

Voice Gender Recognition Using Acoustic Features, MFCCs and SVM

Fadwa Abakarim(✉) and Abdenbi Abenaou

Research Team of Applied Mathematics and Intelligent Systems Engineering, National School of Applied Sciences, Ibn Zohr University, 80000 Agadir, Morocco
fadwa.abakarim@gmail.com

Abstract. This paper presents a voice gender recognition system. Acoustic features and Mel-Frequency Cepstral Coefficients (MFCCs) are extracted to define the speaker's gender. The most used features in these kinds of studies are acoustic features, but in this work, we combined them with MFCCs to test if we will get more satisfactory results. To examine the performance of the proposed system we tried four different databases: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), the Saarbruecken Voice Database (SVD), the CMU_ARCTIC database and the Amazigh speech database (Self-Created). At the pre-processing stage, we removed the silence from the signals by using Zero-Crossing Rate (ZCR), but we kept the noises. Support Vector Machine (SVM) is used as the classification model. The combination of acoustic features and MFCCs achieves an average accuracy of 90.61% with the RAVDESS database, 92.73% with the SVD database, 99.87% with the CMU_ARCTIC database and 99.95% with the Amazigh speech database.

Keywords: Signal processing · Gender recognition · Acoustic features · Mel-Frequency Cepstral Coefficients · Zero-crossing rate · Support Vector Machine

1 Introduction

A voice gender recognition system is one of the most relevant and useful systems in applications such as sorting telephone calls by gender, security systems, human-machine interaction, targeted advertising, and voice recognition in crime scenarios [1–3].

It is easy for a human to differentiate the male voice from the female voice and vice versa. On the other hand, it is a difficult task for a machine. Therefore, several methods are developed to provide the machine with an optimal result. One of the most commonly used approaches is the extraction of the voice frequency features, because the female voice has higher frequency features than those of the male voice. Moreover, the most common of these features is the fundamental frequency (F0). The voice fundamental frequency of a male is from 85 Hz to 180 Hz while that of a female is from 165 Hz to 255 Hz [4, 5].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022, LNCS 13375, pp. 634–648, 2022.
https://doi.org/10.1007/978-3-031-10522-7_43

Archana and Malleswari [3] presented gender identification and performance analysis of speech signals. They mentioned that pitch is the most used feature for voice gender identification, but it is not useful in cases when female and male pitch are similar. Therefore, they used other features such as Mel-Frequency Cepstral Coefficients (MFCCs), energy entropy and frame energy estimation. The classification methods used are the Artificial Neural Network (ANN) and the Support Vector Machine (SVM). The database used in this work is collected from 10 speakers, 5 female and 5 male, and 15 audio files are recorded by each speaker. The test phase is done by using 10 utterances. The experimental results show that SVM performs better than ANN. SVM reaches an accuracy of 80.00% and ANN reaches an accuracy of 40.00%.

Ahmad et al., [6] developed a gender identification system for telephone applications using Mel-Frequency Cepstral Coefficients (MFCCs). This work presented a comparative study between five classification methods: K-Nearest Neighbor (KNN), Naïve Bayes (NB), Multilayer Perceptron (MLP), Random Forest (RF) and Support Vector Machine (SVM). The voice features are extracted by MFCCs. The database used is collected from Korean speakers, and it contains two types of speech recordings: a short speech that lasts a few seconds and long telephone conversations that last between 5 and 8 min. The authors test the performance of their system by varying the size of the training dataset. When the training dataset is small, the classification performance becomes very low. As the size of the training dataset increases, the classification performance increases as well. The experimental results show that SVM performs better than other classification models for both short speech and long telephone speech.

Shareef, Abd and Mezaal [7] realized a gender voice classification with a high accuracy rate. The features are extracted by the well-known method, Mel-Frequency Cepstral Coefficients (MFCCs), and then the Vector Quantization (VQ) is applied for feature selection. In addition, other features are extracted from each voice: the mean, standard deviation (STD), zero-crossing (ZC) and amplitude (AMP). A set of 16 features is collected (12 MFCCs features and 4 features of mean, STD, ZC and AMP). Classification is performed using the machine learning algorithm (J 48). The database used in this work is collected from the CMU_ARCTIC database. The number of voice recordings used is 2770, including 1138 male and 1132 female. The proposed system reaches a high accuracy rate of 99.80%.

Buyukyilmaz and Cibikdiken [8] presented a voice gender recognition system using deep learning. A total of 21 acoustic features is used in this study. The classification step is performed by one of the best neural network models, the Multilayer Perceptron (MLP). The system is examined by 3168 utterances, including 1584 male and 1584 female. The authors succeeded in achieving a high accuracy of 96.74%.

Ramdinmawii and Mittal [9] realized a system of gender identification from speech signals. The features used in this work are Mel-Frequency Cepstral Coefficients (MFCCs), pitch and signal energy. The classification is done by linear Support Vector Machine (SVM). The performance of the proposed system is tested using the Texas Instruments Massachusetts Institute of Technology (TIMIT) database. 120 speech recordings are used, and 6 utterances are recorded by each of 10 male and 10 female speakers. The experimental results show that MFCC coefficients give a higher accuracy of 69.23% compared to the other features, which obtained an accuracy below 60.00%.

Uddin et al., [10] created a gender voice recognition system. The authors use a multilayer architecture with two layers. The first one contains the following features: fundamental frequency, spectral entropy, spectral flatness, mode frequency and mapping. The second one contains the MFCCs features. Two different classification models are used in this work: K-Nearest Neighbors (KNN) and Support Vector Machine (SVM). The performance of the proposed system is tested by three different databases: TIMIT (Texas Instruments Massachusetts Institute of Technology), RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) and a BGC database created by authors. The proposed system achieved a higher accuracy of 96.80% using the TIMIT database and KNN classifier.

These research studies showed that the commonly used methods for voice gender identification are acoustic features, Mel coefficients, and SVM classifier [3, 6–10]. In this paper, we combine acoustic features and MFCCs. In order to test the performance of the proposed system, we leave the speech signals unfiltered. Only the silence is removed but we keep the noises and other disturbances. Four different databases are used with different accents, languages, emotions, voice pitches, diseases and ages. Moreover, only half of each database is used for training and the other half for testing, in order to determine whether the system can reach satisfactory results. Indeed, the larger the training dataset, the better the performance [6, 11].

The rest of the article is organized as follows: the second section describes the set of approaches used for voice gender identification, then in the third section, the experimental results are presented with a discussion and finally the conclusion comes in the fourth section.

2 Applied Methods

In our study, we conducted a three-step process to identify the voice gender:

The first step is pre-processing, in which we removed silence from the input voice signals to compress their length using Zero-Crossing Rate (ZCR). However, we kept the noises and any disturbances found to evaluate the performance.

The second step is feature extraction, in which we used two different approaches. The first approach is to extract acoustic features from the input signals by calculating the mean frequency, standard deviation, median frequency, first quartile, third quartile, interquartile range, skewness, kurtosis, spectral entropy, spectral flatness and mean voice fundamental frequency. In addition, the second approach is to extract MFCCs features.

Finally, the last step is classification, in which two methods are used. The first method is to determine a threshold from the acoustic features of the training dataset. Then, we test whether the average of the acoustic features of the test signals is below or above the threshold. If it is below, then the voice gender is male, if it is the opposite, then the voice gender is female. This method is only valid for acoustic features because they are extracted from the voice fundamental frequencies and frequency spectrums. Moreover, the second method is one of the widely used classifiers in the field of speech recognition, the SVM model.

Figure 1 shows the three-step process used in our study.

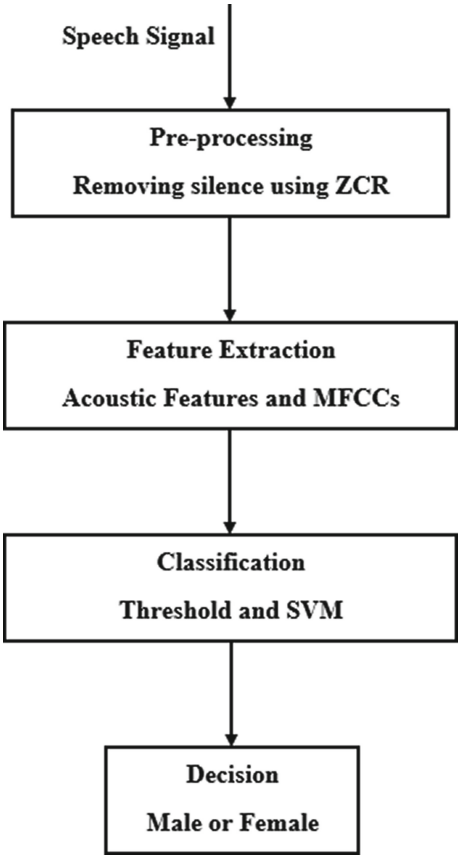

Fig. 1. Steps of the proposed system.

2.1 Pre-processing

Signal pre-processing is an important step in speech recognition. It allows us to compress the signals and to keep their qualities by removing the useless information. In our study, we used the Zero-Crossing Rate (ZCR) method to detect the unvoiced parts of the signal and remove them to obtain a 100% voiced signal.

The Zero-Crossing Rate is a measure of the number of times the amplitude of the speech signals crosses the value zero in a given frame. When ZCR is high, the speech signal is unvoiced and when it is low, the speech signal is voiced, as shown in Fig. 2 [12, 13].

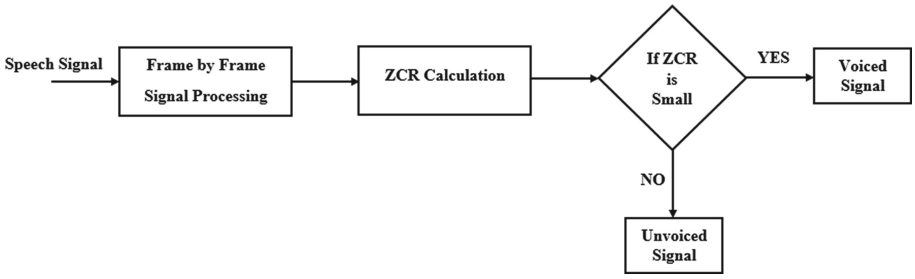

Fig. 2. Operation of ZCR.

The calculation of ZCR is done as follows [13, 14]:

$$Z(n) = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]| w(n-m) \tag{1}$$

where

$$sgn[x(n)] \begin{cases} 1 \ if \ x(n) \geq 0 \\ -1 \ if \ x(n) < 0 \end{cases} \tag{2}$$

with $w(n)$ is the windowing function with a window size of N samples:

$$w(n) \begin{cases} \frac{1}{2N} \ for \ 0 \leq n \leq N-1 \\ 0 \ for \ otherwise \end{cases} \tag{3}$$

Figure 3 presents the speech signal before and after using ZCR.

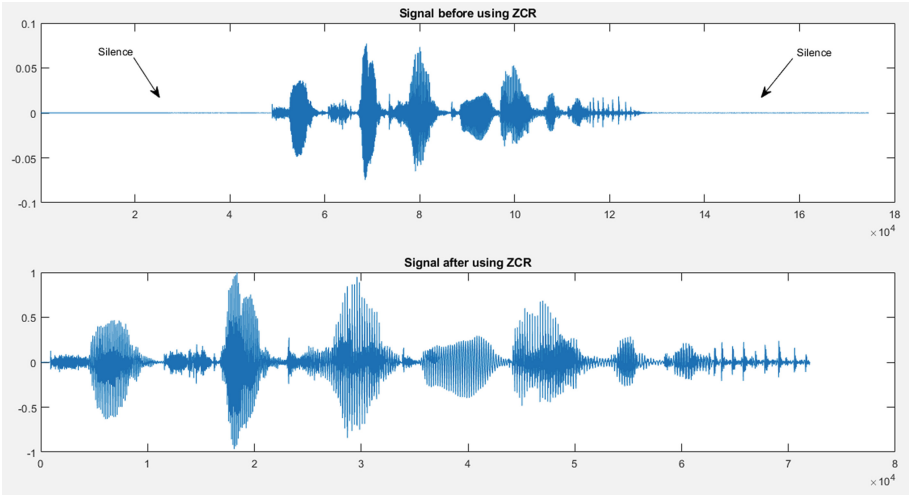

Fig. 3. Removing silence from the speech signal by using ZCR.

2.2 Feature Extraction

Feature extraction has a significant impact on the identification of processed signals. There are several methods for extracting important information from signals. Among them are acoustic features and Mel-Frequency Cepstral Coefficients.

Acoustic Features. Acoustic features are a set of properties that identify the voice, including the frequency features that are extracted after transforming the signal from the time domain to the frequency domain.

In our study, we extracted the following 11 features [4, 10]:

Mean Frequency. The mean frequency of a speech is the average value of the speech frequency in a specific interval of the frequency spectrum. It is calculated as the sum of the frequencies in the spectrum divided by the total number of frequency frames in a spectrum. Equation (4) presents the calculation:

$$f_{mean} = \frac{\sum_{i=0}^{n} f_i}{n} \tag{4}$$

where f_i is the frequency of ith frame and n is the number of frequency frames in the spectrum.

Standard Deviation. The standard deviation of a speech frequency is a measure of the frequencies in a spectrum that differ from the mean frequency. It is calculated as follows:

$$f_{sd} = \sqrt{\frac{\sum_{i=0}^{n} (f_i^2 - f_{mean}^2)}{n}} \tag{5}$$

where f_i is the frequency of ith frame, n is the number of frequency frames in the spectrum and f_{mean} is the mean frequency mentioned at Eq. (4).

Median Frequency. The median frequency of a speech is the median value of the speech frequency in a specific interval of the frequency spectrum. It is calculated as follows:

$$f_{median} = \frac{\sum_{i=0}^{n} f_i}{2} \tag{6}$$

where f_i is the frequency of ith frame and n is the number of frequency frames in the spectrum.

First Quartile. To find the first quartile called Q1, we have to sort the frequency values obtained from the spectrum, and then it is the median of the lower half of the speech frequency set.

Third Quartile. To find the third quartile called Q3, we have to sort the frequency values obtained from the spectrum, and then it is the median of the upper half of the speech frequency set.

Interquartile Range. To find the interquartile range called IQR, we first have to find the first quartile (Q1) and the third quartile (Q3), and then calculate the difference between them.

Skewness. Skewness is a measure of the lack of symmetry. A distribution is symmetric if it looks the same to the left and right of the mean. It is calculated as follows:

$$f_{skewness} = \frac{\frac{1}{n}\sum_{i=0}^{n}(f_i - f_{mean})^3}{f_{sd}^3} \tag{7}$$

where f_i is the frequency of ith frame, n is the number of frequency frames in the spectrum, f_{mean} is the mean frequency mentioned at Eq. (4) and f_{sd} is the standard deviation value mentioned at Eq. (5).

Kurtosis. Kurtosis determines the heaviness of the distribution tails. It is calculated as follows:

$$f_{kurtosis} = \frac{\frac{1}{n}\sum_{i=0}^{n}(f_i - f_{mean})^4}{f_{sd}^4} \tag{8}$$

where f_i is the frequency of ith frame, n is the number of frequency frames in the spectrum, f_{mean} is the mean frequency mentioned at Eq. (4) and f_{sd} is the standard deviation value mentioned at Eq. (5).

Spectral Entropy. Spectral entropy is a measure of spectral power distribution of a speech signal. it is calculated as follows:

$$f_{spent} = \frac{-sum(\sum_{i=0}^{n} y_i * \log(\sum_{i=0}^{n} y_i))}{\log(n)} \tag{9}$$

where y_i is the relative amplitude value of the frequency at the ith frame and n is the number of frequency frames in the spectrum.

Spectral Flatness. Spectral flatness is the ratio of geometric mean to the arithmetic mean of a power spectrum.

Mean Voice Fundamental Frequency. The fundamental frequency is also known as the voice pitch. It is the frequency of the vocal folds and to calculate it we used autocorrelation.

Autocorrelation is a time domain method and a mathematical function that used to determine the similarity degree. Moreover, it is one of the most used methods for the extraction of the pitch, and the following equation shows the calculation:

$$A(\tau) = \frac{1}{T}\sum_{t=0}^{T-1} S(t)S(t+\tau) \tag{10}$$

with $\tau = [0, T-1]$.

where $A(\tau)$ is the autocorrelation function at lag τ, T is the size of the frame of a speech signal, $S(t)$ is the speech signal and τ is the lag number or delay time. When lag, $\tau = 0$ the autocorrelation value always represents the maximum value.

After calculating the autocorrelation value, we used Eq. (11) to calculate the fundamental frequency f_0, then we calculated the mean value.

$$f_0 = \frac{1}{T} \tag{11}$$

Mel-Frequency Cepstral Coefficients. The Mel-Frequency Cepstral Coefficients is one of the most commonly used methods by researchers for speech recognition. Since speech is a non-stationary signal, short-term spectral analysis is the most common way to characterize the speech signal [3].

In our study, we obtained 14 Mel coefficients by following these steps [1, 4, 7, 10, 15]:

The first step to extract MFCCs features is to segment the speech into multiple frames, the size of each frame should be neither shorter nor longer. If it is shorter, it will not be enough to extract special information from the signal. On the other hand, if it is longer, the signal will be highly changed throughout the frame. The optimal solution is to segment the signal into a length of 20–40 ms for each frame. Then we used the Hamming Window to minimize the discontinuities of the speech signal before and after each frame of the window. Equation (12) shows the calculation of the Hamming Window:

$$W(n) = 0.54 - 0.46\cos[\frac{2\pi n}{N-1}] \tag{12}$$

with $0 \leq n \leq N-1$.

where $W(n)$ is the window operation, n is the number of each individual sample and N is the total number of speech samples.

The next step is to transform each frame of the speech signal from the time domain to the frequency domain. Therefore, the Fast Fourier Transform (FFT) is applied to obtain the spectrum of each frame. The frequency range of the FFT spectrum is very wide and the speech signal does not follow the linear scale. Therefore, the bank of filters is processed according to the Mel scale. Equation (13) is used to convert a given frequency in Hz to Mel scale:

$$Mel(f) = 1125 * \ln(1 + \frac{f}{700}) \tag{13}$$

where f is the given frequency in Hz and $Mel(f)$ is the perceived Mel frequency.

The final step is to convert the log Mel to the time domain by using the Discrete Cosine Transform (DCT). The result of this operation is called the Mel-Frequency Cepstrum Coefficient and the set of these coefficients is the acoustic vectors.

Figure 4 shows the steps of MFCCs extraction.

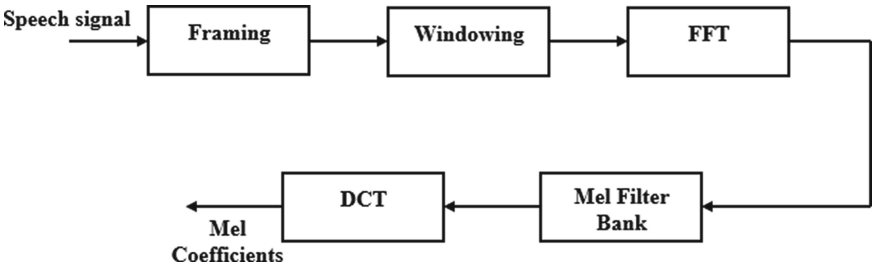

Fig. 4. Steps to get MFCC coefficients.

2.3 Classification

Classification is the next step that comes after feature extraction. The choice of a classification model is very important as it affects the final decision and the performance of the system. In this study, we used two methods of classification. The first one consists in classifying the test signals by using a threshold based on the average of the training dataset features, and the second one is the Support Vector Machine (SVM) model [16].

Threshold. The frequency features help the machine to differentiate the male voice from the female voice since the frequency data of the female voice is higher than of the male voice. Therefore, we calculated the mean value of the training dataset features for male and female. As a result, we obtain a threshold. If the mean value of the test signal features is above the threshold, it is a female voice and if it is below, it is a male voice [4, 5]. This classification approach is only valid with frequency features and not with MFCCs.

Support Vector Machine. Support Vector Machine (SVM) is one of the most powerful classifiers in the field of speech recognition. SVM supports binary classification. In this work, we have the female class and the male class. SVM goes through two steps, in the first step, it is very important to choose a kernel function. We chose the Radial Basic Function (RBF) kernel, which performs a low to high dimensional feature transformation. This transformation allows non-linearly separable data to be linearly separable at a higher dimension. Then, in the second step, a maximum margin hyperplane is constructed to draw the decision boundary between the classes [6, 17, 18].

3 Experimental Results and Discussion

In this section, we describe the four databases used in this study and we present and discuss the results obtained.

3.1 Data

To test the performance of our work we used four different databases: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [19], the Saarbruecken Voice Database (SVD) [20], the CMU_ARCTIC database [21] and the Amazigh speech database (Self-Created).

RAVDESS. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a widely used database for speech emotion recognition. It contains 1440 audio files in .wav format with a frequency of 48 kHz, recorded by 24 authors, including 12 male and 12 female. The database is composed of 8 emotions: calm, happy, sad, angry, fearful, surprised, neutral and disgusted [22]. The language of the audio files is English. This database is chosen to test the performance of our system with speech signals having different tones. We divided the database in two, the first half for training and the second for testing as shown in Table 1.

SVD. The Saarbruecken Voice Database (SVD) is a widely used database for voice pathology detection. It is a German database, which contains a set of voice recordings from more than 2000 people, including males and females of different ages. The voice recordings contain the vowels /a/, /i/, and /u/ in a normal, high, low, and low-high-low pitch. Moreover, they contain the sentence "Guten Morgen, wie geht es Ihnen?" which means "Good morning, how are you?" in English. The length of each voice recording is between 1 and 4 s with a frequency of 50 kHz. The database contains 71 type of voice pathologies. In our study, we used the audio files of the recorded sentence "Guten Morgen, wie geht es Ihnen?". The SVD is chosen to test the performance of our system with pathology speech signals [23]. We divided the database in two, the first half for training and the second for testing as shown in Table 1.

CMU_ARCTIC Database. The CMU_ARCTIC database is widely used for voice gender recognition. It contains 1132 audio files in.wav format with a frequency of 32 kHz for female gender and 1138 audio files in.wav format with a frequency of 16 kHz for male gender. The language of the audio files is English. We chose to test the performance of our system with the CMU_ARCTIC database because it is commonly used by researchers for voice gender recognition [4, 24]. We divided the database in two, the first half for training and the second for testing as shown in Table 1.

Amazigh Speech Database. The Amazigh speech database is a self-created database that contains 300 audio files of numbers from one to five in the Amazigh language (yan, sin, krad, kuz, smmus) [12, 25, 26]. The duration of each audio file is between 1 and 3 s with a frequency of 16 kHz. A total of 60 people participated in the recording, including 30 male and 30 female. We created this database to test the performance of our system with a new type of language and accent. We divided the database in two, the first half for training and the second for testing as shown in Table 1.

Table 1. Information about databases used in this study.

Database name	Number total of audio files		Training dataset		Test dataset	
	Male	Female	Male	Female	Male	Female
RAVDESS	720	720	360	360	360	360
SVD	880	1108	440	554	440	554
CMU_ARCTIC database	1138	1132	569	566	569	566
Amazigh database	150	150	75	75	75	75

3.2 Results

The aim of this work was to detect the gender of the speech: whether it is male or female. This system can be used in several fields including security systems and human-machine interaction. In our study, we extracted a total of 25 features using two approaches, 11 acoustic features and 14 MFCCs features. The classification is done by two methods, the first method is to use a threshold as mentioned in Sect. 2 part 3, and the second method is to use the SVM classifier. To test the performance of the proposed system, we did not filter the input signals; we only removed the silence. The experiments were performed on a MATLAB 2019 platform using an Intel Core i7 processor with 8 GB of RAM and a 64-bit operating system.

Table 2 shows the accuracy rates of voice gender recognition using acoustic features. As mentioned in Sect. 2, part 3, a threshold was created to perform the classification. We obtained good results using this classifier, but they are not as good as those obtained using the SVM classifier. We also observe that the RAVDESS database has the lowest results compared to the other databases. With the threshold, RAVDESS achieved a rate of 78.06% for male recognition and 80.28% for female recognition. Furthermore, with the SVM classifier, it achieved a rate of 88.61% for male recognition and 90.22% for female recognition. On the other hand, the Amazigh speech database obtained the best results compared to the other databases. With the threshold, it reached a rate of 96.15% for male recognition and 95.01% for female recognition. Moreover, with the SVM classifier, it reached a rate of 99.80% for male recognition and 99.83% for female recognition.

Table 2. The accuracy rates of voice gender recognition using acoustic features.

Classifier	Databases							
	RAVDESS (%)		SVD (%)		CMU_ARCTIC (%)		Amazigh database (%)	
	Male	Female	Male	Female	Male	Female	Male	Female
Threshold	78.06	80.28	83.86	89.35	82.60	86.11	96.15	95.01
SVM	88.61	90.22	90.27	94.43	99.47	99.65	99.80	99.83

Table 3 shows the accuracy rates of voice gender recognition using MFCCs features. The threshold mentioned in Table 2 is only valid for frequency features. Since females have higher frequency information than males, only the SVM classifier is used for the MFCCs features.

We can see in Table 3 that the results obtained are lower than the ones obtained in Table 2 using the acoustic features with the SVM classifier. However, it is not the same with the CMU_ARCTIC database. We can observe that it has higher results than those in Table 2, where it reached an accuracy rate of 99.47% for male recognition and 99.65% for female recognition. On the other hand, with MFCCs it achieved an accuracy rate of 99.73% for male recognition and 99.77% for female recognition. This is a difference of 0.26% for male recognition and a difference of 0.12% for female recognition.

Table 3. The accuracy rates of voice gender recognition using MFCCs features.

Classifier	Databases							
	RAVDESS (%)		SVD (%)		CMU_ARCTIC (%)		Amazigh database (%)	
	Male	Female	Male	Female	Male	Female	Male	Female
SVM	87.22	88.39	90.15	93.50	99.73	99.77	99.23	95.22

After using the acoustic and MFCCs features individually, we combined them to see if this would provide better results than the ones obtained in Tables 2 and 3.

Table 4 shows that the results of the RAVDESS database are always the lowest, whether the acoustic and MFCCs features are used separately or in combination. Indeed, the signals in the database present 8 different emotions: calm, happy, sad, angry, fearful, surprised, neutral and disgusted. Therefore, they have different tones and each of them varies in a different way. In addition, for the four databases, we did not remove noises and other disturbances that also affect the results. However, in the end, we obtained higher results even though we did not filter the signals and used different databases with different standards.

The Amazigh speech database reaches an almost 100% accuracy for the recognition of male and female gender. The same goes for the CMU_ARCTIC database, while the SVD and RAVDESS databases achieve more than 90.00% accuracy.

Table 4. The accuracy rates of voice gender recognition by combining acoustic features and MFCC coefficients.

Classifier	Databases							
	RAVDESS (%)		SVD (%)		CMU_ARCTIC (%)		Amazigh database (%)	
	Male	Female	Male	Female	Male	Female	Male	Female
SVM	90.44	90.78	90.68	94.77	99.85	99.89	99.91	99.98

Table 5 presents the results obtained in our study with the results of other existing studies.

Table 5. The highest results obtained in our study with other existing studies

State of the art	Proposed method	Database	Accuracy (%)
[7]	MFCC, VQ and machine learning algorithm (J 48)	CMU_ARCTIC	99.80
[10]	Fundamental frequency, spectral entropy, spectral flatness, mode frequency, MFCCs features and KNN classifier	TIMIT	96.80
[27]	GMM-SV based SVM developed with PLP features	aGender	90.40
[28]	Pitch, energy and MFCC	TIMIT	96.45
[29]	Pitch, MFCC	OGI multilingual corpus	96.53
[30]	VAD, MFCC and PCA	RAVDESS	98.88
[31]	MFCC and BiLSTM	QDAT	94.01
Current study	Acoustic features, MFCC coefficients and SVM	Amazigh speech database (with highest results)	99.91 (Male) 99.98 (Female)

4 Conclusion

In this work, gender voice recognition goes through three steps: pre-processing, feature extraction and finally classification. First, ZCR is used to compress the input signals by removing silence. Then we extracted the acoustic and MFCCs features. Finally, we used threshold and SVM as classifiers. As mentioned in the introduction section, the researchers used a limited number of signals for training and testing, as it was stated that a small training dataset would lead to a low system performance. Therefore, we want to experiment the opposite and used 50% of the database for testing and the other 50% for training. As a result, although we used unfiltered signals and different databases with different ages, accents, languages, emotions and diseases, our system performed perfectly, especially when we combined the acoustic features with the MFCC coefficients. The accuracy rate was almost 100% for the Amazigh speech and CMU_ARCTIC databases, as well as over 90% for the RAVDESS and SVD databases. The execution time was fast in the second and third steps, but the pre-processing step takes more time. The system can for example be used for human-machine interaction, security systems, and digital advertisements by changing the graphical interface if the detected voice is male or female. In the future, we will improve our work by realizing a gender and emotion recognition system with an Amazigh speech database, which is rarely used in the field of speech recognition. In addition, we will use cross-training and cross-testing to perform the system.

Acknowledgment. A special thanks to Mr. T. Hobson from the Anglosphere English Center for reviewing for spelling and grammatical mistakes. In addition, to all participants who recorded their voices for this research.

References

1. Alkhawaldeh, R.S.: DGR: Gender recognition of human speech using one-dimensional conventional neural network. Sci. Program. (2019). https://doi.org/10.1155/2019/7213717
2. Ng, C.B., Tay, Y.H., Goi, B.M.: Vision-based human gender recognition: A survey. In: Proceedings of the Computer Vision and Pattern Recognition (2012). https://doi.org/10.48550/arXiv.1204.1611
3. Archana, G.S., Malleswari, M.: Gender identification and performance analysis of speech signals. In: Proceedings of the 2015 Global Conference on Communication Technologies (GCCT), pp. 483–489. IEEE (2015). https://doi.org/10.1109/GCCT.2015.7342709
4. Hong, Z.: Speaker gender recognition system, Master's thesis, degree programme in wireless communications engineering. University of Oulu, Oulu, Finland, p. 54 (2017)
5. Titze, I.R.: Measurements for voice production: Research and clinical applications. J. Acoust. Soc. Am. (1998)
6. Ahmad, J., Fiaz, M., Kwon, S.I., Sodanil, M., Vo, B., Baik, S.W.: Gender identification using MFCC for telephone applications - a comparative study. Int. J. Comput. Sci. Electron. Eng. **3**(5), 351–355 (2015). https://doi.org/10.48550/arXiv.1601.01577
7. Shareef, M.S., Abd, T., Mezaal, Y.S.: Gender voice classification with huge accuracy rate. Telkomnika **18**(5), 2612–2617 (2020). https://doi.org/10.12928/TELKOMNIKA.v18i5.13717
8. Buyukyilmaz, M., Cibikdiken, A.O.: Voice gender recognition using deep learning. In: Proceedings of the 2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA), vol. 58, pp. 409–411. Atlantis Press (2016)
9. Ramdinmawii, E., Mittal, V.K.: Gender identification from speech signal by examining the speech production characteristics. In: Proceedings of the 2016 International Conference on Signal processing and Communication (ICSC), pp. 244–249. IEEE (2016). https://doi.org/10.1109/ICSPCom.2016.7980584
10. Uddin, M.A., Hossain, M.S., Pathan, R.K., Biswas, M.: Gender recognition from human voice using multi-layer architecture. In: Proceedings of the 2020 International Conference on Innovations in Intelligent Systems and Applications (INISTA), pp. 1–7. IEEE (2020). https://doi.org/10.1109/INISTA49547.2020.9194654
11. Garg, D., Kaur, S., Arora, D.: Comparative analysis of speech processing techniques for gender recognition. Int. J. Adv. Electr. Electron. Eng., 278–283(2012)
12. Abakarim, F., Abenaou, A.: Amazigh isolated word speech recognition system using the adaptive orthogonal transform method. In: Proceedings of the 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), pp. 1–6. IEEE (2020). https://doi.org/10.1109/ISCV49265.2020.9204291
13. Bachu, R.G., Kopparthi, S., Adapa, B., Barkana, B.D.: Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In: Proceedings of the American Society for Engineering Education (ASEE) zone conference proceedings, pp. 1–7 (2008)
14. Shete, D.S., Patil, S.B., Patil, S.: Zero crossing rate and energy of the speech signal of devanagari script. IOSR J. VLSI and Signal Process. **4**(1), 01–05 (2014). https://doi.org/10.9790/4200-04110105

15. Muda, L., Begam, M., Elamvazuthi, I.: Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. J. Comput. **2**(3), 138–143 (2010)
16. Noble, W.S.: What is a support vector machine? Nat. Biotechnol. **24**(12), 1565–1567 (2006). https://doi.org/10.1038/nbt1206-1565
17. Fokoue, E., Ma, Z.: Speaker gender recognition via MFCCs and SVMs. Rochester Institute of Technology RIT Scholar Works, pp. 1–9 (2013)
18. Jena, B., Mohanty, A., Mohanty, S.K.: Gender recognition and classification of speech signal. In: Proceedings of the 2021 International Conference on Smart Data Intelligence (ICSMDI), pp. 1–7. SSRN (2021)
19. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American english. PLoS ONE **13**(5), e0196391 (2018). https://doi.org/10.1371/journal.pone.0196391
20. Barry, W.J., Putzer, M.: Saarbruecken voice database. http://www.stimmdatenbank.coli.uni-saarland.de/. Accessed 01 Mar 2022
21. Kominek, J., Black, A.: The CMU Arctic speech databases for speech synthesis research. Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMULTI-03–177 (2003). http://www.festvox.org/cmu_arctic/. Accessed 20 Feb 2022
22. Bhavan, A., Chauhan, P., Hitkul, Shah, R.R.: Bagged support vector machines for emotion recognition from speech. Knowl.-Based Syst. (2019)
23. Abakarim, F., Abenaou, A.: Voice pathology detection using the adaptive orthogonal transform method, SVM and MLP. Int. J. Online Biomed. Eng. **17**(14), 90–102 (2021)
24. Livieris, I.E., Pintelas, E., Pintelas, P.: Gender recognition by voice using an improved self-labeled algorithm. Mach. Learn. Knowl. Extr. **1**(1), 492–503 (2019). https://doi.org/10.3390/make1010030
25. Idhssaine, A., El Kirat, Y.: Amazigh language use, perceptions and revitalisation in Morocco: The case of rabat-sale region. J. North Afr. Stud. **26**(3), 465–479 (2021). https://doi.org/10.1080/13629387.2019.1690996
26. Zaid, H., El Allame, Y.E.K.: The place of culture in the Amazigh language textbooks in Morocco. L1-Educ. Stud. Lang. Lit. **18**, 1–20 (2018). https://doi.org/10.17239/L1ESLL-2018.18.01.01
27. Yücesoy, E., Nabiyev, V.V.: A new approach with score-level fusion for the classification of a speaker age and gender. Comput. Electr. Eng. **53**, 29–39 (2016). https://doi.org/10.1016/j.compeleceng.2016.06.002
28. Chaudhary, S., Sharma, D.K.: Gender identification based on voice signal characteristics. In: Proceedings of the 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), pp. 869–874. IEEE (2018). https://doi.org/10.1109/ICACCCN.2018.8748676
29. Keyvanrad, M.A., Homayounpour, M.M.: Improvement on automatic speaker gender identification using classifier fusion. In: Proceedings of the 2010 18th Iranian Conference on Electrical Engineering, pp. 538–541. IEEE (2010). https://doi.org/10.1109/IRANIANCEE.2010.5507010
30. Nashipudimath, M.M., Pillai, P., Subramanian, A., Nair, V., Khalife, S.: Voice feature extraction for gender and emotion recognition. In: Proceedings of the ITM Web Conferences, vol. 40, p. 03008. EDP Sciences (2021). https://doi.org/10.1051/itmconf/20214003008
31. Mohammed, A.A., Al-Irhayim, Y.F.: Speaker age and gender estimation based on deep learning bidirectional long-short term memory (BiLSTM). Tikrit J. Pure Sci. **26**(4), 76–84 (2021)