138474192. OCLC 1008875245
17. Gouyon, F., Pachet, F., Delerue, O.: On the use of zero-crossing rate for an application of classification of percussive sounds (2002)
18. Fundamental frequency, Pitch, F0. In: Li, S.Z., Jain, A. (eds.) Encyclopedia of Biometrics. Springer, Boston (2009)
19. Tan, L., Karnjanadecha, M.: Pitch detection algorithm: autocorrelation method and AMDF (2003)
20. Drugman, T., Huybrechts, G., Klimkov, V., Moinet, A.: Traditional machine learning for pitch detection. IEEE Sig. Process. Lett. **PP**(99), 1 (2018). https://doi.org/10.1109/LSP.2018.2874155
21. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Commun. **54**(4), 543–565 (2012). https://doi.org/10.1016/j.specom.2011.11.004
22. Rehr, R., Krawczyk, M., Gerkmann, T.: A posteriori voiced/unvoiced probability estimation based on a sinusoidal model. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, pp. 6944–6948 (2014). https://doi.org/10.1109/ICASSP.2014.6854946.
23. Sahidullah, M., Chakroborty, S., Saha, G.: On the use of perceptual Line Spectral pairs Frequencies and higher-order residual moments for Speaker Identification. Int. J. Biometr. **2**, 358–378 (2010). https://doi.org/10.1504/IJBM.2010.035450
24. Sandri, M., Zuccolotto, P.: Variable selection using random forests (2006). https://doi.org/10.1007/3-540-35978-8_30
25. Alotaibi, Y., Meftah, A.: Review of distinctive phonetic features and the Arabic share in related modern research. Turk. J. Electr. Eng. Comput. Sci. **21**, 1426–1439 (2013). https://doi.org/10.3906/elk-1112-29