School of Information Technology and Engineering at the
ADA University

School of Engineering and Applied Science at the
George Washington University

# AGE-GROUP AND GENDER IDENTIFICATION FROM SPEECH IN AZERBAIJANI LANGUAGE

A Thesis
Presented to the Graduate Program of Computer Science and Data Analytics
of the School of Information Technology and Engineering
ADA University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computer Science and Data Analytics
ADA University

By
Farida Aliyeva

April, 2022

THESIS ACCEPTANCE

This Thesis by: Farida Aliyeva
Entitled: *Age-group and Gender identification from Speech in Azerbaijani Language*

has been approved as meeting the requirement for the Degree of Master of Science in Computer Science and Data Analytics of the School of Information Technology and Engineering, ADA University.

Approved:

Abzatdin Adamov
(Adviser) (Date)

Abzatdin Adamov
(Program Director) (Date)

Sencer Yaralan
(Dean) (Date)

# ABSTRACT

Our Speech comprises of paralinguistic features such as: identity, age, gender, accent etc. The objective of this paper is to identify, the age-group and gender of the speaker from Azerbaijani speech. The paper will be focused on both the adult and children speech identification. The identification of the age and gender of children speech is more complex than the adult's speech as the voice of boys and girls before puberty coincide and additionally the puberty complicates the distinction between an adult and a teenager which results in possible errors with age-group identification. Moreover, the existence of numerous accents of Azerbaijani data an additional milestone. To identify age and gender from the speech the data should be pre-processed and then the features extracted. Next step is to classify according to results obtained. Various approaches are going to be tested such as x-vectors and i-vectors that are based on Deep Neural Network architecture. Then there is MFCC - a feature extraction technique a part of Automatic voice processing for unique feature extraction. On top of that the GMM-SVM model which is a Gaussian mixture model is run. KNN and MLP are another prominent approach to be used as a classifier for age and gender identification problem. Another feature extraction technique called SDC – shifted delta cepstral coefficient will be tested and compared with the MFCC results. The music and audio analysis package called Librosa and a PyAudio library are used to enable the record and play of an audio for demonstration purposes in the future. There outcome of the model is going to be classified into 4 age groups which are: Children (7-14), Young aged (15-24), Middle aged (25-54) and Seniors (55-80) and 2 genders: Male and Female. A proper identification of age and gender is sometimes a hard task for a human being as well which complicated the identification process for the machines.

**Additional Keywords and Phrases:** GMM, MFCC, SDC, PyAudio, Librosa, DNN, KNN, MLP, Paralinguistics, X-vectors, I-vectors.

## TABLE OF CONTENTS

## LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| Abbreviation | Explanation |
|---|---|
| GMM | Gaussian Mixture Model. |
| KNN | K nearest neighbors. |
| SVM | Support Vector Machine. |
| DTC | Decision Tree Classifier. |
| RFC | Random Forest Classifier. |
| DNN | Deep Neural Network. |
| MLP NN | Multilayer Perceptron Neural Network. |
| MFCC | Mel-frequency cepstrum. |
| RMSE | Root mean square. |
| ZCR | Zero Crossing Rate. |
| PCA | Principal Component Analysis. |
| CNN | Convolutional Neural Network. |
| WAV | Waveform Audio file. |

# 1 INTRODUCTION

Speech is a complex phenomenon that involves numerous anatomical systems moving in unison to influence the overall speech and voice quality. [2] Speech is the most common and straightforward mode of communication. It also includes speaker-dependent para-linguistic data such as the speaker's identity, emotional state, health, age, and gender, in addition to linguistic data. [8] Daily communication with people is improved by utilizing this sources of data.

With the rapid technological advancements, Multimedia Retrieval, and Human-computer interaction (HCI) technologies are becoming increasingly vital. For that reason, paralinguistic analysis, where age and gender detection are of primary focus, becomes a fast-emerging topic of research. [11] Automatic speech extraction systems could be beneficial in a variety of applications. These are personal identification in financial banking systems, customer service applications such as call centers, voice bots, interactive and intelligent voice assistants. There are numerous global companies, namely Google, Amazon, and others, providing applications for such services of speech processing in English. Another possible application of using the information about the age and gender of the speaker is the IVR which stands for Interactive Voice Response system. [19] The system is aimed to redirect the speaker to an appropriate consultant for the given age/gender group, or it can be used to send speakers to a particular branch of an interactive voice response (IVR) scenario that is tailored to the speaker's age. [19] Client profiling is performed in call centers by categorizing or ordering speakers into age groups, which is the foundation for critical applications such as targeted advertising, and service customization. In terms of voice-bots, the para-linguistic information can be helpful to change the potential behavior of the bot. Advertisements play a crucial role in today’s market and age/gender information that is retrieved from the speaker’s speech is extremely helpful in this matter. [9] This knowledge can be used to target the ads or navigate through the search that is specific for the given age/gender group. Altogether, the process of exploitation of para-linguistic data can result in a more advanced user experience, which in its turn leads to high revenue for the company that decides to deploy such systems. By means of gender information, gender-dependent acoustic modules are created to foster and enhance the speech recognition process. It is used in broadcast news subtitling systems to change the color of subtitles. This helps people with inabilities to differentiate between the speakers and corresponding subtitles. [9] Another concern is that the majority of gender classification systems are only trained to differentiate between male and female adult voices. Children's voices are comparatively rare. The reason behind this might be that most detectors don't attempt a three-class distinction because it is challenging to gather huge corpora of children's voices. However, the recognition of children's voices is very crucial in some applications, such as the automatic detection of child abuse films on the internet.

The range of possible benefits ranges not only in the Marketing and Sales fields but can also be applied in security checks and forensic science. It has always been complicated process to confirm the identity of an individual. It required a high level of credibility as well as accuracy. Therefore, the law enforcement authorities have been concerned about several biometric procedures. [27] Fingerprint patterns, facial traits, hand geometry, signature dynamics, and voice patterns are all biometric characteristics that can be utilized for forensic identification. The reliability of a method in a certain application, as well as the data provided, are factors to consider. The evidence in some of the criminal cases can appear in the form of recorded conversations. For that reason, speech patterns are considered as crucial information in investigation process according to law enforcement officers. A person's speech pattern, for example, can reveal information about his or her age, gender, accent, emotional or psychological state, and social or regional group membership. As a result, speech may be used for speaker identification, which is critical in many situations like kidnapping, threatening calls, and false alarms. [27] The following study is mainly focused on speaker gender detection and age estimation from Azerbaijani Speech. According to the resources, the following study has been applied and tested in English, German, and Turkish languages. The paper is providing a novice resolution to the same problem but for the Azerbaijani language specifically. The language specifications overcomplicate the work as there is a lot to consider during the data pre-processing, specifically the accents, speed of talk, manner of talk, etc. Although the Turkish language is the closest to Azerbaijani among all others listed above, the following research will require the collection of Azerbaijani-specific data for the purpose of achieving higher and more accurate results. Moreover, gender and age perceptions have a substantial mutual impact on one another; hence these two features have been investigated jointly in several articles. From various perspectives, computerized speech-based age assessment is tricky. First of all, there is frequently a distinction between a speaker's perceived age, also called a perceptual age, and their actual age (chronological age). [9] Second, creating a reliable age recognition system necessitates a well-labeled, diverse, and balanced database. For that specific reason, the data collection process requires the collection of an even amount of gender and age-specific audio samples. In other words, the dataset should not have more voice recordings dedicated to a young (15-24 y.o) group rather than seniors (55-80 y.o). Finally, speech patterns are also affected by numerous factors such as weight, height, and emotional state. This implies that there is a significant variability among speakers. The variabthat is not connected to or merely correlated with age.

Several approaches have been developed and used up until now for the following topic. The starting point of the following task is the pre-processing of the given speech signal. Speech processing is mainly dependent on the features that are correctly extracted from the voice signal using a proper feature extraction tool. One of the most used spectral feature sets is MFCCs, which can simulate the vocal tract filter in a short time power spectrum [23]. Then there is an SDC feature which is known as useful for language identification tools, speaker recognition, and verification [24]. SDC can be thought of as a development of the delta cepstral features. Its main objective is to outperform the derivative features in the cepstral features.

Next in turn the training and classification of the pre-processed data. One of the first and most influencing models is a Deep Neural Network (DNN) which is widely used in a variety of disciplines. Although DNN is well-known for applications in computer vision and image processing problems, it has a great influence on the following topic as well which is its deep architecture. SVM – Support Vector Machine is used to classify age groups and emotions. For the following problem, there are going to be 7 groups: Male Young/ Female Young, Male Middle/ Female Middle, Male Adult/ Female Adult, Male Senior/Female Senior, and Children. Groups will help to find the proper matching age/gender group for the speaker. [22] The GMM super vector model is used in SVM work. Until recently the Gaussian mixture models (GMM) have become the majoring modeling strategy for speaker verification problems. GMM- UBM which is the Gaussian Mixture Model with Universal Background Model is another method that takes the training data of individual classes to train the GMM. Additionally, it uses all training data or another dataset to build the UBM. Thus, it creates a fast speaker/speech or class-dependent model for each voice recording. SVM in its turn uses high-dimensional super vectors produced by GMM. The suggested combination of GMM-UBM/SVM technique combines GMM UBM's generative with SVM's discriminative capacity.

There have been several models used for this specific domain which are MLP, KNN, SVM, GMM/SVM, DNN, and others but investigating the results obtained, the highest accuracy results were obtained with the GMM/SVM. However, the following paper will discuss several possible implementations and combinations and test the models which will finally give the overview and comparison along with the best results obtained.