# Analysis of Voice Data towards Biometric Identification: Gender, Age and Identity Classification

**João Carlos Ramos Gonçalves de Matos,** *up201704111*
**Maria Jorge Miranda Loureiro,** *up201704188*
**Maria Manuel Domingos Carvalho,** *up201706990*

*Abstract*—**Voice data plays an important role in some high-tech applications nowadays. In this work, features are extracted from a voice audios dataset and, upon those, Machine Learning classification strategies are developed, towards biometric identification of gender, age, and identity. For gender classification, both linear SVM and Random Forest classifier show good performance, with a F1-Score of about 76%. For age classification, a combined strategy of Ridge Regression with costs, followed by a mapping of ages into 3 classes shows to be a fair approach, with a balanced accuracy of 40%, due to data imbalance. As for subject identification, open and closed sets approaches were performed. The maximum accuracy values were in a database of 10 identities, achieving, in the closed set approach, 98% for audio-dependent, and 55% for audio-independent; in the open set case, 83% and 55%. Subject-dependent and subject-independent comparisons also took place, and it is concluded, in general, that the learning and testing processes become easier when subjects are already known by the model. Finally, audios with speeches in native languages and English are compared throughout this work, and, overall, both utterances seem to have similar performances.**

*Index Terms*—**Machine Learning, Biometrics, Model Selection, Data Imbalance, Closed and Open Set**

## I. Introduction

VOICE analysis has gained importance over the years with applications in the most diverse areas, from biomedical devices, to security and authentication systems [1]. These systems should work in a variety of environments and languages without decreasing its performance. Within voice recognition, there are several areas of interest with different tasks, all relevant in specific scenarios.

One of them is gender classification based on voice signals. Although it seems to be a simple task, it is not as easy for an alogrithm as it is for humans. This classification can enhance the performance of several applications: from human-computer interactions to online advertisement and access control. Through the years, several works have proposed solutions for this classification problem, using different Machine Learning models, namely Random Forest, Support Vector Machine (SVM) and Neural Network (NN) [2], [3]. Considereing this, 4 different models were tested to evaluate their performance in gender classification using voice signals, testing each model with datasets using mother tongue or lingua-franca utterances of bilingual speakers. This analysis will evaluate the performance of already tested models in different languages, checking applicability in real world applications, where models should be equally efficient in different languages.

Age classification is also addressed in this project. Similarly to gender classification, age classification shows to have several potential applications, for instance in security systems or for biometrics authentication. It is possible to do this classification since voice changes throughout a person's life, specially after adolescence and when reaching older ages [4]. To classify age based on voice, some studies have been performed, with design of classifiers based on supervised learning, including SVM and Random Forest models for identification of children, as well as unsupersived learning, [5], with mixture of Gaussians [6]. In this work, for age classification, two main strategies were used: multi class classification with a SVC classifier, with three classes, defined based on biological reasoning [4], and considering the data imbalance. An alternative and experimental strategy consisted of fitting the data into a regression, and then mapping the age estimations into the three defined classes. The performances of these models are also assessed for cross-linguistic matters.

Speaker recognition is a process that recognizes individuals based on their voice. Speaker identification can be performed in a closed or open-set. A closed-set classification assumes that the identity of the tested voice is already present in the database and only needs to be identified. In open-set case, it is not known if the identity of the voice in analysis is present in the database, so the algorithm needs to decide if the voice is present in the database and, if it is, then it can identify it. This decision will require a threshold so that false matching of voices does not occur. Though the last approach is more challenging, it also presents a wider range of applications.

In recent years, several studies have shown that the energy distribution of human voice utterances follows a Gaussian Model. Thus, Gaussian Mixture Models are identified as one of the main approaches to model human voice signals [9]. A simplified approach of these models is assessed in this paper: instead of using Gaussian Mixtures, the developed algorithm uses Gaussian Naive Bayes to model each identity of the database, with the goal of performing closed and open-set classification in a subset of identities of the dataset.

## II. Data Preparation

### A. Dataset

The dataset used was *BVC Voice Dataset*, comprising voice utterances from 526 individuals of which 336 are males and 190 are females. The total number of voice utterances are 3,964 consisting of 2,149 male and 1,815 female. The dataset contains five different speeches of English and the equivalent translated native languages. Each subject has 1 to 5 voice recordings in English language and their native languages. The native language set is made up of 28 different languages [8].

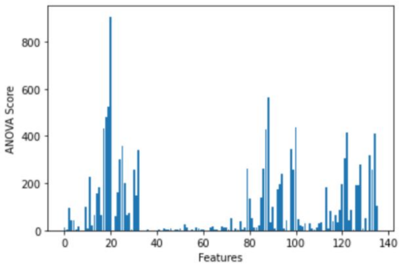


Fig. 1. ANOVA scores for each of the 136 extracted features.

### B. Feature Extraction

In this work, voice signal features were used. Considering that different features are considered relevant in literature [8], [10], a set of signal features was chosen and extracted from all the utterances in the database [11]. The pyAudioAnalysis (v.0.3.6) Python library was used to extract all audio features.

### C. Dataset Split and Cross-Validation

After extracting the features from the audios of the dataset, the data was split into training and test sets. Models were trained with 10-fold cross-validation, in which a portion of the training set is reserved for validation. Therefore, to train and test with as much data as possible, the training data was divided into training and validation subsets in 10 different pseudo-random ways. At the end, the dataset consisted of: a test set containing 15% of the total dataset, and 10 different combinations of training and validation subsets, with training comprising 70% of the total data, and validation the remaining 15%. This split was applied on the entirety of the data.

In addition, the dataset split could be dependent or independent on the subjects' ID. When subject-dependent, the split is done with no consideration of the ID, which means that, since each person has multiple utterances, audios corresponding to the same person could be present in the training, validation, or test subsets. If the split is subject-independent, features corresponding to audios of the same person are grouped together and only appear either on the training, validation or test set. Additionally, the dataset could also be separated by gender, in order to evaluate its influence in age classification. At the end, each one of these combinations (entirety of the data, data separated by gender, and data considering subject-dependency or not), prepared for 10-fold cross-validation, were saved into Pickle files, thus assuring data split was done the same way, for all experiments, throughout this work.

## III. Experiments, Results and Discussion

### A. Gender Classification

As previously mentioned, one of the goals of this work is to compare the performance of gender classification using English or native language utterances of bilingual speakers. With this in mind, different models were trained with the whole dataset and tested with 3 types of test sets: containing all data, only containing English utterances, or only containing native language utterances. Models were trained following the cross-validation strategy already described.

#### 1) *Pre-Processing Strategies*

The effect of Principal Component Analysis (**PCA**), at the beginning of each model's pipeline, was assessed, with the goal of dimension reduction and computation cost decrease. The variance was set to 95%, considering literature's previous studies and after a quick optimization process [12], [13].

As a feature selection procedure, Analysis of Variance (**ANOVA**) was assessed. In figure 1, one can observe features' ANOVA scores which suggests that not all features are equally significant, and, thus, a feature selection approach may be helpful in classification tasks for this dataset. With ANOVA, F-values were computed to assess features' scores, and thus, select the most significant ones, as an alternative approach for data pre-processing, at the beginning of the machine learning pipelines. The number of most significant features was set to 3, since the achieved results were interestingly high, considering the whole dataset contains 136 features [14].

All models were trained in 3 ways: with all features and full dimension; with dimension reduction with PCA 95% variability; with ANOVA for selection of 3 best features.

In this section, we compare linear and nonlinear models, preceded by the aforementioned pre-processing strategies. Linear models: SVM and Logistic Regression, compared with nonlinear models: SVM with Gaussian kernel and Random Forest Classifier, for which results are in table 1.

#### 2) *Linear Models*

A **linear SVM** was trained and optimized with a grid search. The optimized achieved values were C = 100, for regularization, and the hyperparameter $\gamma$ = 0.01.

After testing the influence of PCA and ANOVA in the pre-processing, the best linear SVM model was with PCA, for which accuracy was 74.3%, F1-Score 75.5%, AUROC 76.1%.

**Logistic Regression** model was trained with the raw features, until realizing that features normalization benefited the classification results. Mean was subtracted, and variance was made unitary for all features. The function used for Logistic Regression contained cross validation incorporated. The parameter Cs, which is inversely proportional to the regularization strength, was tuned to obtain better validation results, and the optimized final value was 50.

The best version of Logistic Regression model was with ANOVA selection of 3 best features, for which accuracy was of 72.2%, F1-Score 75.1%, and AUROC 75.3%. It is interesting to notice that a dataset with 136 features has such a robust performance for gender classification when the number of features is reduced to only 3, which highly decreases the computational cost of classification.

Comparing both linear models, the linear SVM performed slightly better than the Logistic Regression. As for English vs. native utterances, linear SVM shows similar results in both cases, while Logistic Regression achieves slightly better results for native utterances.

#### 3) *Nonlinear Models*

A **SVM with Gaussian kernel** model was trained and the optimized parameters were, once again, C = 100 and $\gamma$ = 0.01.

Table 1. Summary of results for models tested for gender classification

| Tested Models | | | | Model Assessment Metrics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Complete Test | | | English Language only | | | Native Language only | | |
| | Classifier | Linearity | Pre-Processing | Accuracy | F1-Score | AUROC | Accuracy | F1-Score | AUROC | Accuracy | F1-Score | AUROC |
| Subject-Independent | SVM | Linear | None | 0.743 | 0.745 | 0.756 | 0.748 | 0.749 | 0.754 | 0.739 | 0.742 | 0.757 |
| | | | with PCA (95% variability) | 0.743 | 0.755 | 0.761 | 0.742 | 0.755 | 0.750 | 0.744 | 0.755 | 0.771 |
| | | | with ANOVA (3 features) | 0.732 | 0.754 | 0.754 | 0.719 | 0.745 | 0.738 | 0.744 | 0.763 | 0.771 |
| | Logistic Regression | Linear | None | 0.737 | 0.747 | 0.755 | 0.745 | 0.753 | 0.756 | 0.730 | 0.740 | 0.755 |
| | | | with PCA (95% variability) | 0.727 | 0.745 | 0.730 | 0.734 | 0.752 | 0.749 | 0.721 | 0.739 | 0.774 |
| | | | with ANOVA (3 features) | 0.722 | 0.751 | 0.753 | 0.711 | 0.743 | 0.738 | 0.733 | 0.760 | 0.771 |
| | SVM | Non Linear (RBF) | None | 0.750 | 0.756 | 0.779 | 0.745 | 0.751 | 0.770 | 0.756 | 0.762 | 0.788 |
| | | | with PCA (95% variability) | 0.747 | 0.757 | 0.780 | 0.742 | 0.750 | 0.771 | 0.753 | 0.764 | 0.789 |
| | | | with ANOVA (3 features) | 0.737 | 0.758 | 0.754 | 0.725 | 0.749 | 0.738 | 0.750 | 0.767 | 0.772 |
| | Random Forest | Non Linear | None | 0.750 | 0.757 | 0.784 | 0.754 | 0.760 | 0.775 | 0.747 | 0.754 | 0.791 |
| | | | with PCA (95% variability) | 0.706 | 0.728 | 0.764 | 0.699 | 0.721 | 0.748 | 0.713 | 0.735 | 0.779 |
| | | | with ANOVA (3 features) | 0.729 | 0.737 | 0.765 | 0.711 | 0.720 | 0.758 | 0.747 | 0.754 | 0.773 |
| Subject-Dependent | SVM | Linear | with PCA (95% variability) | 0.761 | 0.773 | 0.799 | 0.784 | 0.787 | 0.814 | 0.740 | 0.761 | 0.786 |
| | Logistic Regression | Linear | None | 0.766 | 0.776 | 0.816 | 0.777 | 0.780 | 0.839 | 0.756 | 0.772 | 0.797 |
| | SVM | RBF | with PCA (95% variability) | 0.764 | 0.775 | 0.816 | 0.780 | 0.783 | 0.824 | 0.750 | 0.768 | 0.809 |
| | Random Forest | Non Linear | None | 0.773 | 0.780 | 0.833 | 0.791 | 0.792 | 0.861 | 0.756 | 0.770 | 0.809 |

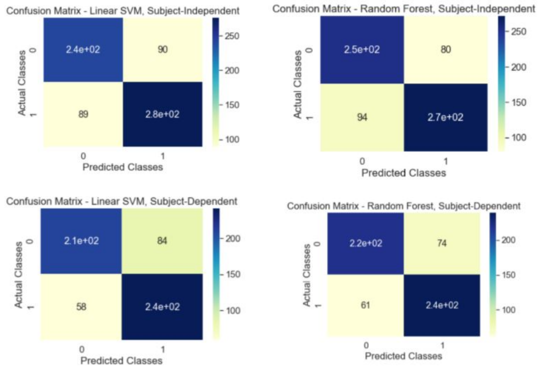


Fig. 2. Confusion Matrices for gender classification on test set. On the left, linear SVM (with PCA); on the right, Random Forest Classifer. On top, considered datasets are subject-indepent; on the bottom, subject-dependent. Class 0 for Female and class 1 for Male.

The nonlinear SVM showed better performances when PCA was performed at the beginning of the pipeline, achieving an accuracy of 74.7%, F1-Score of 75.7%, and AUROC of 78%.

A **Random Forest Classifier** was also trained and, with grid search for parameters optimization, the best number of estimators (trees in the forest) was 250; maximum depth (of the tree) was 10; minimum number of samples required to be at a leaf node 20; and the number of features to consider when looking for the best split was maximum.

The resultant model had the best performance when no pre-processing took place, with accuracy 75%, F1-Score 75.7%, and AUROC 78.4%.

Comparing nonlinear models, one can conclude that the best performance is achieved by the Random Forest Classifier. As for audio language analysis, again, it is not clear which one performs better: for non linear SVM, native utterances seem to have better F1-Scores, while for Random Forest Classifier, the trend is opposite.

Comparing the best linear model with the best nonlinear models, performances are similar: 76% F1-Score. The confusion matrix for the test of each model is in figure 2. One can observe that males are more accurately classified than females.

Finally, regarding subject-dependency, all models show to have better performances when subjects are all mixed in train and test sets, which makes sense, since the model can perform better for already known persons. In all models, the F1-Score is 2-3% better for subject-dependent datasets.

Literature results, which apply convolutional neural networks (CNN) to this dataset, present overall accuracies of about 83% for subject-dependent gender classification. Literature's results also report better performances for native utterances, due to higher competence and fluency of the subjects speaking their mother tongue [8]. The results obtained in this work are encouraging, as for the same classification task, we could achieve an accuracy of 77.3%, with the Random Forest Classifier. However, as for the cross-linguistic effects, the developed work and results did not lead to clear conclusions.

## *B. Age Classification*

For the classification of subject's age, the first step was defining how to segment data into 3 classes, according to the ages, which are within the range 14 to 65 years. Biological reasoning, such as the analysis of voice changes throughout life, were taken into account [16], and thus the considered intervals were 14 - 16 for class 0; 17 - 25 for class 1; 26 - 65 for class 2.

A concern inherent to this task was class imbalance: the ages' distribution is not uniform, as can be seen in the histogram in figure 3, and ages between 20 and 25 years were richer in terms of available amount of data, accounting for about 89% of the data. This non-uniform distribution is particularly unfortunate considering the classes defined. To tackle this problem, some well-known techniques were explored, such as data pre-processing, in which the minority classes can be oversampled, through algorithms like SMOTE (Synthetic Minority Oversampling Technique); and training with costs, in each the cost of misclassifying a class is inversely proportional to its frequency [17].

### *1) **Classification Task***

As a first approach, a classification task was performed with an optimized Support Vector Classifier (SVC), with a

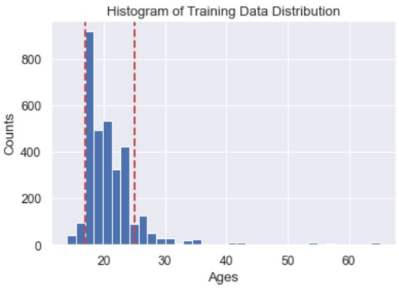
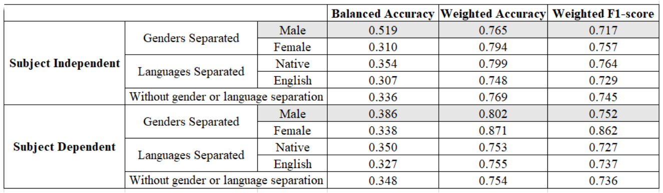


Fig. 3. Histogram of training data distribution, for each age label. The red dashed lines represent the classes separation.

Table 2. Summary of results for age classification



| | | | Balanced Accuracy | Weighted Accuracy | Weighted F1-score |
|---|---|---|---|---|---|
| **Subject Independent** | Genders Separated | Male | 0.519 | 0.765 | 0.717 |
| | | Female | 0.310 | 0.794 | 0.757 |
| | Languages Separated | Native | 0.354 | 0.799 | 0.764 |
| | | English | 0.307 | 0.748 | 0.729 |
| | Without gender or language separation | | 0.336 | 0.769 | 0.745 |
| **Subject Dependent** | Genders Separated | Male | 0.386 | 0.802 | 0.752 |
| | | Female | 0.338 | 0.871 | 0.862 |
| | Languages Separated | Native | 0.350 | 0.753 | 0.727 |
| | | English | 0.327 | 0.755 | 0.737 |
| | Without gender or language separation | | 0.348 | 0.754 | 0.736 |

Gaussian kernel, regularization parameter C equal to 0.1 and kernel coefficient gamma equal to 1.8. The SMOTE algorithm was applied to the training data before training the models, so that the model could be trained with equally-distributed and significant data from all classes. The SVC was then trained in two ways: firstly, with the entirety of the training set; secondly, in training sets in which features where separated by gender. The reason behind classifying age after splitting the dataset by gender was the hypothesis that the gender could influence age classification. For example, younger male voices may be similar to older female voices. [4].

Additionally, the classification model trained with the entire dataset was tested in datasets split by native and English language. Moreover, all these sets had two variants: spliting considering subject dependency and independence. The results for the classification with SVC, following pre-processing with SMOTE, can be found in the table 2.

As it can be observed, the accuracy results differ a lot in general when comparing the weighted accuracy (among all classes), with balance accuracy, which takes into account the imbalance of the classes in the data by averaging the recall value for each class. It is possible to conclude that the high values in the other metrics are due to the fact that the data is so imbalanced that a great majority of it belongs to class 1, being easily identified. This conclusion can be confirmed by analysing the confusion matrices in figure 4. The balanced accuracy shows that, in reality, the developed model does a poor job in most cases, not being able to classify correctly data that does not belong to the majority class 1. It is interesting to notice that the best results occur in the classification of men, both in datasets with subject dependency and independence. Although the results for females are not as high (possibly due to the fact that the training dataset was smaller), this shows that gender separation might be a good approach before age classification. In addition, it is also possible to notice that the classification results are slightly better in data corresponding to native utterances. However, despite the not great overall results, it is important to notice that if the imbalance of the training data had not been corrected with the SMOTE algorithm, the model would most likely predict class 1 for every point, while in this case the model was trained with balanced datasets, so theoretically wouldn't have the tendency to lean towards one major class, and was able to classify correctly some data points outside of class 1. But, all in all, it can be concluded that the test set used doesn't provide a good evaluation of the model due to its imbalance, since its evaluation becomes extremely dependent on the classification of one single class in the set, and not on the overall multiclass classification.

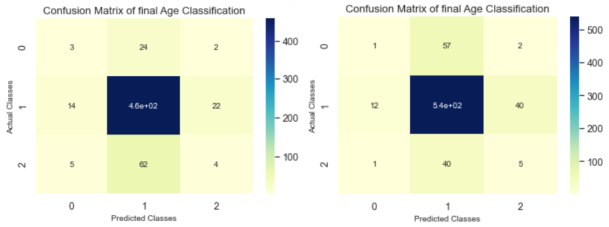


Fig. 4. Confusion matrices for the age classification with test split with subject dependency (left) and dependency (right)

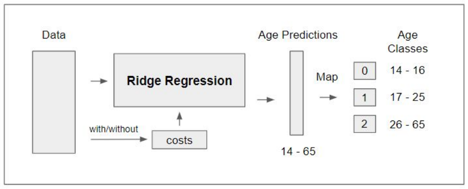


Fig. 5. Diagram representing data flow of the proposed model, with Ridge regression followed by a mapping of ages predictions into ages classes.

### 2) *Regression Task followed by class mapping*

Not fully satisfied with the previous classification results, an alternative approach was performed: the task was now primarily faced as a regression task, and the age estimations were then mapped to the already defined classes (figure 5). Though the results were not the expected ones, in this section, we will present our experiments and main results for this task.

In the chosen approach, samples' weights were firstly computed, as being inversely proportional to classes frequencies, and then included in the loss function of a Ridge regression, which was then trained with $\lambda = 0.1$, for regularization. The cross-validation approach was similar to previous ones. Only subject-independent datasets were now considered. SMOTE algorithm, for an equalization of frequencies, could not be applied for this classification strategy. The raw age labels, in range 14 - 65, often lack samples to make the implementation of SMOTE possible, which needs a minimum amount of data per label (number of samples per label must be higher than number of neighbours, usually set as 5).

The evaluation metrics for the test of the model, with and without the costs, can be found in table 3, and the computed confusion matrix, for each scenario, in figure 6. For the analysis of the stand-alone regression, the obtained scores ($R^2$) were -0.158 without costs, and -7.64 with costs; the confusion matrices for the regression age predictions are in figure 7.

In fact, the results for stand-alone regression are not very

Table 3. Summary of results for Ridge Regression followed by class mapping, for age classification

| Model | Model Assessment Metrics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Complete Test | | | English Language | | | Native Language | | |
| Ridge Regression | Balanced Accuracy | Weighted Accuracy | Weighted F1-Score | Balanced Accuracy | Weighted Accuracy | Weighted F1-Score | Balanced Accuracy | Weighted Accuracy | Weighted F1-Score |
| Without Costs | 0.332 | 0.844 | 0.776 | 0.332 | 0.845 | 0.777 | 0.331 | 0.842 | 0.775 |
| With Costs | **0.401** | 0.311 | 0.390 | 0.404 | 0.315 | 0.394 | 0.398 | 0.307 | 0.389 |

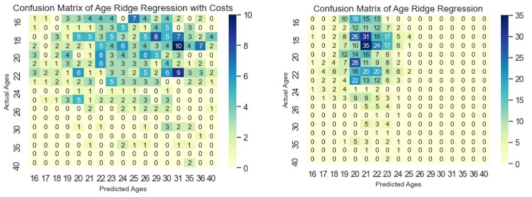


Fig. 6. Confusion Matrices for stand-alone Ridge regressions, with (left) and without (right) costs.

enthusiastic. The test set's R$^2$ values are negative, indicating that the model does not even follow the data trend. The confusion matrices show low capability of correctly predicting ages, whatever the label, with very few predictions perfectly matched with the actual age. In general, the regression with costs has a worse performance, with an even more negative score, and more dispersed values in the confusion matrices. This may be justified by an increase of noise, when samples with age labels with very low frequencies are highly valued in the loss function, during learning process.

However, when the regression with costs is combined with the mapping into the 3 age classes, the results become a bit better (table 3), with a balanced accuracy of 40.1% for the complete test set. Though the weighted F1-scores steeply decreases (39%, which makes sense, taking into consideration the poor stand-alone regression results), this is a result that can be considered better than the one achieved by the SVM classification approach, considering the data imbalance. Once again, the model without costs presents high values of weighted F1-score (77.6%), but low balanced accuracies (33.2%), which is also within the expected, since the predictions are strongly leaning for class 1.

When analysing the confusion matrices in figure 7, it is clear that data imbalance is a real limitation for this task, since there is not a single accurate prediction for class 0 and class 2, without costs considered. A closer look at the confusion matrices for the whole model with costs shows that samples are mainly predicted as class 2, when the actual class is 1. Though the weighted metrics of the model with costs highly decrease, costs make the model capable of accurately predicting some more samples as class 0 or 2, which was not possible otherwise.

All in all, this approach slightly improves the results for age classification, when comparing to the previously reported strategy, with SVM, in which a maximum balanced accuracy of 33.6% for the whole dataset (subject-independent) was obtained. The data imbalance is not fully tackled with success, though, since the model is only mainly capable of identifying ages encompassed in the range 17 - 25, which represents about 88% of the dataset. Values outside this range are rarely well identified. This composed strategy could eventually be improved with some kind of nonlinear regressor, such as a MLP regressor, and an enhanced approach to class imbalance, for instance, combining oversampling of minority classes and undersampling of majority classes [17]. Also, considering the results regarding gender separation for age classification, obtained in the SVM model for age classification, this could also be considered to further enhance the results obtained with the regression approach.

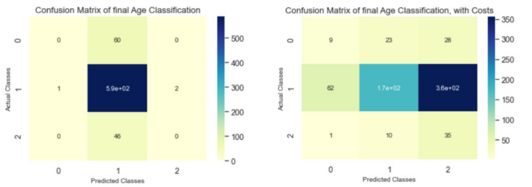


Fig. 7. Confusion Matrices for age classification, after Ridge regressions, with (left) and without (right) costs.

Although the results obtained in this work are not extremely satisfying, results in literature regarding for age classification with a SVM classifier can be a bit more promising. In a work done with a (age-imbalanced) dataset of voice calls, SVM models could classify age with a recall of up to 60% [7]. It is important to refer, too, that the authors of the BVC dataset did not explore age classification [8].

### C. Closed Set and Open Set Speaker Identification

Towards the identification of specific subjects in a subset of voice utterances, two distinct pipelines, to perform closed and open set classification, were designed. Figure 8 illustrates a representative scheme of both approaches. The closed set classification will test if the algorithm can correctly identify a subject within a database of subjects, assuming he already integrates this dataset. The open set approach, on the other hand, will validate if the developed algorithm can properly recognize if a voice utterance is present in a pool of identities and correctly identify it.

In both approaches, Gaussian Naive Bayes was used to model each identity class, present in the database, employing a diagonal co-variance matrix to each model, assuming features independence. After the creation of the Gaussian database, with both English and native utterances for each subject, the audio track in analysis was tested in each model and the probability of the sample belonging to each class was computed. This process is common for both approaches.

In the closed set pipeline, if the model with the highest probability, previously computed, corresponded to the subject in analysis, the track was considered identified. In the open set pipeline, when the probability of belonging to a Gaussian model was equal or above a given threshold, the track was considered identified.

The experiments performed consist in a increasing number of identities in the database from 10 to 25, 50 and 100. For the open-set classification, two thresholds were tested: 1.0 and 0.7. In the first case, the algorithm will only consider the identity in analysis to be present in the database if the probability of belonging to a Gaussian Model from the database is equal to 1. The second case is a bit more flexible, in which this probability will only have to be higher than 70%. In addition, two different scenarios were also tested: audio dependent and

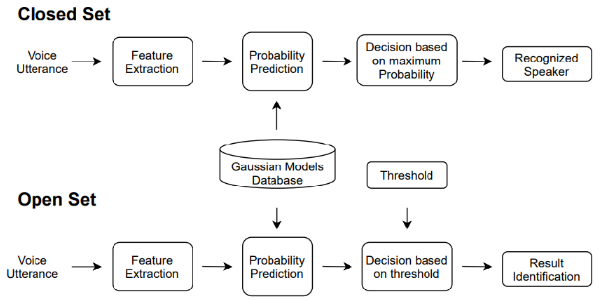
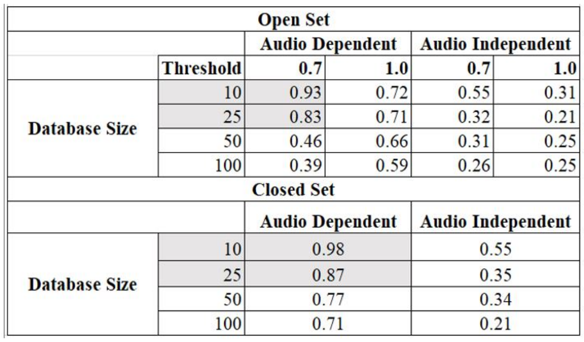


Fig. 8. Scheme of Closed and Open-Set Identity Classification Pipeline.

Table 4. Summary of accuracy results for closed and open set classifications



| Open Set | | | | | |
|---|---|---|---|---|---|
| | | Audio Dependent | | Audio Independent | |
| | Threshold | 0.7 | 1.0 | 0.7 | 1.0 |
| Database Size | 10 | 0.93 | 0.72 | 0.55 | 0.31 |
| | 25 | 0.83 | 0.71 | 0.32 | 0.21 |
| | 50 | 0.46 | 0.66 | 0.31 | 0.25 |
| | 100 | 0.39 | 0.59 | 0.26 | 0.25 |
| Closed Set | | | | | |
| | | Audio Dependent | | Audio Independent | |
| Database Size | 10 | 0.98 | | 0.55 | |
| | 25 | 0.87 | | 0.35 | |
| | 50 | 0.77 | | 0.34 | |
| | 100 | 0.71 | | 0.21 | |

audio independent classification; in the first case, the audio tested was used in the training of the Gaussian Database;in the second, it was not. The obtained accuracy for each of the tests performed is present in table 4.

Regarding the open set classification, in particular, the accuracy values obtained when the threshold is 1.0 (maximum value of 72% for audio dependent with a pool of 10 identities) are generally lower when compared to 0.7 (maximum value of 93% for audio dependent with a pool of 10 identities), which makes sense, taking into consideration that the latter is more flexible, and considers that the probability of the audio fitting a specific model does not have to be exactly 1.

Comparing different database sizes, for both identification tasks, better results were generally obtained for databases with a smaller number of identities, where for the best scenario - closed set, audio dependent - the accuracy varied from 98% (10 subjects) to 71% (100 subjects).

Comparing both audio dependent and audio independent classifications, the first has higher values of accuracy for all test sets, which was foreseeable since the model has used the audio being tested to create the database of identities, making the identification task much easier. Indeed, to insure higher values of accuracy for the audio independent task, which has greater potential in real world applications, the number of recorded audios per individual would have to be higher, since in this dataset a maximum of 10 audios exist per subject, which is very limited and does not allow for the creation of robust models to characterize each identity, specially given that we are using simple Gaussian Naive Bayes models. In fact, reported literature has shown values of around 90% of accuracy for this type of classification, but using a dataset containing 20 samples per subject, more than double, in average, than what we worked with [18].

Finally, addressing the differences in closed and open set identification, the latter is more challenging and, thus, it was expected to result in lower accuracy values when compared to closed set. In this regard, the results here reported, in the case of audio dependent analysis, are in agreement with the predictions, with the maximum accuracy values for the open set identification being 83% and 93% (threshold = 0.7, database size equal to 25 and 10, respectively) as opposed to 87% and 98% in the closed set scenario (database size equal to 25 and 10, respectively).

## IV. Conclusion

The work here reported had the aim to develop simple but effective Machine Learning strategies to achieve gender, age and identity classification, in a multilingual dataset. Regarding gender classification, good results were obtained with both SVM and Random Forest Classifier, which generally maintained their performance when testing with both English only and native language sets (with differences bellow 1%). This validates universal applicability of these models, independently of the language. For age classification, a pipeline with Ridge Regression and age mapping into 3 classes showed to be a validate approach to achieve this task. The use of a SVC model is also considered a solid approach for age classification, but the testing results are a bit inconclusive due to data imbalance. Finally, for test of open and closed set approaches towards speaker identification, Gaussian Naive Bayes was used to model each identity and showed to be a fair and simple approach, though further tests should be performed in datasets with a higher number of audios per subject.

## References

[1] Abozaid, A. et al. Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion. Multimed Tools Appl 78, 16345–16361 (2019).

[2] P. Gupta et al. ”A Stacked Technique for Gender Recognition Through Voice,” 2018 11th (IC3), Noida, 2018, pp. 1-3.

[3] S. Chaudhary and D. K. Sharma, ”Gender Identification based on Voice Signal Characteristics,” 2018 (ICACCCN), Greater Noida (UP), India, 2018, pp. 869-874.

[4] M. Lee Hummert et al. Journal of Nonverbal Behavior, ”Vocal Characteristics of Older Adults and Stereotyping”, vol. 23, pp. 111-132, 1999.

[5] Katerenchuk, Denys. (2018). Age Group Classification with Speech and Metadata Multimodality Fusion.

[6] J. Přibil et al ”GMM-based speaker gender and age classification after voice conversion,” 2016 (SPLINE), Aalborg, 2016

[7] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt and E. Noth, ”Age and gender recognition for telephone applications based on GMM supervectors and support vector machines,” 2008 IEEE International Conference on Acoustics, Speech and Signal Processing

[8] O. Iloanusi et al., Voice Recognition and Gender Classification in the Context of Native Languages and Lingua Franca. 2019, pp. 175-179

[9] S. Chakraborty and R. Parekh, ”An improved approach to open set text-independent speaker identification (OSTI-SI),” 2017 (ICRCICN), Kolkata, 2017, pp. 51-56.

[10] Rami S. Alkhawaldeh, ”DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network”, Scientific Programming, vol. 2019, 12 pages, 2019.

[11] https://github.com/tyiannak/pyAudioAnalysis/wiki/3.-Feature-Extraction

[12] https://www.mikulskibartosz.name/pca-how-to-choose-the-number-of-components/

[13] https://medium.com/datadriveninvestor/principal-component-analysis-pca-a0c5715bc9a2

[14] S. Gajawada, ”ANOVA for Feature Selection in Machine Learning”, Medium, 2021.

[15] R. Gandhi, ”Support Vector Machine — Introduction to Machine Learning Algorithms”, Towards Data Science, 2018.

[16] H. Liu et al. ”Age-related differences in vocal responses to pitch feedback perturbations: A preliminary study”, The Journal of the Acoustical Society of America, vol. 127, no. 2, pp. 1042-1046, 2010.

[17] R. Cruz, K. Fernandes, J. Cardoso and J. Costa, ”Tackling Class Imbalance with Ranking”, 2016.

[18] H. B. Kekre and V. Kulkarni, ”Closed set and open set Speaker Identification using amplitude distribution of different Transforms,” 2013 (ICATE), Mumbai, 2013, pp. 1-8.