# Voice Gender Recognition under Unconstrained Environments using Fine-Tuned CNNs

Jorge JORRIN-COZ, Mariko NAKANO[1],
Jonathan FLORES-MONROY and Hector PEREZ-MEANA
*Mechanical and Electrical Engineering School, Instituto Politecnico Nacional*
ORCiD ID: Mariko Nakano https://orcid.org/0000-0003-1346-7825

**Abstract.** Automatic voice gender recognition (VGR) offers several real-world applications, including recommender system, human-robot interaction, and forensic application. VGR systems become challenging when these operate under unconstrained environments. In this study, we evaluate the performance of VGR systems using different fine-tuned pretrained Convolutional Neural Networks (CNNs), in which the speech signals under unconstrained environments are introduced as input data. First, preprocessing is applied to the original speech signal, which consists of noise attenuation based on low-pass filter and silence part removal based on sound amplitude. Then, the time-frequency features, such as Spectrogram, Mel-Spectrogram and Mel Frequency Cepstral Coefficients (MFCC) are extracted, which are converted into RGB images and processed by CNN models. Our research utilizes the VoxCeleb dataset, which is the largest video-audio dataset recorded under unconstrained environments. The results obtained by several fine-tuned CNN models provide higher accuracy compared with the state-of-the-art techniques on this topic. The best accuracy achieved is 98.58% using fine-tuned MobileNet, which is higher than the best accuracy provided by previous works.

**Keywords.** Voice gender recognition, Unconstrained environment, Mel-Spectrogram, MFCC, VoxCeleb, CNN, Fine-tuning

## 1. Introduction

Speech signals have been considered as the most useful communication medium among human beings. Recently, these speech signals have been used for human-machine communication, such as human-robot interaction, automatic question-answering systems among others [1]. In these applications, speaker characteristics such as gender, age, mood, etc. are very important factors to obtain satisfactory results. Therefore, the acquisition of individual characteristics through speech signal analysis has been the subject of various research effort in recent decades, especially due to the increase in computational power that have enabled greater processing power to execute more complex algorithms. In addition to the dominant linguistic information, speech also carries paralinguistic information such as speaker identity, emotional estate, and ethnicity.

[1] Corresponding Author: Mariko Nakano, mnakano@ipn.mx

Using this paralinguistic information, the Automatic Speaker Profiling (ASP) becomes an active research field, which estimates the speaker characteristics, such as age, race, gender, weight, height, mood, biomarker, among others, from the speech signal. Applications of ASP span many fields, and these are used in a wide range of real-world applications, including surveillance, forensics, human-robot interaction, and commercial applications [2]-[4]. The gender-dependent advertisements, customized services, and caller-agent pairing are principal commercial applications of the ASP. Nowadays, voice signal is utilized in several assistant devices, such as mobile phones, smart homes, and smart buildings. These applications have accelerated the development of various ASP systems.

The automatic Voice Gender Recognition (VGR) is a subset of the ASP that infers speaker gender information from his/her speech signal. The VGR can be applied in several commercial and no-commercial voice-based systems, such as voice-based recommender system and voice-based survey system. Gender information obtained from voice signal can be used to offer better and customized service. In security and surveillance environments, such as airports or public transportation stations, the ability to identify the gender of individuals can be useful for enhancing security and implementing appropriate control measures. In the healthcare field, an accurate gender classification system can aid in medical and epidemiological research, allowing for a better understanding of how certain diseases or health conditions affect different demographic groups.

Although the VGR is relatively easy for humans to recognize the gender of the speaker in controlled environment, this task becomes difficult under unconstrained environment, where noise and/or other environmental sound such as background music, other human voice etc. are presented together with speech signal of interest. Generally, under the unconstrained environment, the performance of automatic VGR systems is significantly degraded. As well as ASP systems, automatic VGR systems offer several important applications, such as gender-sensitive recommender systems and gender-specific customized services. These applications are often used via smartphones in unconstrained environments where the desired speech signal is mixed with some unwanted environmental sounds.

Until now, several VGR systems have been developed and they provide good classification performance. However, many of them used dataset in controlled environment, such as Vowel Dataset [5] and the Texas Instruments/Massachusetts Institute of Technology Acoustic-Phonetic Continuous Speech Corpus (TIMIT) [6]. The audio signals in these datasets are clear and free of interfering sounds such as music, other people's voices, and environmental sounds. As mentioned above, considering that VGR systems are often used in unconstrained environments, there is a need to develop VGR systems that are robust against unwanted environmental noise.

In this work, we provide an efficient automatic VGR system under unconstrained environment. Intensive performance evaluation of several fine-tuned CNNs resulted in the best model. As a dataset, we used VoxCeleb, developed by the University of Oxford [7]. This dataset contains over 100,000 unconstrained utterances by 1,251 speakers with different nationalities and different languages [8]. VoxCeleb consists of voice signals and corresponding video frames obtained from YouTube. Some of the data were interviews recorded outdoors in a very noisy environment and included multiple voice signals and background music.

The proposed system first performs preprocessing to attenuate the noise signal and remove silence parts from the original input signal. The preprocessed signals are converted into time-frequency data such as Spectrogram, Mel-spectrogram, and Mel Frequency Cepstral Coefficients (MFCC), and then these time-frequency data are stored as RGB color images, which serve as input data for fine-tuned CNNs. The best accuracy is achieved using fine-tuned MobileNet, which is 98.58%. This accuracy is higher than the best accuracy obtained by the previous published works with VoxCeleb dataset. Additionally, we carried out GradCAM-based visual analysis [8] about attention to be paid by the fine-tuned CNN models in the time-frequency representation for its classification. From this analysis, we can see that the CNN models pay attention to some harmonic frequency bands to determine the gender, which coincides with some voice research [9, 10].

The remainder of this paper is organized into four sections as follows. In Section 2, brief descriptions of related works are provided. Section 3 describes the proposed methodology in detail, and Section 4 presents the performance obtained by several CNNs together with a comparison among previous works, and a visual analysis to interpret the decision made by the CNNs. Finally, Section 5 provides the conclusions of this work.

## 2. Related Works

There is a large amount of literature on VGR systems, with some promising results. As mentioned earlier, most papers present results based on datasets consisting of controlled audio signals. The related works can be basically divided into two categories. The first category uses conventional classifiers, such as support vector machine (SVM) and Logistic Regression, introducing hand-crafted features extracted from speech signals as input data. The second one uses deep neural networks as a classifier and speech signal, or its time-frequency representation is applied directly as input data. In general, the second category provides a better performance than the first one.

As a work belonging to the first category, Pahwa and Aggarwal used MFCC and delta features to extract gender-related features from audio signals [5]. They employed support vector machine (SVM) as a classifier and achieved an accuracy of 93.48% when the system applied to a controlled Vowel dataset [5]. Chaudhary and Sharma achieved 96.45% accuracy using TIMIT [6], another controlled dataset. They utilize pitch, energy, and MFCC as extracted features, and SVM is used as a classifier [6].

Recently, several VGR systems based on deep neural networks have been developed with good results. In [11], the author proposed a VGR system based on Long-Short Term Memory (LSTM) and this proposed system was evaluated using his own dataset and an accuracy of 98.40% was achieved. Uddin et al. extracted MFCC and Linear Predictive Coding (LPC) from input speech signals and used a one-dimensional CNN as classifier. In the VGR task, the authors worked with TIMIT dataset and obtained an accuracy of 93.01% [12]. Alnuaim et al., using a pretrained ResNet 50 with fine tuning, obtained a result of 98.57% using the Voice Common Dataset. They used MFCC, Mel-Spectrogram and Chroma as gender-related features [13]

Some few studies used VoxCeleb dataset for evaluation. Nasef et al. proposed two VGR systems using self-attention mechanism [3], which was inspired by Transformer [14]. They used Logistic Regression as classifier and the best accuracy obtained using VoxCeleb was 96.23%. Hechmi et al. used a combination of MFCC and i-vector for speech features and logistic regression as a classifier [15]. They achieved a commendable result of 98.29% F1-Score. This study was conducted over the VoxCeleb dataset.

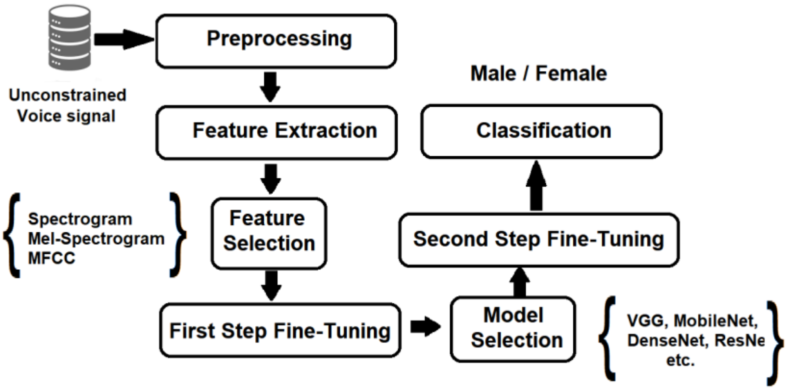


**Figure 1.** Block diagram of the proposed method.

## 3. Proposed Method

A block diagram of the proposed method is shown in Fig. 1. The proposed method is composed of a preprocessing stage, feature extraction stage, two fin-tuning stages using pre-trained CNN models, and classification stage. The following subsections provide detailed descriptions of each stage.

### *3.1. Preprocessing Stage*

In the VGR tasks, signal preprocessing played a crucial role in enhancing the robustness of the signal analysis. To deal with the presence of unwanted noise, a low-pass filter with cut-frequency 4kHz was applied to attenuate high-frequency components, effectively reducing noise interference. Additionally, silence parts were removed from the original signal to focus only on active speech segments, enhancing the relevance of the processed data. The application of a pre-emphasis filter is realized through a first-order high-pass filter, simplifies the amplification of high-frequency components, and facilitates the subsequent feature extraction process.

Energy normalization is applied to increase the consistency of different speech signals by adjusting their amplitude. This step is particularly useful in speech processing applications, where it is important to focus on the spectral characteristics of the signal rather than its absolute amplitude. By normalizing the energy, variations in the overall loudness of different speech signals are mitigated, and the emphasis is placed on the spectral characteristics and patterns of the speech signal. The energy normalization of signal is given by (1).

$$x_{normalized}(t) = \frac{x(t)}{\sqrt{E}} \tag{1}$$

where $x(t)$ represents the original speech signal, $E$ is the energy of the signal, and $x_{normalized}(t)$ is the energy-normalized signal. Figs. 2 and 3 show some examples of the preprocessing stage.

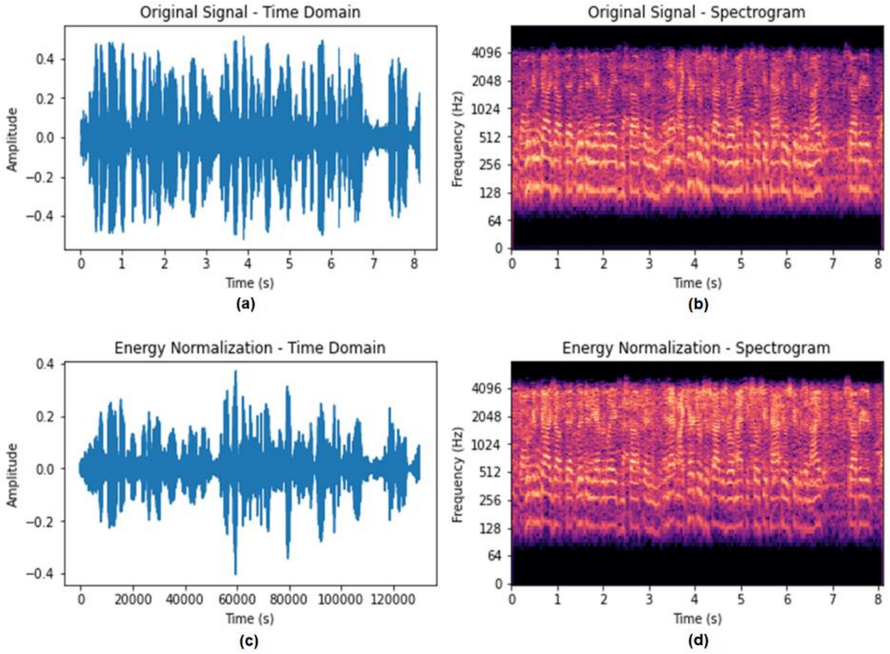


**Figure 2.** Preprocessing of voice signal and their time-frequency representation. (a) Original signal, (b) Spectrogram of (a), (c) Normalized signal and (d) Spectrogram of (c)

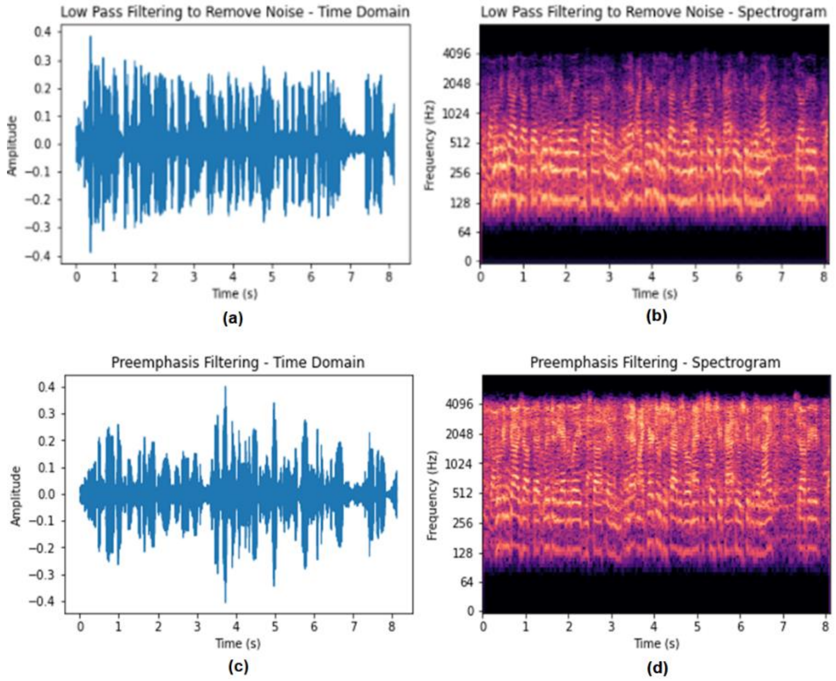


**Figure 3.** Preprocessing of voice signal and their time-frequency representation. (a) Filtered signal by low-pass filter (b) Spectrogram of (a), (c) Filtered signal by pre-emphasis filter and (d) Spectrogram of (c)

### *3.2. Feature Extraction Stage*

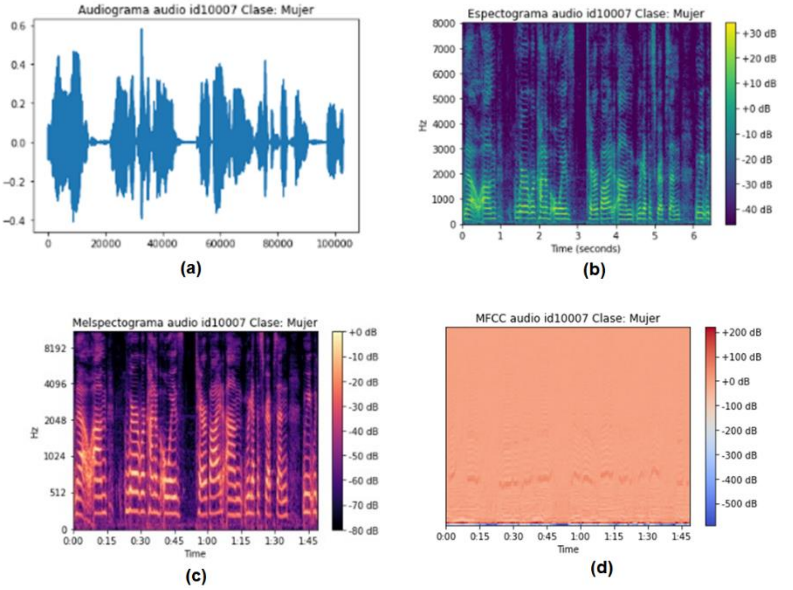


**Figure 4.** (a) Audio signal, (b) Spectrogram, (c) Mel-Spectrogram, (d) MFCC, using audio signal id10007, which corresponds to female voice.

In the proposed method, we used different techniques for extraction of time-frequency representation of raw audio signal, such as Spectrogram, Mel Spectrogram and MFCC. A spectrogram is a two-dimensional graph that represents the intensity of frequency components in each time frame in terms of brightness or color tone. In this graph, the X-axis represents the time frame, and the Y-axis represents the frequency components. If this two-dimensional graph is interpreted as a color image, several pretrained 2D CNN can be used for classification task [13]. The Spectrogram is obtained by applying Short Time Fourier Transform (STFT) to the speech signal and the Mel Spectrogram is obtained by applying Mel-scale to the frequency component of the Spectrogram. The MFCC is a method for extracting more compressed elements from a Mel-Spectrogram, and it is possible to extract low-dimensional data by applying cosine transformation to a logarithm of Mel-Spectrogram.

The following parameters were used to calculate the Spectrogram, Mel-Spectrogram, and MFCC. The sampling rate: 16 KHz, number of data: 48000, hop length: 160, window length: 1024, number of coefficient bins are 124 and 40 for Mel-Spectrogram and MFCC, respectively. Fig. 4 (b)-(d) shows an example of time-frequency representation of audio signal given by Fig.4(a).

### *3.3. Pre-trained CNN Models Adjustment and Classification Stage*

This stage is composed of two steps. In the first step, we applied transfer learning to several pre-trained CNN models, such as Alexnet, Resnet50, Vgg16, Mobilenet, Resnet18 and Densenet, for VGR task. Using pretrained CNN models for audio classification tasks is highly beneficial due to the transfer of learning, the reduction in training time and resources, and the improved performance they offer. These models, trained on large datasets like ImageNet, can detect general features useful for analyzing audio spectrograms, allowing them to be efficiently adapted to new tasks with less labeled data [17]. Additionally, pretrained models such as ResNet, VGG16, MobileNet, and AlexNet, widely validated in the scientific community, provide a high degree of confidence in their performance and results [16]. The conversion of audio signals to spectrograms and the fine-tuning of these models allows for efficient and effective handling of various audio classification tasks, such as gender and age identification. In this step, we applied three time-frequency representations as input data. All these models were trained over the first subset, which is ten times smaller subset than the original one. The subset composes of 10,000 data for training and 1,000 data for testing, with the objective of finding the best results in a small dataset to reduce the trained time and making a preliminary estimation to select the best time-frequency representation and some promising CNN models. In the second step we used the best time-frequency representation and three CNN models with the better performances obtained in the first step. In this second step, full dataset is used to realize fine-tuning to these better models to further improve performance. The metric used to measure performance of the CNNs is accuracy given by

$$Accuracy = \frac{True\ positive + True\ negative}{Total\ number\ of\ datos} \tag{2}$$

The most used loss function for binary classification problems such as VGR is binary cross-entropy loss. The binary cross-entropy loss function takes the actual result (Ground True) and the estimated result provided by the SoftMax activation function and calculates the loss value. The Softmax activation function provides the probability that the system considers the output data to be positive. The optimizer uses a gradient descent method, such as a backpropagation algorithm, to try to reduce the loss value by changing the kernel values of the CNN or the connection weight values of fully connected layers. Adam (Adaptive Moment Estimation) is often used as an optimizer algorithm due to its generally better performance. In the proposed method, we used a binary cross-entropy function, given by (3), and an Adam optimizer with a learning rate of 0.0001, considering the literature information.

$$Cross\ Entropy\ Loss = -(ylog(\hat{y}) + (1-y)\log(1-\hat{y})) \tag{3}$$

Where $y$ is the true label, which can be 0 or 1 and $\hat{y}$ is the model prediction, representing the estimated probability of the instance belonging to the positive class.

Once loss function value is converged, the system considers that the training stage is finished. In the inference, test data is introduced into the adjusted CNNs. The output of this stage is probability of each gender class, which is calculated by Softmax function.

To determine which models, perform best, we conducted a selection process based on their performance. In the first stage, we used the transfer learning technique, modifying only the final layer of the models to match the number of classes in our classification task: male or female. In the second stage, we applied the fine-tuning technique to the top three models from the first stage. These models were frozen from the first layer up to a specific layer, depending on the model selected. Generally, we unfroze the last two layers, as this tends to yield better performance based on the machine used for training. Figure 5 (a) – (c) illustrates the layers unfrozen for each model.

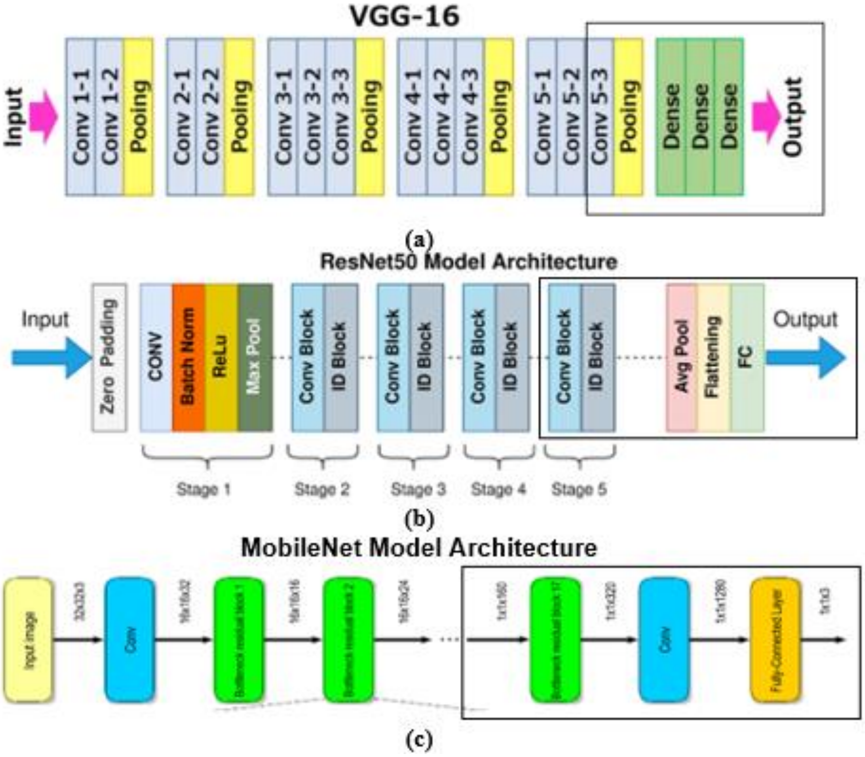


**Figure 5.** (a) Freeze and Unfreeze Layers VGG16, (b) Freeze and Unfreeze Layers ResNet50, (c) Freeze and Unfreeze Layers MobileNet. The part of box of each architecture is re-trained.

## 4. Proposed Method

### *4.1. Dataset*

We used the VoxCeleb dataset developed by the University of Oxford, which are extracted from several uploaded videos in YouTube [7]. We use this dataset because it is the largest dataset with 100,000 audio signals from 1,251 speakers, reflecting real-world audio signals including unwanted noise, background music, other people's voices, laughter, etc. Recording locations range from quiet studios to very noisy outdoor stadiums. In this dataset, the gender classes are nearly balanced, with 55% male speakers and 45% female speakers. Speakers have a variety of characteristics, including nationality, ethnicity, accent, occupation, and age [7].

In this approach, we generated two subsets from the VoxCeleb dataset as mentioned before. The first subset has 10,000 audio signals for training and 1,000 audio signals for testing. The subsets are balanced in male and female gender classes. The second subset is a larger subset consisting of 87,500 audio signals for training and 12,500 audio signals for testing, balanced by gender class like the first subset. The first subset is used to quickly select the best time-frequency representation and three promising CNN models, and the second subset is used to perform fine-tuning on the selected three models.

### *4.2. Results in the First Step*

**Table 1.** Performance of best three models selected in the first step.

| Model | Extracted feature | Accuracy |
|---|---|---|
| Vgg16 | Spectrogram | 93.4 % |
| Vgg16 | Mel-Spectrogram | **95.2 %** |
| Vgg16 | MFCC | 95.1 % |
| MobileNet | Spectrogram | 92.9 % |
| MobileNet | Mel-Spectrogram | 92.4 % |
| MobileNet | MFCC | 92.6 % |
| ResNet50 | Spectrogram | 93.5 % |
| ResNet50 | Mel-Spectrogram | 95.1 % |
| ResNet50 | MFCC | 92.9 % |

This section presents the performance results obtained in the first step. In this step, we use the first subset to select the three CNN models with the better performance. To perform this stage quickly, we trained all CNN models under transfer learning paradigm in just 10 epochs. As a result, we get three promising CNN models: Vgg16, MobileNet, and ResNet50. Table 1 shows the accuracy of these three models using three time-frequency features of the audio signal.

At this step, Vgg16 with Mel-Spectrogram as input feature was the best model with 95.2%, as shown in Table 1. We decided to use the Mel-Spectrogram as the best time-frequency representation of the input data because it gives better results in almost all CNN models.

### *4.3. Results in the Second Step*

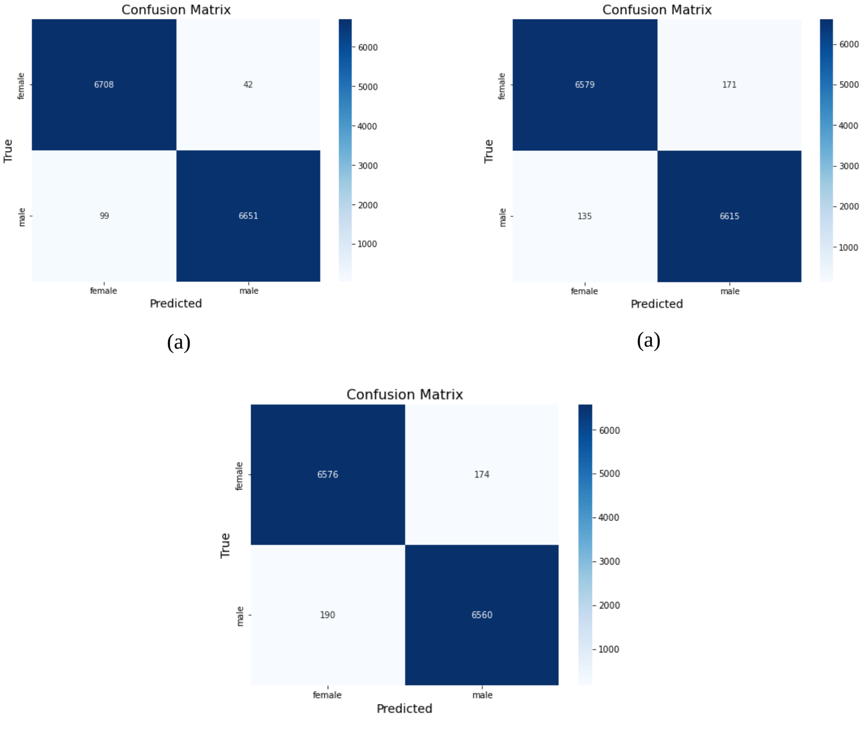
As mentioned before, this step uses the full dataset, and uses Mel-Spectrogram as input data to further improve the performance of three selected models: Vgg16, MobileNet, and ResNet50. Table 2 shows the final performance obtained after this stage.

**Table 2.** Final results after second step using Mel-Spectrogram as input data.

| Model | Total number of parameters | Model Size (Mb) | Accuracy |
|---|---|---|---|
| Vgg16 | 134268738 | 512.19 | 97.30% |
| MobileNet | 414722 | 1.58 | **98.96%** |
| ResNet50 | 23511230 | 376.82 | 97.73% |

From Table 2, we can observe that the performance of all three models improves after further training and fine-tuning using the complete VoxCeleb dataset. In particular, MobileNet shows the best result of 98.96%. Considering that MobileNet is designed for mobile devices such as smartphones, this high performance of MobileNet allows for compact VGR systems that can be implemented in mobile devices.

We show the confusion matrix for each model in the Figs. 6 (a – c).



(a)

(a)

(c)

**Figure 6.** (a) Confusion Matrix of Mobile Net, (b) Confusion Matrix of ResNet50 (c) Confusion Matrix of VGG16.

### *4.4. Performance Comparison*

**Table 3.** A comparison between the proposed system and the state-of-the-art results

| **VGR system** | **Objective** | **Method** | **Accuracy** | **F1-score** |
|---|---|---|---|---|
| Nasef et al. [3]<br>2021 | Gender | MFCC, Attention, LR | 96.23% | ---- |
| Hechmi et al. [15]<br>2021 | Gender and age | MFCC, i-vecor, LR | ---- | 98.29% |
| **Proposed method**<br>2024 | Gender | MS, Fine-tuned MovileNet | **98.96%** | **98.95%** |

LR: Logistic Regression. MS: Mel-Spectrogram

The proposed methodology refers to determine the gender of a person from their voice audio signal. Although some proposals provide good results, most of them used controlled environment datasets, such as Vowel Data Set and TIMIT. In the proposed approach, we preprocessed raw audio in the unconstrained VoxCeleb dataset and carried out fine-tuning several pretrained CNN models to get best model that provides a highest accuracy. Table 3 shows a performance comparison among VGR systems that use VoxCeleb dataset.

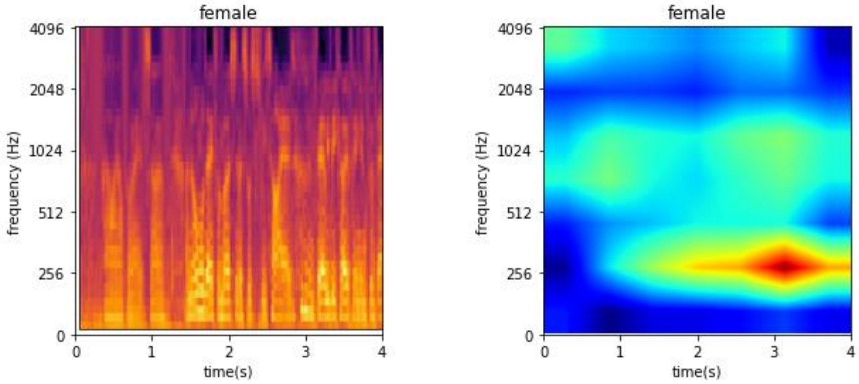
The proposed system achieves a better performance with accuracy of 98.96% and F1-score of 98.95%, compared with the state-of-the-art methods using datasets obtained under unconstrained environments (VoxCeleb). Another advantage of our system is the portability, because MobilenNet is the lightest model of CNN networks, which was designed to be efficient on mobile or resource-constraint devices. Its main goal is to achieve fast and efficient inference in terms of power consumption and model size. Mobilenet achieves this advantage using depth wise separable convolutions which reduces the number of parameters and operations.

### *4.5. Visual Analysis*

In this section, we provide visual analysis about the decision-making performed by the trained CNNs to classify Mel-Spectrogram representation of voice signal into its gender (male or female). For this purpose, we applied GradCAM [8] to the CNNs introducing Mel-Spectrogram as input data. Fig. 7 shows some examples of heatmap image provided by GradCAM together with their corresponded Mel-Spectrogram images. Figs. 7(a) and (b) correspond to a female voice and Figs. 7(c) and (d) corresponds to a male voice.

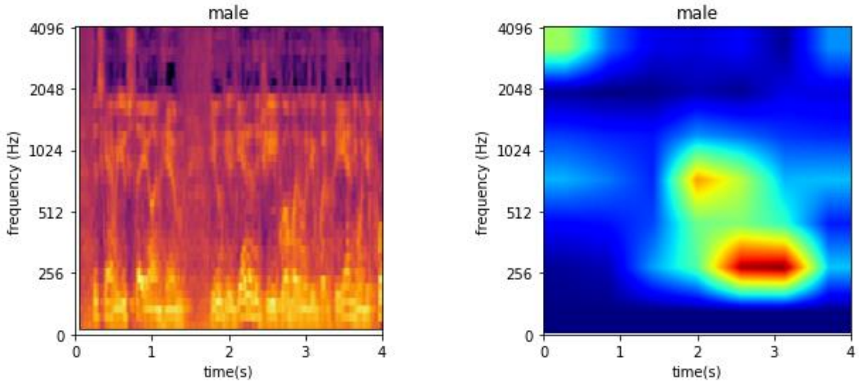
As observed from the figure of heatmaps, the trained CNN pays attention to some frequency bands to determine the gender of voice signal. These frequency bands correspond to some harmonic frequencies such as H1 and H2 of voice signal, which were used for several voice analysis, such as language analysis, speaker analysis and phonetic analysis [9, 10].

All evaluations were performed on GPU, NVIDIA Geforce RTX 3070Ti , and CPU, Ryzen 7, 3700 x 8 cores with 3.60GHz, 32GB of RAM.



(a)

(b)



(c) (d)

**Figure 7.** (a) Mel-Spectrogram of a female voice signal, (b) The region to where CNN pays attention (red region) for its classification, (c) Mel-Spectrogram of a male voice signal, (d) The region to where CNN pays attention (red region) for its classification.

## 5. Conclusions

Voice Gender Recognition (VGR) estimates gender of speaker using his/her voice signals. VGR has been extensively studied by researchers in the last two decades because of its importance and wide range of applications, including commercial and non-commercial fields. Many successful VGR systems have been proposed in a controlled environment where there is not unwanted noise nor other people's voices. However, considering the practical applications of VGR, it is often necessary to be operated in unconstrained environments. In this study, we proposed an efficient VGR system in an unconstrained environment where unwanted noise, background music, and other people's voices are present.

The proposed method consists of four stages. These four stages are a preprocessing stage where noise is attenuated and silence removed from the original signal, a feature extraction stage where time-frequency features such as Spectrogram, Mel-Spectrogram, and MFCC are extracted from the input signal, and CNN training and fine-tuning stage, and finally, the gender classification stage using the trained and fine-tuned CNN.

The proposed method was evaluated using the VoxCeleb dataset, which is the largest speech dataset generated under unconstrained conditions. Intensive evaluation reveals that Mel-Spectrogram is the best time-frequency feature for the proposed VGR, and fine-tuned MobileNet provides the best performance. Considering that MobileNet was developed to work on mobile devices with limited capacity, the proposed VGR system can be efficiently implemented on mobile devices such as smartphones. The result of visual analysis shows the region of the Mel-Spectrogram to where the trained CNN pays attention for classification. These regions correspond to harmonic frequencies H1 and H2, which are used for several speech analyses.

As future work, we may consider including some different features such as X-Vectors and D-Vectors and obtain higher accuracy by retraining the CNN model using these additional features. It is also possible to obtain better results by using deeper CNN networks with fine-tuned techniques, recurrent neural networks, such as LSTM, or Transformer.

Considering that automatic speaker profiling (ASP) is a more powerful tool for many applications of VGR, such as voice-based recommendation systems and voice-based survey systems, an extension of this work to include at least an age estimation mechanism is necessary. Another direction for future research could be to implement the proposed method on mobile devices.

## References

[1] Singh P, Rani P. An approach to extract features using MFCC. IOSR Journal of Engineering. 2014 Aug; 4(8): 21-6.

[2] Jaid, UH, Hassan AKA. Review of Automatic Speaker Profiling: Feature, Methods, and Challenges, Iraqi Journal of Science. 2023 Dec;64(12): 6548-23, doi: 10.24996/ijs.2023.64.12.36.

[3] Nasef, MM, Sauber AM, Nabil MM. Voice gender recognition under unconstrained environments using self-attention. Applied Acoustic. 2021 Apr;175, 107823, doi: 10.1016/j.apacpust.2020.107823.

[4] Ghahremani P, Nidadavolu PS, Chen N, Villalba J, Povey D, Khudanpur S, Dehak N. End-to-end Deep Neural Network Age Estimation. Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH). 2018 Spt. 2-6, Hyderabad, p. 227-4, doi: 10.21437/Interspeech.2018-2015

[5] Pahwa A, Aggarwal G. Speech feature extraction for gender recognition. International Journal of Image Graphics and Signal Processing. 2016 Aug;8(9):17-8, doi:10.5815/ijigsp.2016.09.03

[6] Chaudhary S, Sharma DK. Gender identification based on voice signal characteristics. Proceeding of the International conference on advances in computing, communication control and networking (ICACCCN); 2018 Oct. 12-13, Greater Noida, India, p. 869-6. doi: 10.1109/ICACCCN.2018.8748676.

[7] Nagrani A, Chung JS, Xie W, Zisserman A. Voxcelb: Large-scale speaker verification in the wild, Computer Speech & Language. 2020 Mar;60: 101027, doi:10.1016/j.csi.2019.101027.

[8] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parkh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization, Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22-29; Venice, Italy, p. 618-8 doi. 10.1109/ICCV.2017.74

[9] Petrovsky A, Azarov E. Instantaneous harmonic analysis: Techniques and applications to speech signal processing. In: Ronzhin, A., Potapova, R., Delic, V. (eds) Speech and Computer. SPECOM 2014. Lecture Notes in Computer Science, 8773. Springer, Cham. doi: 10.1007/978-3-319-11581-8_3

[10] Chen G, Feng X, Shue YL, Alwan A. On using voice measures in automatic gender classification of children's speech. Proceedings of the 11 th Annual Conference of the International Speech Communication Association (INTERSPEECH). 2010 p. 673-4, doi: 10.21437/Interspeech.2010-251.

[11] Ertam F. An effective gender recognition approach using voice data via deeper LSTM networks. Applied Acoustics. 2019 Dec;156: 351-7, doi:10.1016/j.apacoust.2019.07.033.

[12] Uddin, MA, Pathan RK, Hossain MS, Biswas M. Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN. Journal of Information and Telecommunication. 2022 Jan;6(1): 27-17, doi: 10.1080/24751839.2921.1983318.

[13] Alnuaim AA, Zakariah M, Shashidhar C, Hatamleh WA, Tarazi H, Shukla PK, Ratna R. Speaker gender recognition based on deep neural networks and ResNet50. Wireless Communications and Mobile Computing. 2022, 4444388, doi: 10.1155/2022/4444388

[14] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomes AN, Kaiser L, Polosukhin I. Attention is all you need, Proceeding of the 31th Annual Conference on Neural Information Processing Systems. 2017 Dec. 4-9, Long Beach, p. 5999-10.

[15] Hechmi K, Trong TN, Hautanmäki V, Kinnunen T. Voxceleb enrichment for age and gender recognition. Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2021 Dec. 13-17. Cartagena, Colombia, p. 687-7. Doi: 10.1109/ASRU51503.2021.9688085.

[16] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*.

[17] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems* (pp. 1097-1105).