Gender classification and speaker identification using machine learning algorithms

Emmanuel de J. Velásquez-Martínez[1],
Aldonso Becerra-Sánchez[2],
José I. de la Rosa-Vargas[3],
Efrén González-Ramírez[4],
Gustavo Zepeda-Valles[5],
Armando Rodarte-Rodríguez[6]
Unidad Académica de Ingeniería Eléctrica
Universidad Autónoma de Zacatecas
Zacatecas, México
[1]iemmanuelvm@gmail.com,
[2]a7donso@uaz.edu.mx,
[3]ismaelrv@ieee.org,
[4]gonzalezefren@uaz.edu.mx,
[5]gzepeda@uaz.edu.mx,
[6]armandorodarte19@gmail.com

Nivia I. Escalante-García[7],
J. Ernesto Olvera-González[8]
Laboratorio de Iluminación Artificial
Tecnológico Nacional de México
Campus Pabellón de Arteaga
Zacatecas, México
[7]aivineg82@gmail.com,
[8]e.olvera.Itp@gmail.com

***Abstract*—The speech is a unique biological feature to each person, and this is commonly used in speaker identification tasks like home automation applications, transaction authentication, health, access control, among others. The purpose of the present work is to compare gender classification and speaker identification experiments in order to determine the machine learning algorithm that shows the best metrics performance based on Mel frequency cepstral coefficients (MFCC) as speech descriptive features. In this process, the machine learning algorithms implemented were logistic regression, random forest, k-nearest neighbors and neural network, which were evaluated with accuracy, specificity, sensitivity and area under the curve. The schemes that revealed the best performance were random forest and k-nearest neighbors, reflecting an AUC (area under the curve) of 1, which indicates that the models have robust capacity of classification both in isolated samples and in complete audio files. The results obtained open guidelines to carry out another type of experimentation using the MFCC features with audios where the environment noise factor is included to measure the performance with these classification algorithms. The experimentation proposed for this work seeks to be applied in the future in different areas, where MFCC are used to describe the voice to perform another type of classification.**

Keywords—gender classification, machine learning algorithms, MFCC, speaker identification.

I. INTRODUCTION

Human beings express their feelings, points of view and notions orally through speech, whose production process includes articulation, voice signals and fluency [1]. Speech is an infinite information signal, thus a direct analysis of it is required due to this fact; therefore, digital signal processes are carried out to represent the voice. Gender classification and speaker identification achieve a task similar to that performed by the human brain; this begins with speech, where generally the classification process is accomplished in three main steps: feature extraction, acoustic processing and classification [2].

Automatic detection of a person gender has many applications from the point of view of automatic speech recognition (ASR), such as classification of telephone calls by gender, in ASR system to improve adaptability, multimedia semantic information, answering machine, automatic dialogue system and other applications. In the case of speaker identification, it is used as a biometric mechanism for identification in some information system, as well as the classification of the person by age range through the voice signal, detection of nationality or language, home automation applications, transaction authentication, access control and other applications [3]. For these identification tasks, an audio feature extraction approach is required to represent the voice signal. There are several feature extractions approaches that usually produce a multidimensional feature vector for the speech signal. Some of the options available to parametrically represent the speech signal are Perceptual Linear Prediction (PLP), Linear Predictive Coding (LPC) and Mel Frequency Cepstrum Coefficients (MFCC) [1].

In this paper, MFCC were proposed as characteristics to represent a voice signal through a predetermined number of signal components. The classification of acoustic observations and audio files is carried out with logistic regression, neural networks, random forest and k-nearest neighbors. To evaluate the performance of each algorithm, accuracy, specificity, sensitivity and ROC curve (specifically the area under the curve) were calculated, where random forest and k-nearest neighbors obtained the best results. We propose a comparison of supervised machine learning algorithms, where the implementation of gender classification and speaker recognition through voice was carried out to perform an active identification. Since this model focuses on audio features, it is independent of language, accent, context, and can perform speaker identification. Before implementing the classification models, an analysis of characteristics to be considered was proposed, this in order to determine the MFCC feature vector length extracted from the audios, determining the minimum number and optimal MFCC to perform both classifications. In addition, since for each audio that represents the speaker, there is a set of vectors with the MFCC characteristics, a classification by sample was carried out, and to improve the performance of the classification algorithms, it was proposed to implement a technique called majority vote, which consists of keeping a count of the feature vectors that were correctly classified with respect to those that were erroneously classified. Where the quality of the identification was experimentally evaluated with an available database. We were

able to classify the genre and identify a speaker given the metrics increased when majority voting was applied.

The rest of the paper is organized as follows, in section 2 related works are presented, in section 3 materials and methods used are described. Section 4 depicts the experimentation and results procedures, while section 5 discusses the conclusions and future work.

II. Related Works

Here, the state-of-the-art of some related works is reviewed, performing classification using MFCC features and machine learning algorithms applied to gender or speaker classification. Under this guideline, Badhon et al. [4] developed a research that carries out the speaker gender recognition based on MFCC and support vector machines, employing a Chinese-speaking database (Mandarin); their main results are summarized with an accuracy of 98.7%. While, Tejale et al. [5] performed gender identification using MFCC using logistic regression, random forest and algorithms with gradient increase; reporting an accuracy of 99.13% in their male/female voices dataset. In addition, Sedaaghi [6] reported a classifier to recognize age, gender and accent, which employed a combination of a Gaussian mixture model classifier, support vector machines and vector quantization. The work used Australian speech data to train and test the system, obtaining an accuracy between 97.96% and 98.68%. In this sense, Kim et al. [7] reported an age and gender group recognition system that can be used for human-robot interaction by using support vector machines and decision trees classifiers taking as input MFCC and LPCC (Linear Predictive Cepstral Coefficient) features. On the other hand, Rajeshet al. [8] developed a DNN-based (Deep Neural Network) speaker recognition system using standard MFCC, normalized power cepstral coefficients and perceptual linear prediction. This approach was based on speaker embeddings and probabilistic linear discriminant analysis. By the same line, Nguyen et al. [9] did experiments for gender classification purposes using superposition, elongation and the reference acoustic feature, as well as MFCC observations. In this research support vector machines and recurrent neural networks were used to classify the genre. Results showed an accuracy of 89.61% with RNN using a set of functions that includes MFCC, overlay and lengthening at the same time.

On the other hand, there are more specific works for speaker identification tasks, e.g. Luque-Suárez et al. [10] modeled the speaker as a high-dimensional point cloud of entropy-based features extracted from the voice signal, modeling the classification using k-nearest neighbors. The work reported an accuracy of 97% when the recording environment is not controlled, and 99% for controlled recording environments. Along the same line Bose et al. [11] established the combination of two different feature sets such as MFCC and LPCC and the use of ensemble classifiers together with principal component transformation and GMM, using on it the NTIMIT reference speech corpus. They carried out two experiments varying the amount of audio per speaker for the training and test phase. The results of combining the features in both experiments improved the performance of the algorithm. They reported a baseline of 72.3% for the first experiment varying the amount of audio, that was in a 6:4 speaker dataset ratio; while 67.3% under an 8:2 ratio. Besides, Nagrani et al. [12] proposed a fully automated pipeline based on computer vision techniques to create the dataset from open-source media. They obtained the audios through the YouTube video platform, carrying out the active verification of the speaker using a convolutional neural network (CNN) and as network training characteristics, where each audio was extracted from its spectrogram. CNN reported two accuracy metrics given the data set they proposed, so their accuracies were 80.5% and 92.1%. Furthermore, Adetoyi's paper [13] presented a text-independent speaker identification system that employs MFCC for feature extraction and k-nearest neighbor (KNN) for classification. The maximum cross-validation precision they obtained was 60%. While Afonja et al. [14] performed a proposal on audio classification models, where the feature extraction used was the naive extraction approach for each audio. They aim to mimic the behavior of the victim model trained to identify a speaker, proposing the use of a generative model to create a sufficiently large and diverse pool of synthetic attack queries. They achieved a test accuracy of 84.41% with a total of 3 million queries for the model. Finally, Shahin et al. [15] developed experiments to capture Emirati-accented speech in each of the neutral and noisy conversational environments in order to improve speaker identification performance in noisy environments. They carried out experiments with hidden Markov models and MFCCs as features. Their results showed that the average identification performance of speakers with an Emirati accent in a neutral environment was 94%, 95.2% and 95.9%, respectively; while, the mean performance in the noisy environment was 51.3, 55.5 and 59.3%.

III. Materials and Methods

A. Methodology

Figure 1 shows the general flow of the methodology used to classify the gender and speaker identification through the proposed algorithms. Phase 1 generates the feature vectors that describe the input audio, depicting the genre and the speaker; three data sets were generated with different numbers of MFCC, where each feature vector is labeled according to the genre as well as the speaker to which it belongs. Subsequently, the MFCC are analyzed through the learning curve and selection of forward features with the aim of determining the number of minimum MFCC features to apply a learning algorithm, which obtains the best result. In phase 2 the machine learning algorithms are implemented (logistic regression, random forest, k-nearest neighbors and neural network). In phase 3 the algorithms are compared by means of accuracy, specificity, sensitivity and area under the curve. Finally, in phase 4 an analysis of the performance of each learning algorithm is carried out to determine the best score.

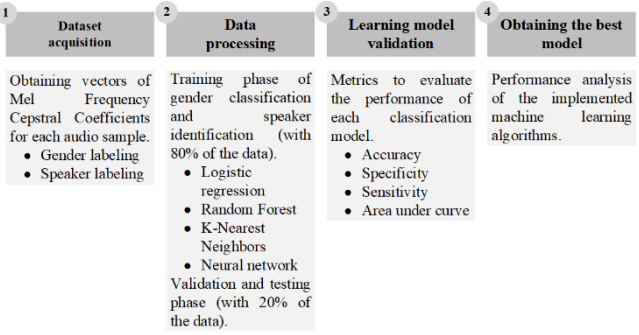

Figure 1. Methodology flowdata for data processing and classification.

B. Dataset Acquisiton

The first step, as shown in Figure 2, parameterizes the speaker speech signals. Each input file was digitized in

discrete values with PCM encoding at 16 bits per sample and with a sampling rate of 16khz in a single channel. As a next step, the initial and final silence in the audio is removed from each digital signal.

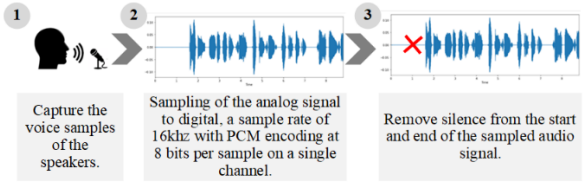

Figure 2. Voice signal pre-processing.

The second step is to extract MFCC vectors from the digital signals, representing their salient features. The procedure involved to extract the MFCC coefficients is shown in Figure 3 [1]; where these steps are:

- Pre-emphasis consists of rewarding the high-frequency part that was suppressed when the speaker produced the sound, making the information from these higher formants more available to the model.
- Windowing is dividing the voice signal into 30 to 20 millisecond windows with an optional 1/3 to 1/2 frame size overlap.
- The windowed speech signal is facilitated with the Hamming window to remove discontinuities in the signal.
- The Discrete Fourier Transform (DFT) obtains how much energy the signal contains in different frequency bands; where Mel filters will now be applied.
- Finally, the Inverse Fourier Transform (IFT) is applied with the Mel scale log cepstrum.
- The features obtained are known as MFCC, which are arbitrary and depend on the parameters that are applied during the described process.

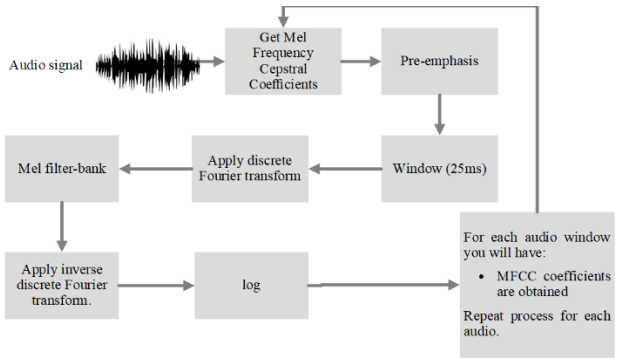

Figure 3. MFCC features extraction process of input speech.

C. Data Processing

In Figure 4 the steps to obtain a classification model are shown. Procedure begins with input MFCC, subsequently, the data are divided into partitions to carry out a cross-validation (k=3); 80% for training and 20% for testing phase. Next step is to obtain the models, in this case logistic regression, neural network, k-nearest neighbors and random forest. For gender classification, the four classifiers were applied, later, based on the metrics obtained for our first classification, k-nearest neighbor, random forest and the neural network were applied to the speaker identification task. Continuing with the workflow, a scheme has to iteratively fit the model with the proper parameters in order to obtain the best classification.

Once these parameters are evaluated, we proceed to assess the model and generate the approach that allows testing data and measure the performance with accuracy, specificity, sensitivity and area under the curve. Thus, it can be determined which models were more optimal for gender prediction and speaker identification based on MFCC features.

D. Classification Models and Metrics

- Logistic regression is an analysis method for classification problems, where it tries to determine if a new sample fits better in a category [16]. This algorithm was implemented only to classify speakers gender; its implementation parameters were based on a saga solver, penalty of l1 and C=0.01 as regularization.

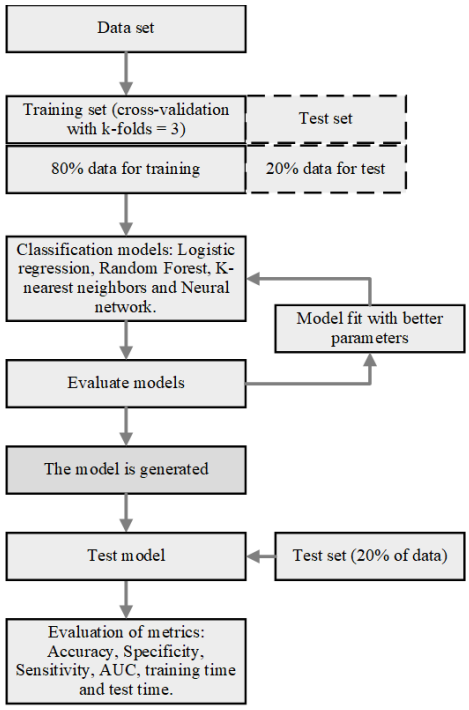

Figure 4. Workflow to obtain the classification models.

- Neural networks are made up of node layers, where each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node wakes up and sends data to the next network layer [16]. In this research, a neural network was used for speaker gender classification and speaker identification. The architecture used has 20 neurons in the input layer, 2048 densely connected neurons in the hidden layer with relu, while softmax activation functions were used in the output layer; where the outputs provided by the classes corresponding to the classification. In addition, 20 epochs were assigned with a block size of 40.
- A random forest is an ensemble learning method for classification, regression and other tasks that operates by building a multitude of decision trees at the time of training. For the classification tasks, the output of the random forest is the class selected by the most trees

[16]. Random forest was used for gender classification and speaker identification tasks. The random forest algorithm was implemented with 1000 trees that the algorithm builds before averaging the predictions of 1000 and with a log2 for the maximum number of features that the random forest considers to divide a node.

- The k-nearest neighbors algorithm is a machine learning algorithm that can be used for both regression and classification tasks. K-nearest neighbors examines the labels of a chosen number of data points that surround a target data point in order to make a prediction about the class to which the data point belongs [16]. This algorithm was used to perform the gender classification and speaker identification tasks. In this, different numbers of k-neighbors were evaluated, this to determine the amount that delivers a better result. For which they were evaluated with k = 2, ..., 9, being k = 9 who delivered a higher accuracy, employing the Manhattan distance as loss metric.

Once the predictor models were obtained, validation metrics were used with the aim of measuring the performance of each model. The indicators used were [16]:

- Accuracy is the fraction of predictions that were correct by our model, i.e.:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (1)$$

- Sensitivity is a proportion measure of actual positive cases that were predicted positive (see (2)). This implies that there will be another proportion of actual positive cases, which would be incorrectly predicted as negative.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

- Specificity is defined as the proportion of actual negatives that were predicted to be negative (see (3)). This implies that there will be another proportion of true negatives, which were predicted as positives and could be called false positives.

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad (3)$$

- An ROC (receiver operating characteristic) curve is a graph that shows the performance of a classification model at all its thresholds. This plots two parameters, the true positive rate and the false positive rate.
- The area under the curve (AUC) provides an aggregate measure of performance at all possible classification thresholds. AUC is shows the probability that the model will rank a random positive example higher than a random negative example.

IV. Experiments and Results

A. Dataset Description

The classification task has been developed using a personalized mid-vocabulary voice corpus of connected-words in Spanish from the northern central part of Mexico in a closed set environment. With the purpose of strengthening the scope of the voice corpus, it was complemented with utterances from audio files generated through online text-to-speech applications with similar natural Mexican sounding voices. The online applications used were ispeech, oddcast (SitePal) and vocalware. In addition, the age of the human participants ranges from 18 and 26. Some of the audio files were recorded using slight ambient noise to robust the dataset.

The speech corpus originally contains 158 classes in 4911 audio files. For our analysis, the corpus was subsampled, taking only a total of 70 classes (speakers, 35 men and 35 women), this due to the balance of classes by gender. Only 20 audio samples with different phrases were taken from each speaker. For the feature selection, 3 data sets were generated; for case 1, a data set with 39 characteristics was generated (13 MFCC, 13 deltas and 13 double deltas); for case 2, 39 MFCC were extracted; while for case 3, 20 MFCC were extracted. In our case, the data set was divided into 80% for the training phase and 20% for the test phase.

B. Software and Hardware

Python was used for the audio processing, employing the librosa library to extract the MFCC features. This is a Python module for audio and music analysis, where it provides the basic components needed to create audio information retrieval systems [17]. For the implementation of the learning algorithms, the Scikit-learn module was implemented, which is a Python library that provides supervised and unsupervised learning algorithms [18]; while for the implementation of the neural network architecture, TensorFlow was used. In addition, the hardware used for this analysis was a Dell Alienware m15, with 2.30 GHz CPU and 16GB RAM running Windows 11.

C. Datasets Features Analysis

In Figure 5 the behavior of the AUC metric for 39 features (13 MFCC, 13 deltas and 13 double deltas) is shown. A total of 39 logistic regression models were generated. Figure 5 shows how the model with the first 13 MFCC provides a higher performance to the classifier, obtaining an AUC slightly higher than 0.75. Since the features corresponding to MFCC deltas and double MFCC deltas were added, the performance of the AUC metric did not show a significant change, where finally the maximum AUC reached was 0.78.

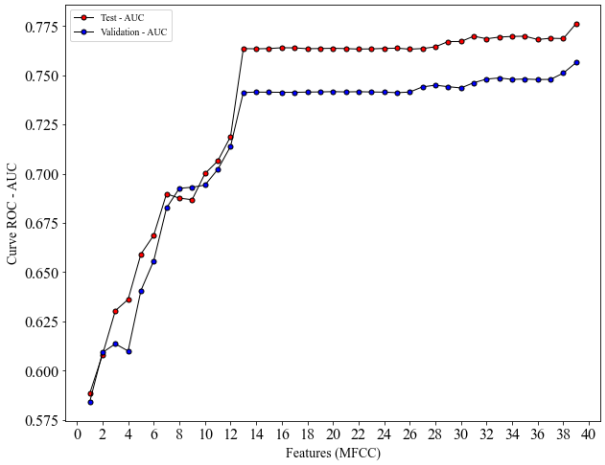

Figure 5. Learning curve with 13 MFCC + 13 deltas + 13 double deltas.

Figure 6 shows the behavior of AUC with 39 MFCC features, where the AUC shows better performance when the classification model is performed with 39 MFCC without applying the first and second derivatives, reaching a maximum area under the curve of 0.90. Although the graph shows the best performance with 39 observations, from feature 20, the AUC does not have a significant increase.

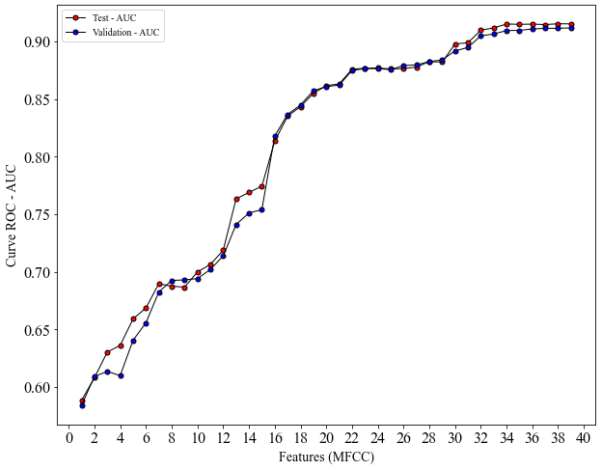

Figure 6. Learning curve with 39 MFCC.

Moreover, Figure 7 shows the behavior of the AUC metric for each logistic regression model applied to each coefficient that was added. Here, the AUC using 20 MFCC was 0.8732, which compared to the model with 39 MFCC, this does not have a considerable increase. Therefore, it was determined that 20 MFCC are sufficient to implement each classification model for both gender and speaker identification.

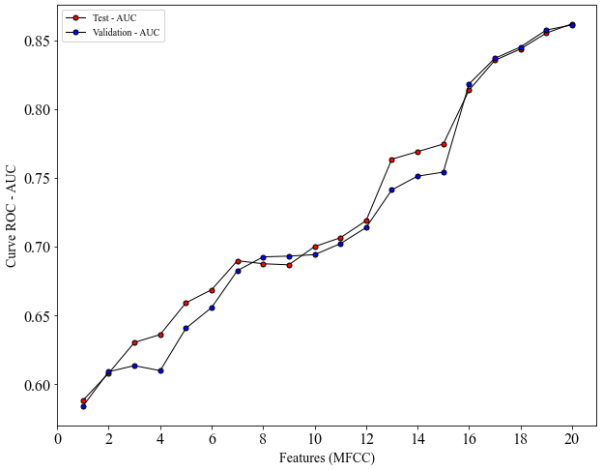

Figure 7. Learning curve with 20 MFCC.

D. Speaker gender classification results with isolated samples

The models with the lowest performance in speaker gender classification were the logistic regression and the neural network (see Table I). On the other hand, k-nearest neighbors and the random forest achieved the highest performance indexes for each metric. For accuracy, 1 was obtained, indicating that the speaker gender samples are being correctly classified. For the specificity and sensitivity metrics, 1 was obtained, which indicates that the speaker gender is being distinguished between male and female. These results offer a interpretation of a core robust framework, i.e. tree-based approaches can present novel variants in pattern classifications environments.

TABLE I. TEST OF SPEAKER GENDER CLASSIFICATION

Classification Model	Speaker Gender Classification			
	Accuracy	*Sensitivity*	*Specificity*	*AUC*
Logistic regression	0.8617	0.8192	0.9031	0.8612
Random forest	1.0	1.0	1.0	1.0
K-nearest neighbor	1.0	1.0	1.0	1.0
Neural network	0.9112	0.8387	0.9819	0.9103

In Fig. 8, 9, 10 and 11 the ROC curves for gender classification are shown. This analysis is a statistical method to determine the accuracy for the test set of each model, indicating the highest sensitivity and specificity, which describes the ability to differentiate the speaker gender. The analysis is achieved by comparing the area under the curve, where it has a value between 0.5 and 1, where 1 represents a perfect classifier and 0.5 indicates that the model does not have the capacity to classify. K-nearest neighbors and random forest presented a greater area under the curve, with AUC of 1. These values are higher compared to the logistic regression and neural network models, which were 0.8612 and 0.9103, respectively. The four classifiers exceed the AUC threshold of 0.5, which indicates that they are statistically significant when carrying out this type of classification, having the ability to distinguish the samples. This provides a good model addressed to future configurations, in which several considerations can be applied with the purpose of determine a correct scheme in this type of tasks.

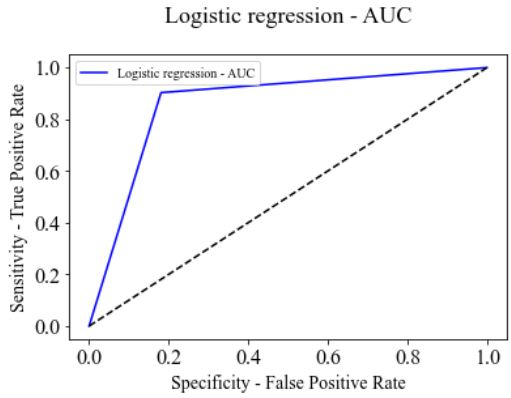

Figure 8. Logistic regression learning curve.

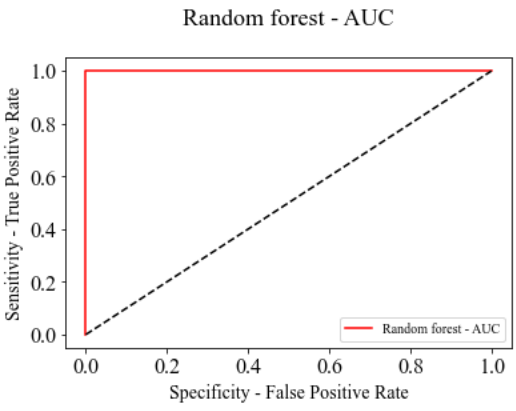

Figure 9. Random forest learning curve.

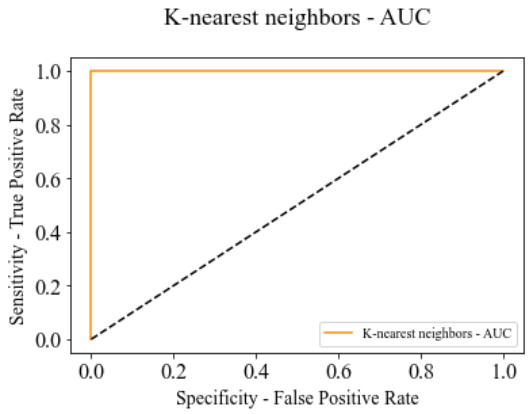

Figure 10. K-nearest neighbors learning curve.

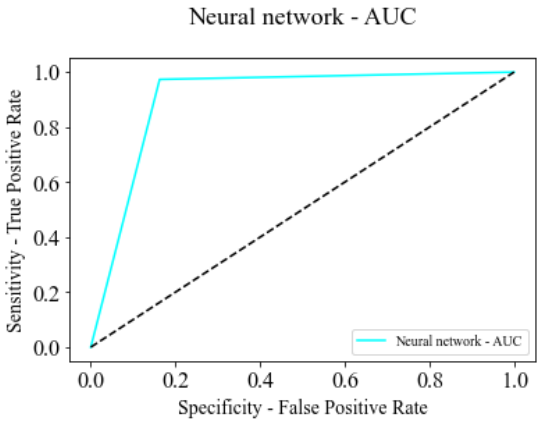

Figure 11. Neural network learning curve.

Models training and testing time for gender classification is shown in Table II, here, the schemes with the longest time are random forest and k-nearest neighbors; while the fastest algorithm was logistic regression.

TABLE II. MODELS COMPUTING TIME FOR GENDER CLASSIFICATION (SECONDS)

Classification Model	Gender Classification Time	
	Training	*Testing*
Logistic regression	80.27	0.01
Random forest	15209.69	48.80
K-nearest neighbor	5493.31	2197.27
Neural network	3494.70	57.16

E. Speaker identification results with isolated samples

Table III shows the average of each metric for 70 classes, where the classifiers with the lowest performance were the neural network and logistic regression. On the other hand, the random forest model and the k-nearest neighbors classifier obtained a performance of 1 in all its metrics, being better classifiers with respect to the previous models. Meanwhile Table IV shows the results of training and testing time for speaker identification, k-nearest neighbors and random forest spent the most time, while neural networks and logistics regression spent the least.

F. Majority vote results to classify complete audio files

Table V shows the results applying the majority vote for the complete audio gender classification. Compared to the results in Table I (isolated samples), the performance metrics of the logistic regression and neural network learning algorithms showed an improvement by increasing their performance. Whereas that Table VI shows the results applying the majority vote to classify complete audios of each speaker. The results obtained in this stage slightly improve the algorithms metrics, such as logistic regression and neural network.

TABLE III. RESULTS FOR SPEAKER IDENTIFICATION

Classification Model	Speaker Gender Classification			
	Accuracy	*Sensitivity*	*Specificity*	*AUC*
Logistic regression	0.8267	0.9974	0.8261	0.9936
Random forest	1.0	1.0	1.0	1.0
K-nearest neighbor	1.0	1.0	1.0	1.0
Neural network	0.7089	0.7887	0.9969	0.8548

TABLE IV. COMPUTING TIME FOR SPEAKER IDENTIFICATION (SEC)

Classification Model	Gender Classification Time	
	Training	*Testing*
Logistic regression	4684.53	0.28
Random forest	36686.36	913.76
K-nearest neighbor	5445.38	2184.57
Neural network	1584.09	30.26

TABLE V. TEST RESULTS FOR SPEAKER GENDER CLASSIFICATION – MAJORITARY VOTE

Classification Model	Speaker Gender Classification			
	Accuracy	*Sensitivity*	*Specificity*	*AUC*
Logistic regression	0.9821	1.0	0.9642	0.9821
Random forest	1.0	1.0	1.0	1.0
K-nearest neighbor	1.0	1.0	1.0	1.0
Neural network	1.0	1.0	1.0	1.0

TABLE VI. TEST RESULTS FOR SPEAKER IDENTIFICATION – MAJORITARY VOTE

Classification Model	Speaker Identification			
	Accuracy	*Sensitivity*	*Specificity*	*AUC*
Logistic regression	0.9571	0.9571	0.9993	0.9782
Random forest	1.0	1.0	1.0	1.0
K-nearest neighbor	1.0	1.0	1.0	1.0
Neural network	0.7142	0.7142	0.9958	0.8550

G. Discussion of the Proposed Model

Unlike the related works, our proposed model performs two tasks, the speaker gender classification, as well as the speaker identification, analyzing the performance of machine learning algorithms in both processes, where the learning algorithms were carried out by tuning hyperparameters. The proposed learning models with the highest performance were KNN and random forest, reaching a performance of 1 in their metrics for gender classification and speaker identification. On the other hand, to improve the performance of the logistic regression algorithms and the neural network architecture, it was proposed to implement the majority voting technique to increase their performance, showing a considerably good increase for both classifications. An analysis of the number of MFCC features was performed before training, where the

Forward Selection algorithm was implemented, which consists of adding feature by feature in a regression algorithm and measuring, by means of a metric, the performance obtained by adding each MFCC. In related works, Kim et al. [7] and Bose et al. [11] proposed the combination of MFCC and LPCC for machine learning algorithms, while other works also used similar types of features. Given that MFCC features are proposed in the state of the art, feature analysis was proposed to implement only this feature extraction process. In the future, we would like to test and improve the performance of these algorithms or even implement deep neural network architectures and add filtering techniques to improve performance in noisy environments. For example, as discussed in the work of Shahin et al. [15], where they reported poor performance in technique learning when speech treatment is conducted in noisy and uncontrolled environments. And also, since voice audio can be easily combined with this type of noise factors, the proposal of a new model for handling these tasks would be ideal for implementation in ASR systems.

A direct quantitative comparison between our results regarding the state-of-the-art cannot be equivalent, since they use different datasets, with different recording circumstances and noise conditions.

V. Conclusions

This article proposed an approach to gender classification and speaker identification using MFCC vectors and learning algorithms. In the first instance, the number of features was determined for each sample of the speaker audio, this preliminary analysis helped to select the number of effective features to obtain a better performance with the least number of possible features that would allow the classification. Where 20 MFCC were used to generate models obtaining an optimal performance, in addition to this analysis, the calculation and computational time for each algorithm are reduced. The algorithms that showed the best performance in classifying samples and complete speaker audios (using majority vote) for both gender classification and speaker identification were k-nearest neighbors and random forest, showing an AUC of 1. As future work, it will be interesting to analyze scenarios with noise factor in input samples.

Although the MFCC features are acceptable with machine learning classifiers, we should try to improve and compare the metrics with additional components in speech, such as external ambient noise and even other signal processing techniques and learning classifiers. The results obtained in the experimentation generates guidelines to apply this type of experimentation in another area. For instance, classifying the people age through voice samples; besides, mood classification could be another alternative. Interesting application areas of this tasks can be in the health sector, identifying any speech-related pathology.

Acknowledgment

We thank CONACYT for their support on this work.

References

[1] D. Jurafsky and J. Martin, Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd ed, Upper Saddle River, NJ: Pearson Prentice Hall, 2008.

[2] R. Kumar, R. Ranjan, S. K. Singh, R. Kala, A. Shukla, and R. Tiwari, “Multilingual speaker recognition using neural network,” Proceedings of the Frontiers of Research on Speech and Music (FRSM), pp. 1-8, December 2009.

[3] G. Manjula and M. S. Kumar, “Stuttered speech recognition for robotic control,” International Journal of Engineering and Innovative Technology (IJEIT), vol. 3, no. 12, June 2014.

[4] S. M. S. I. Badhon, M. H. Rahaman and F. R. Rupon, “A Machine Learning Approach to Automating Bengali Voice Based Gender Classification,” in 8th International Conference System Modeling and Advancement in Research Trends (SMART), 2019, pp. 55-61, doi: 10.1109/SMART46866.2019.9117385.

[5] S. S. Tejale and T. B. Kute, “Performance evaluation of algorithms for gender classification,” International Journal of Engineering Applied Sciences and Technology, 2020, vol. 4, no. 11, pp. 568-573.

[6] M. H. Sedaaghi, “A comparative study of gender and age classification in speech signals,” Iranian Journal of Electrical & Electronic Engineering (IJEEE), March 2009.

[7] H. Kim, K. Bae, and H. Yoon, “Age and Gender Classification for a Home-Robot Service,” in RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication, 2007, pp. 122-126, doi: 10.1109/ROMAN.2007.4415065.

[8] S. Rajesh and N. J. Nalini, “Combined Evidence of MFCC and CRP Features Using Machine Learning Algorithms for Singer Identification,” International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI). July 2020. doi: 10.1142/S0218001421580015.

[9] P. Nguyen, D. Tran, Xu Huang, and D. Sharma, “Automatic classification of speaker characteristics,” in International Conference on Communications and Electronics (ICCE), 2010, pp. 147-152, doi: 10.1109/ICCE.2010.5670700.

[10] F. Luque-Suárez, A. Camarena-Ibarrola, and E. Chávez, “Efficient speaker identification using spectral entropy,” Multimed. Tools Appl., vol. 78, no. 12, pp. 16803–16815, Jun. 2019, doi: 10.1007/s11042-018-7035-9.

[11] S. Bose, A. Pal, A. Mukherjee, and D. Das, “Robust Speaker Identification Using Fusion of Features and Classifiers,” Int. J. Mach. Learn. Comput., vol. 7, no. 5, pp. 133–138, Oct. 2017, doi: 10.18178/ijmlc.2017.7.5.635.

[12] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: a large-scale speaker identification dataset,” 2017, doi: 10.48550/ARXIV.1706.08612.

[13] O. E. Adetoyi, “Text Independent Speaker Identification System for Access Control,” 2022, doi: 10.48550/ARXIV.2209.14335.

[14] T. Afonja, L. Bourtoule, V. Chandrasekaran, S. Oore, and N. Papernot, “Generative Extraction of Audio Classifiers for Speaker Identification,” 2022, doi: 10.48550/ARXIV.2207.12816.

[15] I. Shahin, A. B. Nassif, and M. Bahutair, “Emirati-accented speaker identification in each of neutral and shouted talking environments,” Int. J. Speech Technol., vol. 21, no. 2, pp. 265–278, Jun. 2018, doi: 10.1007/s10772-018-9502-0.

[16] A. Geron, Hands-On Machine Learning with Scikit-Learn and Tensor-Flow: Concepts, Tools, and Techniques to Build Intelligent Systems, 3rd ed, [S.l.]: O’Reilly, 2019.

[17] B. McFee, A. Metsai, et al., librosa/librosa: 0.9.2. Zenodo, June 2022.

[18] L. Buitinck, G. Louppe, et al., ”API design for machine learning software: experiences from the scikit-learn project”, in ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp 108–122.