# 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