ISSN 2007-9737

Gender Recognition of Teen and Adult Voices in Non-Tonal and Tonal Languages in Uncontrolled Environments

Enrique Díaz-Ocampo[1], Areli Karina Martínez-Tapia[1,*], Andrea Magadán-Salazar[2], Raúl Pinto-Elías[2], Máximo López-Sánchez[2], Yael Bensoussan[3]

[1] Colegio Universitario Científico de Datos/COCID, Mexico

[2] Tecnológico Nacional de México/CENIDET, Mexico

[3] University of South Florida, Department of Otolaryngology-Head & Neck Surgery, United States of America

enrique.diaz@cocid.edu.mx, areli.martinez@cocid.edu.mx, andrea.ms@cenidet.tecnm.mx, raul.pe@cenidet.tecnm.mx, maximo.ls@cenidet.tecnm.mx yaelbensoussan@usf.edu

Abstract. Voice gender recognition systems is a term that refers the automatization of gender detection by an acoustic signal of voice. These systems can be trained in uncontrolled environments, whose audios present different types of noises and speaker characteristics. However, the current systems present a bias in the training language, which is usually mainly English. The present work focused on the gender recognition of adult and teen voices in a group of tonal languages and Spanish under uncontrolled environments. The features used were 7 derived from pitch, and two from the mean of the fourth formant and vocal tract length. Two scenarios were built: a training-test scenario on one dataset, and a second validation scenario using the other dataset. The metrics used were accuracy, recall, F1-score, and area under the ROC curve. The algorithms used were Multilayer Perceptron and Random Forest. Despite the bias in the datasets, the biological features and the algorithms were robust to language change.

Keywords. Voice gender recognition, fundamental frequency, vocal tract length, tonal language, Spanish language.

1 Introduction

A person's gender can be studied through two approaches: essentialist or constructivist [29]. In the former, biological traits are considered decisive in distinguishing one gender from another. The word gender is used as a synonym for sex. In the constructivist approach, gender is perceived as a social construction and is associated with certain expectations, conditioning factors, and customs of the social niche to which the person belongs. Systems that automate gender recognition by voice use the essentialist approach. Due to the main characteristic is the differences present in the biology of the vocal tract as well as in their vocal cords [26].

Gender recognition by voice can be classified into two environments: controlled and uncontrolled. The former is usually delimited by a protocol and medical purposes. Recorded voices use specialized microphones. An example of this is the work [6]. On the other hand, it is also usual that these voices only pronounce certain types of phrases. Such is the case of

[1]. Where they proposed a new feature that they called modified voice contour. It consists of the area under the curve of the voice contour for a standardized phrase. This characteristic is based on the differences present between the lengths of the vocal cords between men and women [23, 26].

Voice Gender recognition systems in uncontrolled environments are characterized by the fact that the audios analyzed present the following variations:

— Technical aspects: Recording device, quality of the recording device and audio duration.

— Environment-related: Presence of multiple environmental noise.

— Speaker-related: Whether the speaker reads or is a natural sentence, the age, health, emotional state, accent and language of the person.

The present work focuses on gender recognition in adults and adolescents in tonal languages and in the Spanish language. Subsequently, it validates each trained and tested model in one language with the untrained language. The relevance of this work lies in the following points:

— Two datasets for gender recognition were constructed one in Spanish and the other being a mixture of various tonal languages. Both are available [4, 5].

— The recognition of gender in adults and teens is a variant of the recognition of age. Therefore, a classifier of the adulthood of a voice stacked with an age recognition system will improve its performance.

— The use of biological or biologically robust features in languages allows for a more interpretive study of speakers. Thus, models can be constructed that can perform better in gender recognition.

— Mitigate the bias present in gender recognition systems trained and tested in a single language by disseminating datasets and research in multi-language and uncontrolled environments.

This paper is organized as follows. Section 2 will present a brief state of the art in gender recognition under uncontrolled environments. Given the limited variety of gender recognition work in languages such as Spanish and tonal languages, the state of the art is mainly based on the English language. However, the methodology of gender recognition is similar despite the difference in languages. Section 3 is the Methodology, which is divided into preprocessing, feature extraction, classifiers, and train, test and validation, and finally experiments. Section 4 discusses the results and discussions of them in training, test and validation experiments in both datasets. Finally, Section 5 presents the conclusions and future work.

2 State of the Art

A widely used database for representing this type of environment is Mozilla Common Voice (Mozilla) [3]. Multiple recognitions of gender [1], gender and age [21, 25], gender and accent [2] have been made in this database. However, the approach used in these works is through deep learning, which requires a greater number of elements to be trained and presents a greater computational complexity. Such is the case of [25], where gender and age recognition is performed through a convolutional neural network and a multi-attention module. English language was used and the audios were grouped into the labels F-teens, F-twenties, F-thirties, F-forties, F-fifties, M-teens, M-twenties, M-thirties, M-forties, M-fifties, and M-sixties. Resulting in an accuracy of 76%. Another example is that of [21]. In this work, the classification of gender and age group was carried out using 18 different architectures of temporal convolutional neural networks. The Mozilla English corpus was divided into four classes Young-Male, Young-Female, Adult-Male, Adult-Female, Senior-Male and Senior-Female. Among the 18 architectures, number 4 with 90092 parameters, 81% precision and 76% recall had the best performance.

One of the most widely used corpus from mozilla for gender recognition in the English language is [16]. Among the first works that started to use it is [28]. It focused not only on gender recognition

(male and female), but also on the recognition of age in the categories of young (under thirty years old), matured (between thirty and fifty years old) and old (over fifty years old). Furthermore, by using audios from the Ryerson Audio- Visual Database of Emotional Speech and Song (RAVDESS) [20], their work was able to implement an emotion recognizer in the following categories: happy, sad and angry. Their work consisted of analyzing 6247 audios from the Common voice 5.1 corpus [16] (for gender and age analysis) and 1440 from the RAVDESS corpus (for emotion analysis). Each of these audios were analyzed with Frequency Spectrum Analysis (FSA) [14] and 20 features were extracted. Subsequently, several machine learning algorithms were trained and tested to choose the ones that best detected each of the three features (gender, age and emotion).

Briefly, both the training and the test were used percentage split (80% train set and 20% test set) as 10-fold cross validation. Of all the algorithms used, CatBoost obtained 96.4% and 95.4% accuracy in gender recognition, using percentage split and 10-fold cross validation respectively. In the case of age, the best model was Random Forest with 70.4% accuracy in percentage split and CatBoost with 61.7% in 10-fold cross-validation. Finally, for emotion recognition, XGBoost obtained 66.1% and 58.7% accuracy in percentage split and 10-fold cross-validation respectively.

A variant of gender recognition but using the same dataset is through deep learning algorithms as in [11]. By using cepstral features and a neural network consisting of five dense layers with 512, 256, 128, 128, and 128, respectively. Then, 64 neurons using ReLU as a nonlinear activation function. Consecuently, each dense layer used 30% of dropout. For the last dense layer, two output neurons were used for sex recognition using a sigmoid function. With this system, it was possible to obtain an accuracy of 94.32%.

There has been previous work exploring gender recognition in languages other than English from Mozilla. Such is the case of [10]. By using a Convolutional Neural Network (CNN) as a classifier and a set of cepstral features (Mel-frequency cepstral coefficients (MFCC), Mel-spectrogram and Chroma), significant accuracy metric was obtained in several languages: Irish (98%), Russian (97.5%), Swedish (96.7%), Japanese (94.1%), Chinese (China) (93.5%), German (92.5%) and Chinese (Hong Kong) (88.5%). Thus, there is evidence that there are features that are robust to languages under uncontrolled environments. However, in this work, a validation of performance in a language other than the trained language was not performed.

Despite the results obtained throughout the gender recognition in the English language. There is no significant presence of gender recognition work in languages such as Spanish nor in tonal languages i.e. languages where different tonal inflections can change the word's meaning [27]. On the other hand, one of the problems following gender recognition is the performance of the system in a new language. One way in which the problem can be solved is by studying the biological characteristics of the speakers that are robust in the presence of language diversity. An example of these characteristics is fundamental frequency, i.e. the frequency at which the vocal cords vibrate. This is statistically lower in adult men than in women. Nevertheless, it is not a significant classifying feature when the voices are grouped by age. However, the length of the vocal tract, that is, the length from the space between the vocal cords (called the glottis) to the lips, is a statistically different features between teens and adults [19, 24, 13].

3 Methodology

This paper presents a model of gender recognition of teens and adults by voice. The model consists of three phases: preprocessing, feature extraction and classification. An overview of the system architecture can be seen in Figure 1. The stages of the model will be described in the following subsections.

3.1 Preprocessing

The analysis of voiced and unvoiced speech is essential for gender recognition processes. In this work, the Python programming language was used for voice processing using the Parselmount-Praat

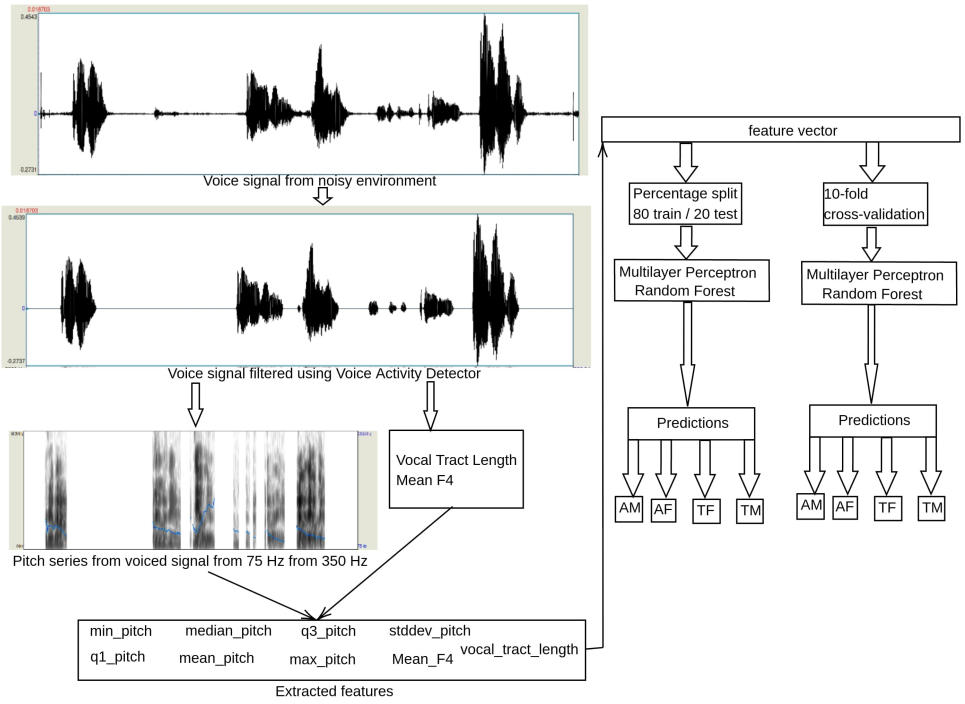


Fig. 1. Proposed methodology for gender recognition. The labels AM, AF, TM and TF denote Adult Male, Adult Female, Teen Male and Teen Female, respectively

library [17]. This library uses the algorithms implemented in the Praat software [9]. The extraction of the voiced parts of each audio follows the Boersman algorithm [9, 12] and autocorrelation pitch calculation proposed in [8]. Also,the parameters used here will be based on those used in [22]. The procedure is the following:

1. A voiced speech detection filter is used throughout the audio. The script first creates a point process object from the sound object. A text grid object is then created from this point process object with information on the voiced/unvoiced parts. Consequently, two copies of the original sound objects are created, one with the voiced parts silenced and one with the unvoiced parts silenced, using the voiced-unvoiced information from the text grid object. For this work, only the copy with voiced speech is considered for the next step.
2. The four standard parameters are provided to calculate a series of pitch candidates per window.
   - Time step (t_s): It is the frame duration in seconds. PRAAT analyzes:
   $$t_s = \frac{0.75}{P_f}, \tag{1}$$
   audios samples per second. The variable P_f is the Pitch floor.
   - Pitch floor (P_f): It is the frequency threshold (in Hertz). Frequencies bellow P_f will not be considered. The standard value used was set on 75 Hz.
   - Window length(W_L): The length of the analyzed window in seconds. W_L is described by the following equation:
   $$W_L = \frac{3}{P_f}. \tag{2}$$
   For this analysis, $W_L = 0.04s$.

— Very accurate: If the value of this option is set to off, the window is a Hanning window with a length of W_L. If is it on, the window is a Gaussian window with a length of $2W_L$. The value used was on.

3. A post-processing algorithm seeks the cheapest path through the candidates according to a functional proposed in [8]. The settings that determine the cheapest path are:

— Pitch ceiling (P_c): Candidates above this frequency will be ignored. The standard value used was $P_c = 350Hz$.

— Silence threshold (S_t): Frames that do not contain amplitudes above this threshold (relative to the global maximum amplitude), are probably silent. The standard value used was $S_t = 0.03$.

— Voicing threshold V_t: The strength of the unvoiced candidate, relative to the maximum possible autocorrelation. If the amount of periodic energy in a frame is more than this threshold, then the frame is considered as a voiced frame; otherwise as unvoiced frame. The standard value was $V_t = 0.45$.

— Octave cost (O_c per octave): Degree of favouring of high-frequency candidates, relative to the maximum possible autocorrelation. The standard value was $O_c = 0.01$ per octave.

— Octave-jump cost (Q_j): Degree of disfavouring of pitch changes, relative to the maximum possible autocorrelation. The value used was $Q_j = 0.35$.

— Voiced/Unvoiced cost (U_c): Degree of disfavouring of voiced/unvoiced transitions, relative to the maximum possible autocorrelation. The value used was $U_c = 0.14$.

3.2 Feature Extraction

The features can be divided into two groups: derived from the pitch and the length of the vocal tract. For the first features, the minimum pitch (min_pitch), first quartile of pitch (q1_pitch), mean pitch (mean_pitch), median pitch (median_pitch), third quartile of pitch (q3_pitch), maximum pitch (max_pitch) and standard deviation of pitch (stddev_pitch) were extracted using the Pitch contours extracted in the preprocessing phase. In the case of vocal tract length (VLT), it was determined by the expression:

$$VTL = \frac{(2n-1)c}{4F_n}, \tag{3}$$

where F_n is the n formant of the human vocal tract, c is an approximation of the speed of sound ($35000\frac{cm}{s}$). This equation is derived from modeling the vocal tract as a tube [18]. For the purposes of this paper, for each of the audio windows containing voice speech, the fourth formant F_4 was calculated and then averaged (mean_F4). Finally, this value was substituted into Equation 3 to obtain a estimation of the vocal tract length (vocal_tract_length). In general, Equation 3 is used for specific vowel sounds in languages. However, It is possible to consider the whole audio as a single sound by averaging F_4 and set maximum formant as 5000 Hz for men, 5500 for women, and 8000 Hz for teens. The above was proposed in this way, since this calculation can be generalized to any sentence spoken by the speaker. It is obvious that it will not approximate the actual value of the person, but statistically it shows a difference between adults and childrens. This gives the vector of 9 features and the labels Adult Male (AM), Adult Female (AF), Teen Male (TM) and Teen Female (TF) as is shown in Figure 1.

3.3 Classifiers

The classifiers used in this work are as follows:

— Multilayer Perceptron: A classifier that uses backpropagation to classify instances. The nodes in this network are all sigmoid. The parameters used were the standard WEKA parameters: Learning rate was set to 0.3, Momentum was set at 0.2, Number of Epochs was set to 500 (see Figure 2).

— Random Forest: Random forest is a robust, general-purpose algorithm that can be used for both regression and classification tasks.

This algorithm works by training multiple decision trees on different subsets of the data, using a technique called bootstrap aggregating, or bagging, to create a diverse set of trees. This diversity helps to reduce overfitting. By averaging or voting the predictions from all the trees, the random forest is able to make more robust predictions than an individual decision tree. WEKA standard parameters were used: Max Depth was set to 0, number of features was set to 0, number of trees was set to 100, and seed was set to 1.

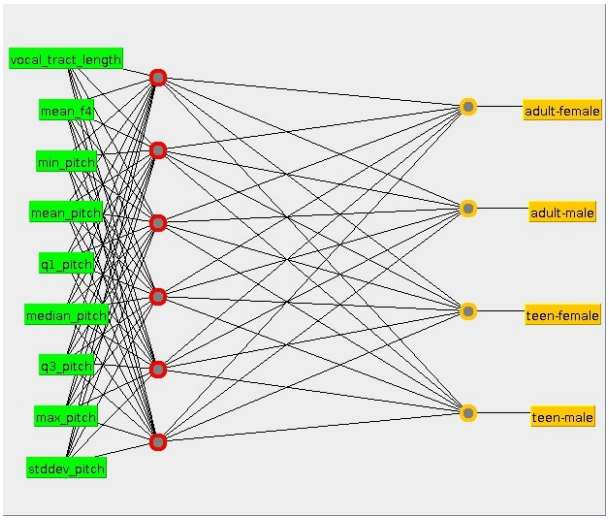


Fig. 2. Multilayer perceptron network architecture. It consists of an input layer of 9 inputs, a hidden layer of 6 neurons with sigmoid activation function, and an output layer with 4 neurons with sigmoid functions per activation function

3.4 Train, Test and Validation

Within uncontrolled environments, it is usual to find datasets with significant imbalances [21]. One way is to use techniques such as SMOTE [7], which allow the construction of artificial instances. In this work, the imbalances in the datasets were preserved to check the separability of the features obtained. In addition, two types of tests were used: Percentage split and cross validation. The Weka software [15] allows automated learning to be performed using the percentage split cross-validation options. The first option consists on divided the dataset into two: traininig set of 80% of the original dataset and test set with the remaining 20%. The second option consists of that the data set is randomly divided into n parts. Then, $n-1$ of those parts are allocated for training and one part is reserved for testing. This procedure is repeated n times and each time a different test set is reserved for testing. For this work, 10-fold cross-validation (90% training set and 10% test set) was used. Finally, for the 10 classification performed, a weighted average of the performance metrics is made. For this work, the metrics used were accuracy, recall (R), F1-score (F1), and Area under ROC curve (ROC Area). In addition, the normalized confusion matrices of the results obtained in each classifier are shown.

3.5 Experiments

The proposed experimentation is illustrated in Figure 3. The first step consists in the choice of an algorithm (MLP or RF). Then, the selected algorithm is trained and tested on one of the datasets (Tonal or Spanish) using one of the two proposed options (10-fold cross validation or Percentage split). Finally, a validation is performed with the dataset that was not chosen for training and testing. To evaluate the eight models obtained, the metrics Recall (R), F1-score (F1), Area under the ROC curve (ROC Area) and accuracy were chosen.

The analyzed audios were taken from Mozilla Common Voice [3]. Tonal languages considered were Thai, Vietnamese, Punjabi and variants of Chinese (China, Hong Kong and Taiwan). In the case of Non-Tonal language, Spanish was considered. A complete description of each language can be found at Table 1. The tonal audios dataset (207622 audios with a mean duration of 4.08 seconds with a standard deviation of 1.64 seconds) consist of 32.95% adult females (AF), 60.24% adult males (AM), 5.30% teen males (TM), and 1.51% teen females (TF). Spanish dataset (213477 audios with a mean duration of 5.06 seconds with a standard deviation of 1.51 seconds)

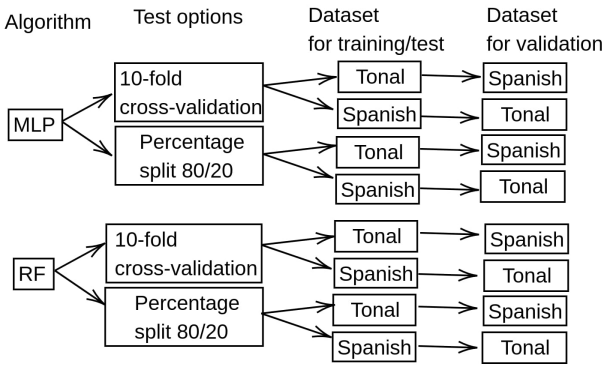


Fig. 3. General scheme of the proposed experiment

has 70.8% AM, 25.21% AF, 2.47 TM, and 1.52% TF.

Figures 4a and 4b show the scatter plot of the proposed characteristics in the two datasets. The statistics derived from the fundamental frequency provides a degree of separation between males and females. However, it does not separate the gender of teens. In the case of vocal tract length and the mean of fourth formant, it does generate a more visible separation in the four classes.

Figure 5a and Figure 5b show the correlation matrix of the chosen characteristics. It can be seen that VTL has an inverse relationship with all the other characteristics. In the case of the statistics derived from pitch and mean F_4, they present a positive correlation. This diagram shows the relevance of VTL.

Since the maturation of the vocal folds is achieved until adulthood, so the frequencies emitted by a child could be within a higher range and could be confused by the classifier. However, the length of their VTL will be significantly shorter than that of an adult.

4 Results and Discussion

For the discussion section, it was divided into four subsections: Results training and testing in the same language, and results training with one language and testing with a different one.

4.1 Results in Training and Test in Tonal Languages

The results of the experimentation in the training and test of gender recognition can be found at Table 2. Given the bias in the datasets constructed, 8 models were built with their respective algorithms and test forms (Percentage split with 80% train and 20% test and 10-fold cross-validation). It is highlighted that the models that used the Random Forest algorithm obtained better accuracy metrics.

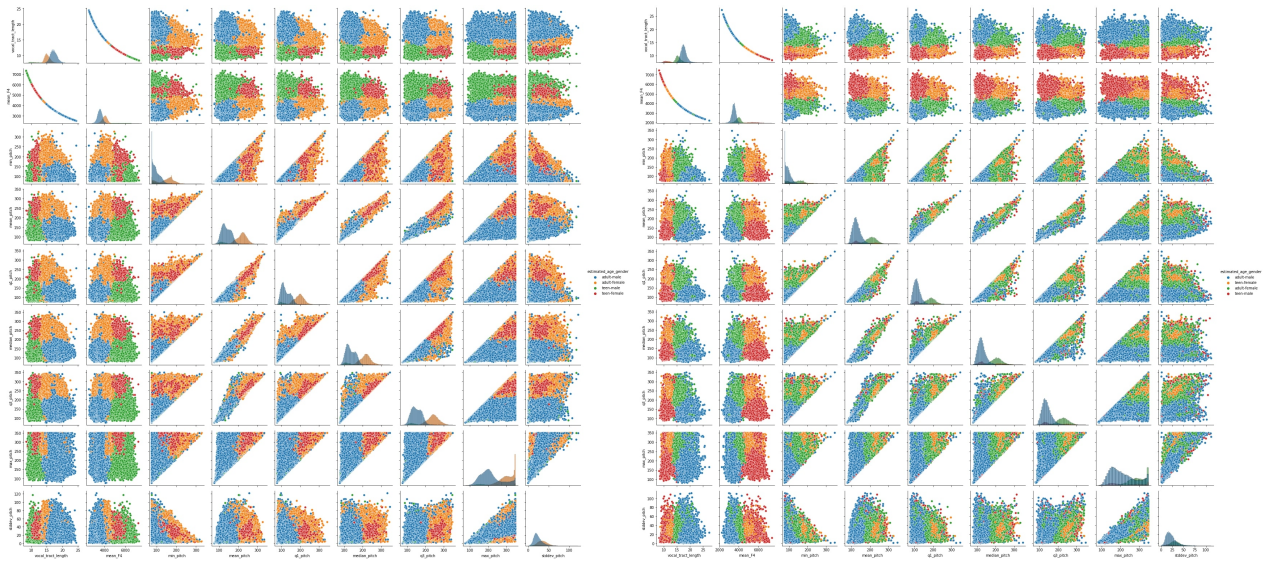
Given the large number of examples of adult voices, the four models of tonal were expected to obtain higher metrics in the adult classes compared to the teen classes. Thus, one way to compare the models is to study detection in teens. Under this approach, for the percentage split models, MLP 80-20 trained-tonal obtained better F1-score metrics (74.1% in F1-TF and 92.2% in F1-TM) than RF 80-20 trained-tonal. On the other hand, RF 80-20 trained-tonal obtained better Recall metrics (75.4% in R-TF and 93% in R-TM). In the case of cross validation models, RF 10-cross trained-tonal obtained better metrics in both F1-score (77.7 % in F1-TF and F1-TM 93.1 %) and Recall (73.7 % in R-TF and 91.6 % in R-TM).

4.2 Results in Training and Test in Non-Tonal Languages

In the case of Spanish (see Table 2), both in the percentage split and cross validation models, RF obtained better metrics in F1-score and Recall. In particular, RF 80-20 trained-spa obtained 88.7 % in F1-TF and 90.2 % in F1-TM and 73.7 % in R-TF and 91.6 % in R-TM. RF 10-cross trained-spa obtained 88.4 % in F1-TF and 88.7 % in F1-TM and 88.7 % in R-TF and 90.2 % in R-TM.

4.3 Results of the Models Trained in Tonal and Validated in Spanish

The results of the classifiers can be found in Table 3. Among the highlights of these results is the fact that there was no significant difference (in terms of metrics) of a specific algorithm model tested in percentage split and in cross validation (see Table 3 and Figure 6). On the other hand, RF 80-20 trained-tonal valid-spa obtained the highest



(a) Scatter diagram of the features in Tonal languages.

(b) Scatter diagram of the features in Spanish.

Fig. 4. Scatter diagram of both datasets

Table 1. Distribution of the audios of adult females (AF), adult males (AM), teen males (TM), and teen females (TF) in the two datasets

Languages	TM Audios	AM Audios	TF Audios	AF Audios	Sum of Audios	TM Speakers	AM Speakers	TF Speakers	AF Speakers	Total of Speakers
Spanish	5277	151133	3254	53813	213477	290	2637	145	892	3964
Tonal languages	11000	125076	3125	68421	207622	275	1867	152	877	3171
Description of Tonal Languages										
Thai	2535	47912	1860	26876	79183	72	417	100	353	942
Chinese (china)	4077	21288	234	4632	30231	135	666	22	149	972
Chinese (HONK KONG)	1134	30688	298	20226	52346	24	325	11	190	550
Chinese (Taiwan)	3220	20410	597	16575	40802	40	416	16	177	649
Vietnamese	34	3579	136	110	3859	4	31	3	7	45
Punjabi	0	1199	0	2	1201	0	12	0	1	13
Sum of Tonal languages	11000	125076	3125	68421	207622	275	1867	152	877	3171

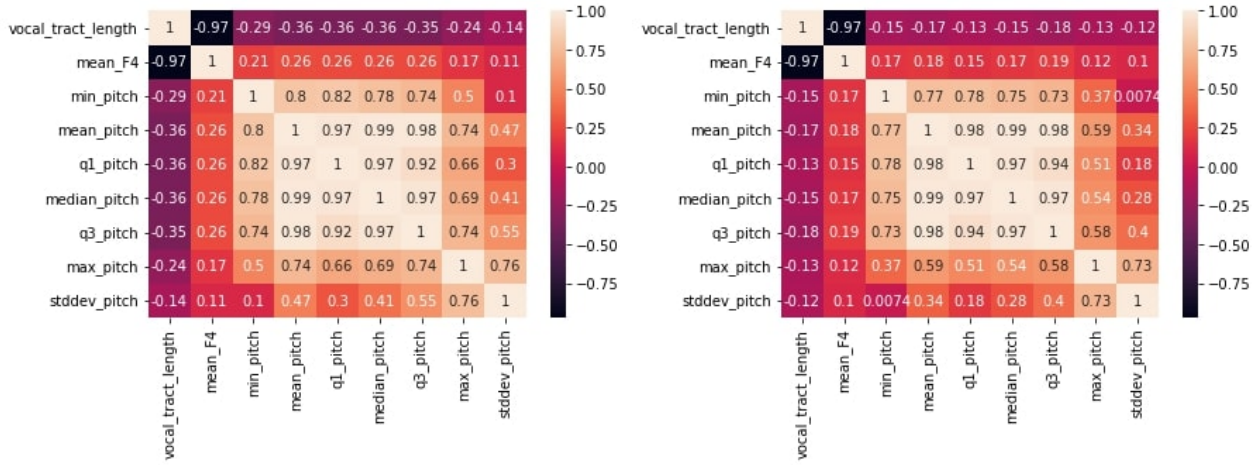
accuracy (93.72%), which is due to the fact that it recognized a larger number of adults.

This can be deduced by analyzing the R-AF, R-AM and F1-AF F1-AM values. However, it was decided to make this phenomenon explicit by analyzing the normalized confusion matrices (see Figure 4a) where 80.14% of AF and 99.28 % of AM were recognized correctly.

It is worth noting that although RF obtained better gender recognition in adults, MLP recognized adolescents better. MLP recognized 82.94% of TF and 96.07% of TM correctly.

4.4 Results of the Models Trained in Tonal and Validated in Tonal Languages

The models trained in Spanish and subsequently validated in the tonal languages provided similar behavior in their gender recognition results (see Table 3). To begin with, there was no significant difference between percentage split and cross validation training. Next, there was no significant difference in the F1-score metric in adolescent recognition for the MLP 80-20 trained-spa valid-tonal (70.4% in F1-TM and 91%F1-TF) and RF 80-20 trained-spa valid-tonal



(a) Correlation matrix of the features of Tonal language dataset. **(b)** correlation matrix of the features of Spanish language dataset.

Fig. 5. Correlation matrix of both datasets

Table 2. Distribution of metrics Recall (R), F1-score (F1), Area under the ROC curve (ROC Area) in the four classes Adult female (AF), Adult Male (AM), Teen female (TF) and Teen male (TM) in the training-test scenario

Classifier	R-AF	R-AM	R-TF	R-TM	F1-AF	F1-AM	F1-TF	F1-TM	ROC Area-AF	ROC Area-AM	ROC Area-TF	ROC Area-TM	Accuracy
MLP_80-20_trained-tonal	0.947	0.977	0.712	0.889	0.946	0.974	0.741	0.922	0.991	0.993	0.985	0.993	0.9590
RF_80-20_trained-tonal	0.949	0.978	0.757	0.93	0.948	0.975	0.72	0.907	0.991	0.993	0.965	0.994	0.9607
MLP_10-cross_trained-tonal	0.944	0.978	0.719	0.905	0.946	0.973	0.767	0.924	0.99	0.993	0.985	0.993	0.9591
RF_10-cross_trained-tonal	0.949	0.979	0.737	0.916	0.951	0.975	0.777	0.931	0.991	0.993	0.973	0.995	0.9621
MLP_80-20_trained-spa	0.958	0.987	0.837	0.919	0.948	0.99	0.891	0.873	0.99	0.992	0.996	0.994	0.9755
RF_80-20_trained-spa	0.964	0.998	0.87	0.933	0.963	0.989	0.887	0.902	0.996	0.996	0.993	0.993	0.9792
MLP_10-cross_trained-spa	0.960	0.987	0.851	0.91	0.958	0.989	0.861	0.868	0.99	0.992	0.996	0.988	0.9761
RF_10-cross_trained-spa	0.965	0.989	0.869	0.923	0.964	0.99	0.884	0.887	0.996	0.996	0.992	0.991	0.9792

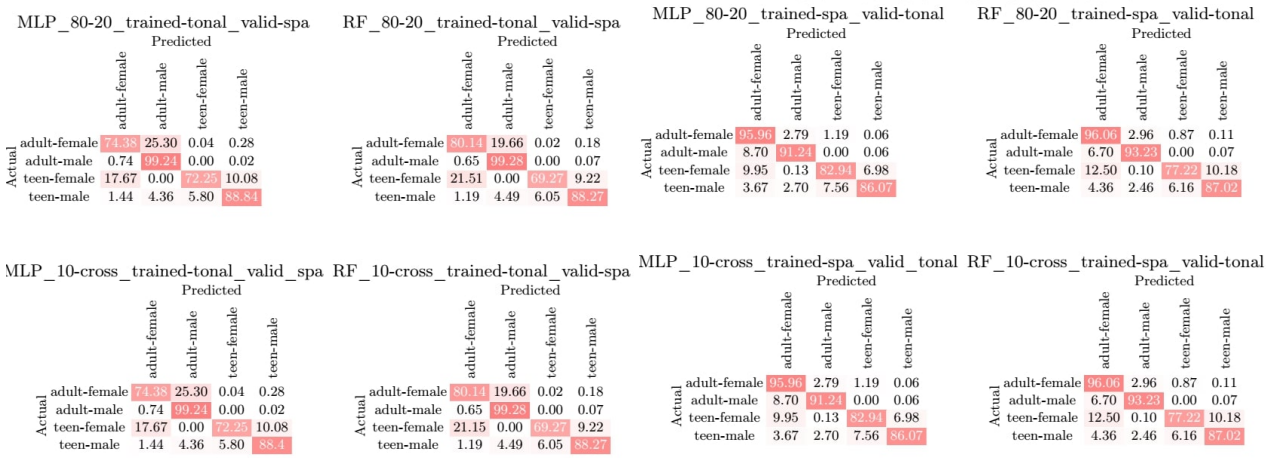
(70.8% in F1-TM and 90.9%F1-TF) models. Finally, analyzing the confusion matrices (see Figure 6b), we can determine that RF 80-20 trained-spa valid-tonal best detected adults (96.06% AF and 93.23% AM) and 87.02% TM, while MLP 80-20 trained-spa valid-tonal detected 82.94% TF. 87.02% of TM, while MLP 80-20 trained-spa valid-tonal detected 82.94% of TF.

5 Conclusions and Future Work

Gender recognition by voice in adults and teens under uncontrolled environments is an open problem due to three types of aspects: those related to the recording equipment (technical aspects), those related to the recording environment, and those related to the speaker.

While progress has been made in the English language, the wide variety of languages present in Mozilla Common Voice have yet to be analyzed. The present work focused on gender recognition of adults and teens from Spanish and Tonal languages of the corpus of Mozilla Common Voice. The classes were Adult-Male, Adult-Female, Teen-Male, Teen-Female. Nine features (7 derived from the fundamental frequency and the last two were mean of fourth Formant and the vocal tract length) in two different datasets were used. The training-test scenarios were studied in a single dataset and their subsequent validation in the second dataset. The metric used to evaluate them were the recall, F1-Score, area under de ROC Curve and accuracy.

The results obtained show that the statistics derived from the pitch as well as the fourth formant and the length of the vocal tract were robust to language change. Futhermore, the estimation of the vocal tract by averaging the fourth formant



(a) Normalized confusion matrix of models validated in Spanish dataset.

(b) Normalized confusion matrix of models validated in Tonal dataset.

Fig. 6. Normalized confusion of the models tested in one dataset and validated in the second dataset

Table 3. Distribution of metrics Recall (R), F1-score (F1), Area under the ROC curve (ROC Area) in the four classes Adult female (AF), Adult Male (AM), Teen female (TF) and Teen male (TM) in the validation scenario

Classifier	R-AF	R-AM	R-TF	R-TM	F1-AF	F1-AM	F1-TF	F1-TM	ROC Area-AM	ROC Area-AF	ROC Area-TF	ROC Area-TM	Accuracy
MLP_80-20_trained-tonal_valid-spa	0.744	0.922	0.722	0.888	0.838	0.952	0.793	0.895	0.978	0.974	0.996	0.992	0.923092
RF_80-20_trained-tonal_valid-spa	0.801	0.993	0.693	0.883	0.874	0.962	0.772	0.893	0.982	0.98	0.973	0.986	0.937244
MLP_10-cross_trained-tonal_valid-spa	0.744	0.992	0.722	0.888	0.838	0.952	0.793	0.895	0.978	0.974	0.996	0.992	0.923092
RF_10-cross_trained-tonal_valid-spa	0.801	0.993	0.693	0.883	0.874	0.962	0.772	0.893	0.982	0.98	0.973	0.986	0.937244
MLP_80-20_trained-spa_valid-tonal	0.96	0.912	0.829	0.861	0.901	0.945	0.704	0.91	0.986	0.973	0.984	0.992	0.923957
RF_80-20_trained-spa_valid-tonal	0.961	0.932	0.772	0.87	0.917	0.956	0.708	0.909	0.987	0.981	0.967	0.985	0.935892
MLP_10-cross_trained-spa_valid_tonal	0.96	0.912	0.829	0.861	0.901	0.945	0.704	0.91	0.986	0.973	0.984	0.992	0.923957
RF_10-cross_trained-spa_valid-tonal	0.961	0.932	0.772	0.87	0.917	0.956	0.708	0.909	0.987	0.981	0.967	0.985	0.935892

along the ventans with voiced speech was shown to be a discriminant feature in the detection of adult and adolescent voices in uncontrolled environments. This is novel, because it opens up research into gender recognition by voice using other biological features like height.

On the other hand, despite the bias in the number of female speakers from the datasets, metrics were obtained that were superior to 72% in F1-TF in both the training-test and validation scenarios. The confusion matrices showed how the RFs performed better in the adult group than in the teens group. While MLPs detected teens better than adults.

In principle, the RF had the highest accuracy metrics, which did not imply a better performance in the recognition of all classes. This suggests that, in the face of such a significant imbalance as presented in this work, the confusion matrix analysis will show the weaknesses of the classifiers.

In future work, we hope to be able to use a suitable combination of cepstral and biological features for gender, age, and accent recognition in a single system and label it as follows accent-age in decades-adult or teen-Male or Female. An example would be Cuban-twenties-adult-male.

Acknowledgments

Enrique Díaz-Ocampo (CVU 919756) would like to thank CONACYT (Consejo Nacional de Ciencia y Tecnología) for the financial support of his Master studies under Scholarship.

References

1. **Alhussein, M., Ali, Z., Imran, M., Abdul, W. (2016).** Automatic gender detection based on characteristics of vocal folds for mobile healthcare system. Mobile Information Systems, Vol. 2016. DOI: 10.1155/2016/7805217.
2. **Angadi, D., R, M. K., S, N. N., Kumar, N. B. (2021).** Voice based age, accent and gender recognition. .
3. **Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., Weber, G. (2019).** Common voice: A massively-multilingual speech corpus. CoRR, Vol. abs/1912.06670.
4. **Autors (2023).** Tonal languages from Mozilla common voice 10. https://goo.su/VL7U. Accessed: 2023-01-16.
5. **Autors (2023).** Voice gender recognition in Spanish language. https://goo.su/T3lt8p. Accessed: 2023-01-16.
6. **Bensoussan, Y., Pinto, J., Crowson, M., Walden, P. R., Rudzicz, F., Johns III, M. (2021).** Deep learning for voice gender identification: Proof-of-concept for gender-affirming voice care. The Laryngoscope, Vol. 131, No. 5, pp. E1611–E1615.
7. **Blagus, R., Lusa, L. (2013).** SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics, Vol. 14, No. 1, pp. 106. DOI: 10.1186/1471-2105-14-106.
8. **Boersma, P. (1993).** Acurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.. IFA Proceedings 17, Vol. 17, pp. 97–110.
9. **Boersma, P., Weenink, D. (2001).** PRAAT, a system for doing phonetics by computer. Glot international, Vol. 5, pp. 341–345.
10. **Chachadi, K., Nirmala, S. R. (2022).** Gender recognition from speech signal using 1-D CNN. **Gunjan, V. K., Zurada, J. M.**, editors, Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications, Springer Nature Singapore, Singapore, pp. 349–360.
11. **Chachadi, K., Nirmala, S. R. (2022).** Voice-based gender recognition using neural network. Lecture Notes in Networks and Systems, Vol. 191, No. Ictcs 2020, pp. 741–749. DOI: 10.1007/978-981-16-0739-4$_7$0.
12. **Corretge, R. (2012-2022).** Praat vocal toolkit. https://www.praatvocaltoolkit.com.
13. **Fitch, W. T., Giedd, J. (1999).** Morphology and development of the human vocal tract: A study using magnetic resonance imaging. The Journal of the Acoustical Society of America, Vol. 106, No. 3, pp. 1511–1522.
14. **Fulop, S. A. (2011).** Speech spectrum analysis. Springer Science & Business Media.
15. **Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. (2008).** The WEKA data mining software: An update. SIGKDD Explor. Newsl., Vol. 11, pp. 10–18.
16. **Henretty, M., Kamp, T., Davis, K. (2017).** Common voice. https:www.kaggle.comdatasets mozillaorgcommon-voice. Accessed: 2023-01-16.
17. **Jadoul, Y., Thompson, B., de Boer, B. (2018).** Introducing Parselmouth: A Python interface to Praat. Journal of Phonetics, Vol. 2018, pp. 1–15. DOI: 10.1016/j.wocn.2018.07.001.
18. **Johnson, K. (2004).** Acoustic and auditory phonetics. Phonetica, Vol. 61, No. 1, pp. 56–58.
19. **Lammert, A. C., Narayanan, S. S. (2015).** On short-time estimation of vocal tract length from formant frequencies. PloS one, Vol. 10, No. 7, pp. e0132193.
20. **Livingstone, S. R., Russo, F. A. (2018).** 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, Vol. 13, No. 5, pp. e0196391.

21. **Sánchez-Hevia, H. A., Gil-Pita, R., Utrilla-Manso, M., Rosa-Zurera, M. (2022).** Age group classification and gender recognition from speech with temporal convolutional neural networks. Multimedia Tools Appl., Vol. 81, No. 3, pp. 3535–3552. DOI: 10.1007/s11042-021-11614-4.
22. **Shagi, G. U., Aji, S. (2022).** A machine learning approach for gender identification using statistical features of pitch in speeches. Applied Acoustics, Vol. 185, pp. 108392. DOI: 10.1016/j.apacoust.2021.108392.
23. **Skuk, V. G., Schweinberger, S. R. (2014).** Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. .
24. **Turner, R. E., Walters, T. C., Monaghan, J. J., Patterson, R. D. (2009).** A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. The Journal of the Acoustical Society of America, Vol. 125, No. 4, pp. 2374–2386.
25. **Tursunov, A., Mustaqeem, Choeh, J. Y., Kwon, S. (2021).** Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. Sensors, Vol. 21, No. 17. DOI: 10.3390/s21175892.
26. **Wu, K., Childers, D. G. (1991).** Gender recognition from speech. part I: Coarse analysis. The journal of the Acoustical society of America, Vol. 90, No. 4, pp. 1828–1840.
27. **Xu, L., Zhou, N. (2011).** Tonal languages and cochlear implants. In Auditory prostheses. Springer, pp. 341–364.
28. **Zaman, S. R., Sadekeen, D., Alfaz, M. A., Shahriyar, R. (2021).** One source to detect them all: Gender, age, and emotion detection from voice. Proceedings - 2021 IEEE 45th Annual Computers, Software, and Applications Conference, COMPSAC 2021, pp. 338–343. DOI: 10.1109/COMPSAC51774.2021.00055.
29. **Zimman, L. (2018).** Transgender voices: Insights on identity, embodiment, and the gender of the voice. Language and Linguistics Compass, Vol. 12, No. 8, pp. 1–16. DOI: 10.1111/lnc3.12284.

Article received on 01/02/2023; accepted on 09/12/2024.
Corresponding author is Areli Karina Martínez-Tapia.