Language Agnostic Voice Recognition Model

Tea Janeva
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering, Skopje
N. Macedonia
tea.janeva@students.finki.ukim.mk

Kostadin Mishev
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering, Skopje
N. Macedonia
kostadin.mishev@finki.ukim.mk

Monika Simjanoska
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering, Skopje
N. Macedonia
monika.simjanoska@finki.ukim.mk

***Abstract*—Voice recognition is the ability of a machine to identify a person based on their unique voiceprint. As this task is becoming more important and dominant in everyday people's lives, this paper is testing different approaches for its implementation. Using a multilanguage database and working with the different frequencies' characteristics, five machine learning models such as Random Forest, XGBoost, MLP, SVM and Gradient Boosting, along with CNN deep learning model were implemented. The models were trained on three different tasks, gender prediction, age range prediction, and combined gender and age range prediction. These models were evaluated using accuracy, F1-score and MCC score. The results showed that Random Forest outperforms other models by achieving an accuracy of more than 0.9 for all the three classification tasks.**

***Index Terms*—Voice recognition, Deep learning, Machine learning, Explainable Machine learning**

I. Introduction

Voice recognition is slowly but surely becoming a key part of the future of communication. This task is the ability of a machine or a program to receive and understand spoken commands from a speaker and later to identify this person based on their unique voiceprint. On the other hand gender recognition and recognition of the age range group of the user are becoming essential information for the interactions of the users, and altogether for the social community. Both of these tasks are something that the human brain can automatically identify, and at the moment these are some of the most explored subjects in the world of voice recognition - the technology behind some of the most popular virtual assistants such as Amazon's Alexa, Apple's Siri, Google's Google Assistant and Microsoft's Cortana.

Most of the voice recognition applications work by analyzing the sound provided by the user and then implementing a variety of methods to understand it and identify the speaker. As this is a technology that first appeared 50 years ago, throughout the process of creating a working voice recognition application, different approaches were tried and were successful. The work that is currently available and working can be separated in 2 categories, text-dependent systems that require predetermined voice passphrases and text independent systems where the subject of analysis is conversational speech. The first category requires the speaker to provide audio of certain key-words and sentences. Some of the most used methods are Dynamic Time Warping (DTW)-based methods and hidden Markov model (HMM)-based methods. Text independent systems given their nature are more popular. The ability to predict without constantly needing to be provided with audios of certain key-words gave the possibility for these systems to be done sequentially. Some of the most used methods are Long-Term-Statistics-Based Methods, VQ-Based Methods, Ergodic-HMM-Based Methods and Speech-Recognition-Based Methods. Both systems have a mutual serious weakness, their security systems can be easily bypassed. Trying to overcome this weakness, another category of voice recognition applications is available, a text-prompted speaker recognition where the idea is that the password sentences are changed every time. Looking more in depth behind the elemental job, voice recognition can be classified into 2 parts, speaker identification and speaker verification. Speaker identification is determining which speaker is the system currently working with, and speaker verification is the process of confirming or denying the identity that is calculated [3].

Analyzing the current work made in this field and for the selected subjects, the majority of speech and later on voice recognition is conducted on single language datasets, mostly working with English speakers. This study investigates different approaches, giving an importance in predicting the speaker on a model that is trained on multiple languages at once. Languages differ in many ways, including the way it is spoken, the way the words are formed and therefore the way the sentences are put together. The difference that is of importance for voice recognition are the different spoken ways, more specifically the features that different languages offer. Focusing on the frequency every language has, there can be seen a lot of differences as each language has its own, with the highest frequency range found in UK English, which peaks at 12000 Hz. Some of the languages with lowest frequency range used in this study are French and Spanish where on average the peak is at 2000 Hz. On the other hand, in general female frequencies are higher than male and correspondingly as the speaker ages the frequency of their speech lowers. [6]

In this paper we are proposing the use of both Machine Learning (ML) and Deep Learning (DL) models. Focusing on the different aspects that working with multiple languages at once offers, we implement and test the models on diverse information and features. Combining both types of models, more ways were explored and more results were acquired for determining the best approach.

The rest of the paper is organized as follows. In Section II different researches on the topic are explored and dis-

cussed. In the methodology Section III an accent is put on explaining the dataset and its creation, the process behind the data preprocessing and also all the ML and DL models. Explainable Machine Learning is introduced as a way to explain the models and the influence of the features on the final prediction. The results are presented and discussed in Section IV and a conclusion is derived in the final Section V.

II. Related Work

Most often the traditional ML models are modified and implemented and later satisfying results are achieved. For example, a system using bootstrapping for audio based gender recognition implemented 5 models such as Naive Bayes, Nearest Neighbor, Decision Tree, Support Vector Machine (SVM) and Logistic Regression on a dataset based on 10 news broadcast excerpts. The classification accuracy for all the models was around 85% [8]. Another example of research in this field combining both gender and age recognition is shown in the models presented in [5]. There were 7 different proposed and tested methods using Gaussian mixture models (GMM) and Support vector machine (SVM) models based on different features such as mel-frequency cepstral coefficient (MFCC), GMM mean supervectors, acoustic, prosodic and voice quality, UBM weight posterior probability supervectors, GMM maximum likelihood linear regression (MLLR) matrix supervectors, polynomial expansion coefficients of the syllable level prosodic and weighted summation of the fusion of all the previously listed. The dataset used is the aGender database. The final results are around 88% weighted accuracy for gender prediction and around 50% for the age prediction. One of the best accuracy for this task was acquired by an approach trying to automatize the Bengali Voice based gender classification [1]. In this study Mel-frequency cepstral coefficient (MFCC) was used for feature extraction and for the ML models Logistic Regression, Random Forest and Gradient Boosting were implemented. Testing on a dataset with more than 250 speakers, an accuracy of 99.13% was obtained.

Using a Multilayer Perceptron on a dataset with more than 3000 samples of both female and male speakers, an accuracy of 96.76% was obtained [2]. The data was preprocessed extracting 22 acoustic parameters from the acoustic signals. Making a combination of acoustic and pitch features and implementing a set of Neural Networks is an example of some of the most accurate approaches for gender classification [4]. The system classifier contains individual experts with 4 MLPs each. Each MLP is separately trained using the Back Propagation algorithm. When testing this model on a Switchboard dataset with 19 male and 19 female speakers an accuracy of 98.5% is reached.

III. The methodology

A. Dataset

This study is conducted on multiple datasets containing audios from different languages. Primarily the dataset contained only English speakers from different backgrounds, but to benchmark the results it was updated and another 15 languages were implemented. The main goal for introducing new languages and working with more than one at a time was to explore the similarities and differences between the voices and their characteristics from different parts of the world. In the final dataset the most dominant language is English represented by approximately 6000 speakers with more than 64000 audios, however, the total number of speakers in the dataset was 13399. The languages that are present in the final, combined dataset are: English, Italian, French, Spanish, Russian, Portuguese, Croatian, Ukrainian, Greek, Turkish, Catalan, Bulgarian, Dutch, Hebrew, Persian and Albanian. The distribution between languages can be seen in Table I. When collecting audios from languages different to English, the same format of the audios was preserved concerning length and number of audios per speaker. All of the data was collected from VoxForge [9] which offers open speech datasets.

TABLE I
Distribution of different languages in the dataset.

Language	Number of speakers
HR	13
FA	15
NB	23
SQ	33
UK	39
CA	41
BG	44
EL	112
TR	154
PT-BR	319
RU	576
NL	682
IT	922
ES	1986
FR	2101
EN	6339

The main focus of this study is towards distinguishing between 6 different classes combined from age and gender.

The numbers between female and male speakers are drastically different, as presented in Table II, implying imbalanced dataset. The first part of this study gives an accent specifically on the gender, and using all the features from the audios of the speakers trying to predict if the speaker in question is male or female.

TABLE II
Representation of gender samples.

Gender	Number
Male	10883
Female	1130

Focusing on the age range, there are 3 age groups implemented as youth, adult and senior. Each of these groups is represented by a different age range. Youth is characterized by 25 years and younger (kids under 18 years are excluded), adult group contains the ages between 25 and 55, and the senior group contains the older than 55. The majority of the dataset is represented by the adult group, then the young group, and the least amount of data is for seniors as can be seen in Table III.

Combining the 2 classes, we created 6 new possible classes: female youth, female adult, female senior, male youth, male adult, and male senior. Within these 6 classes, the

TABLE III
DISTRIBUTION OF AGE RANGE SAMPLES.

Age range	Number
Youth	1643
Adult	10027
Senior	182

dominant is adult males, and the group for which there are not representatives is senior females as can be seen in Table IV. Thus, this class is completely omitted in the further analysis.

The data downloaded to create the dataset is in a specific format, being separated in 2 parts. The first part is the audio of the speaker, split in multiple shorter audios, each of around 10 seconds in duration. The second part is the information about the speaker, such as gender, age range, pronunciation, language, etc. When adding new audios from different languages in the database we followed the same structuring format.

TABLE IV
DISTRIBUTION OF SAMPLES CONSIDERING COMBINED CLASSES.

Age range and Gender	Number
Youth + Male	1507
Adult + Male	8993
Senior + Male	182
Youth + Female	101
Adult + Female	997
Senior + Female	0

B. Preprocessing

Since we are training the classifiers by using both traditional Machine Learning and Deep Learning approach, the dataset was preprocessed in two different manners, accordingly.

For the Machine Learning, the audio files were preprocessed to extract the features describing the frequencies represented in the particular audio, and those are: nobs, mean, skew, kurtosis, median, mode, std, low, peak, q25, q72 and iqr.

For the DL models, raw audios were used as input in the CNN layer of the DL architecture.

The dataset was labeled to correspond to the three different classifiers we intend to create, the first to predict only the gender, the second to predict the age range and the third to predict a label that corresponds to the combination of both gender and age range as presented in Table IV, but without the class for which we do not have any representatives.

C. Machine Learning

Combining 16 languages, we have obtained dataset with 13399 speakers available. After preprocessing the dataset and extracting the features, it was used to build three types of intelligent models capable of predicting the gender, the age range, and the gender and age combined as one label, for a particular speaker.

Five different models were trained for each classification task by using Random Forest, Gradient Boosting, Support Vector Machines, Multilayer Perceptron, and XGBoost. We experimented with different hyperparameters to obtain the best results shown in the following Section IV.

As evaluation metrics we used accuracy, F1 score, and MCC score. N-fold cross-validation was implemented, trying both 5-fold and 10-fold for all the models previously described in this section.

D. Deep Learning

The Deep Leaning model used in this study is the Convolution Neural Network (CNN). Multiple different layers have been implemented such as Conv2D, BatchNormalization, MaxPool2D, Flatten, Dropout and Dense as shown in Figure 1. Adam was used as optimizer, and for the loss we used binary cross entropy. The final result was evaluated by using the same metrics as in the traditional ML approach.

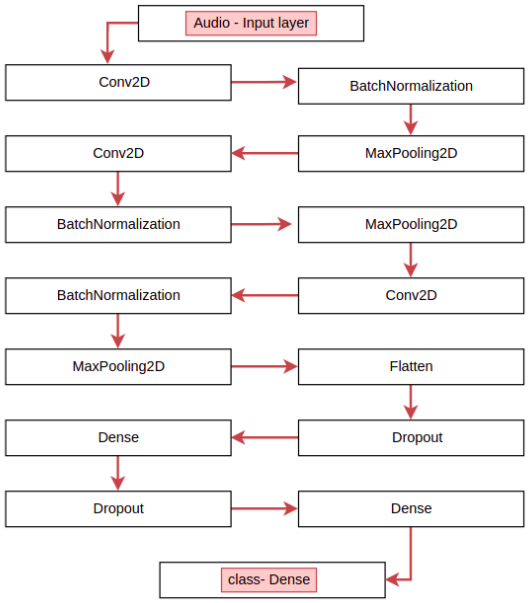

Fig. 1. CNN model- layers

E. Explainable Machine Learning

To understand how the features affect the gender, age and the combined labels classification, an explanation of contribution on each attribute from the dataset to each classification target was done by using SHAP method [7]. To be certain of the impact, Shapley values were calculated for each of the 10-fold round, and consequently averaged. The results are shown in the next Section IV.

IV. RESULTS

This section presents the evaluation metrics obtained from the three types of classifications we presented in the previous sections.

For the traditional ML case, five ML models were trained by using 10-fold cross-validation for all the three types of classifications.

Random Forest performed best at all classification tasks: gender, age range and combined (gender + age range) prediction. Tables V, VI and VII show the performance of all the classification models for the three tasks, respectively.

The MLP and SVM models gave poor results considering the F1-scores and the MCC scores.

TABLE V
RESULTS FROM 10-FOLD FOR ML MODELS FOR GENDER CLASSIFICATION.

Model	Accuracy	F1 score	MCC score
XGBoost	0.895	0.786	0.709
Random Forest	0.972	0.918	0.926
Gradient Boosting	0.949	0.947	0.870
Support Vector Machine	0.781	0.211	0.124
MLP	0.899	0.609	0.243

TABLE VI
RESULTS FROM 10-FOLD FOR ML MODELS FOR AGE RANGE CLASSIFICATION.

Model	Accuracy	F1 score	MCC score
XGBoost	0.907	0.743	0.623
Random Forest	0.977	0.922	0.906
Gradient Boosting	0.895	0.748	0.533
Support Vector Machine	0.848	0.306	0.0
MLP	0.829	0.395	0.188

TABLE VII
RESULTS FROM 10-FOLD FOR ML MODELS FOR COMBINED CLASSIFICATION.

Model	Accuracy	F1 score	MCC score
XGBoost	0.977	0.920	0.854
Random Forest	0.992	0.973	0.946
Gradient Boosting	0.988	0.963	0.931
Support Vector Machine	0.922	0.546	0.139
MLP	0.771	0.372	0.331

For the DL case, the presented CNN architecture in the previous section was used for the three classification tasks denoted as T1, T2 and T3 in Table VIII corresponding to gender, age range and combined classification.

TABLE VIII
CNN EVALUATION FOR THE THREE CLASSIFICATION TASKS.

Task	Accuracy	F1 score	MCC score
T1	0.832	0.767	0.617
T2	0.8	0.692	0.516
T3	0.838	0.359	0.579

Analyzing the results from the CNN evaluation it can be perceived that the DL approach showed worse performance than Random Forest, XGBoost and Gradient Boosting, regarding all the three evaluation metrics at all classification tasks. We assume the results would be better if more data were available for the DL approach.

Analyzing the importance of the features by using the SHAP method as described in the previous section, Figure 2 shows that the median feature has the biggest impact on the gender prediction. Closely behind are impacts from q25 and q75.

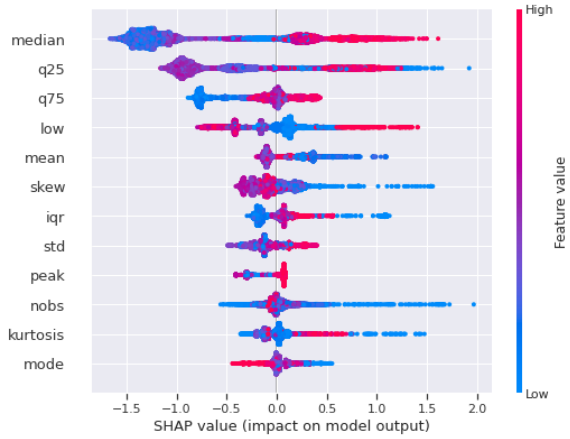

Fig. 2. The most important features for gender prediction.

Looking over the SHAP values for the age range prediction, from Figure 3 it can be seen that the values of skew affect the prediction the most along with Low (meaning the low frequency) and the feature q25.

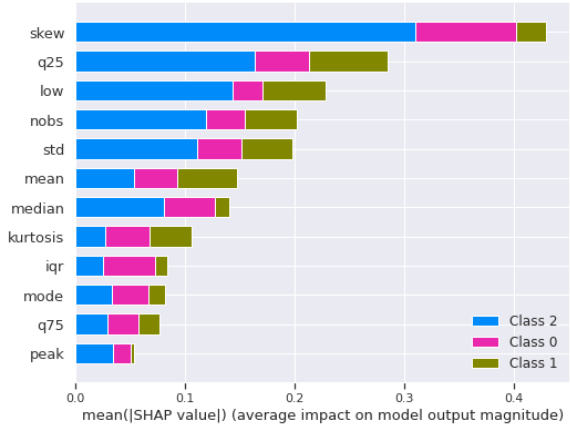

Fig. 3. The most important features for the age range prediction.

Analyzing the SHAP values for the combined gender + age range classification gave significant insight in the features importance overlapping more with the values for gender prediction than age range prediction. The most important feature can be seen as q75, closely followed by the median and q25 values. All the contributions can be observed in Figure 4.

V. CONCLUSION

In this paper, we have presented different approaches to voice recognition by using a multilingual database encompassing 13399 speakers. Using both machine learning and deep learning models, three different classifiers were trained to predict the gender, the age range, and combined label of both gender and age range. The results were evaluated by using accuracy, F1-score and MCC score. Random Forest showed significantly better results when compared to other ML models and DL network.

Additionally, Shapley values were used to explain the influence of the selected features on the model's ability to predict the three types of targets. The frequencies' median and q25 features have shown to have biggest influence on

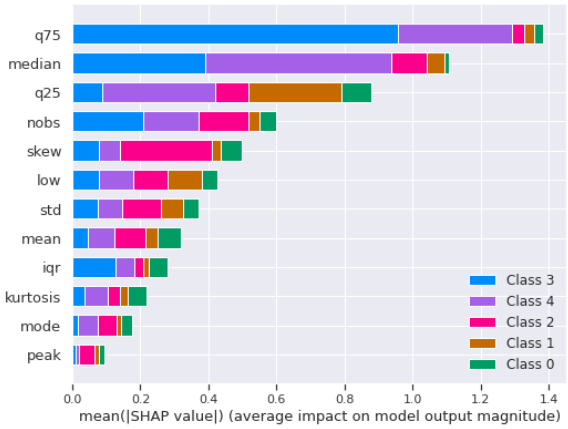

Fig. 4. The most important features for the gender and age range prediction.

the models ability to decide the gender and the age of the speaker.

As there is a disbalance related to gender, the next steps would be to include more speakers so that a more balanced dataset is available as this problem might influence the accuracy of the experiments. Future work would also include implementation of the Macedonian language as part of the dataset.

REFERENCES

[1] S. S. I. Badhon, M. H. Rahaman, and F. R. Rupon. A machine learning approach to automating bengali voice based gender classification. In *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, pages 55–61. IEEE, 2019.

[2] M. Buyukyilmaz and A. O. Cibikdiken. Voice gender recognition using deep learning. In *Proceedings of 2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA2016). Paris, France: Atlantis Press*, volume 10, 2016.

[3] S. Furui. An overview of speaker recognition technology. *Automatic speech and speaker recognition*, pages 31–56, 1996.

[4] H. Harb and L. Chen. Voice-based gender identification in multimedia applications. *Journal of intelligent information systems*, 24(2):179–198, 2005.

[5] M. Li, K. J. Han, and S. Narayanan. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language*, 27(1):151–167, 2013.

[6] J. Lo. Sound frequencies of language. 2018.

[7] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

[8] G. Tzanetakis. Audio-based gender identification using bootstrapping. In *PACRIM. 2005 IEEE Pacific Rim Conference on Communications, Computers and signal Processing, 2005.*, pages 432–433. IEEE, 2005.

[9] Voxforge.org. Free speech... recognition (linux, windows and mac) - voxforge.org. http://www.voxforge.org/, 2022.