GENDER IDENTIFICATION AND AGE ESTIMATION OF ARABIC SPEAKER USING MACHINE LEARNING

[1]FIRAS IBRAHIM, [2]KHALID M.O. NAHAR, [2]MOY'AWIAH A. AL-SHANNAQ

[1]Department of Information Systems and Networks, The World Islamic Sciences and Education University, Amman-11947, Jordan

[2]Department of Computer Sciences, Faculty of Information Technology and Computer Sciences, Yarmouk University, Irbid-21163, Jordan

ABSTRACT

Gender Identification and Age estimation is an important topic in the field of Automatic Speech Recognition (ASR) systems. In the field of robotics, for example, it is important to identify human sex and age for emotion recognition and robot interaction. In this paper, we targeted Arabic speakers by identifying their genders and estimating their ages. Experiments were conducted using six famous learning algorithms such as NB-Tree (Decision Tree), Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Naïve Bayes (NB). We focused on the accuracy with some important classification measurements. Automatic gender identification and age estimation system is proposed based on extracting MFCC features from Arabic speech. The MFCC features with the machine learning algorithms were applied to determine whether the speech sample is male or female and assign the approximate age slice. Two experiments were done, the first one targeted gender identification. The results of the first experiment showed that the learning algorithms SVM and ANN were superior in gender determination with accuracies 98.5% and 96.5% respectively. The second experiment targeted age estimation. The results of this experiment showed that Decision Tree (NB-Tree), and Random Forest (RF) were superior in age estimation with accuracies 95.9%, and 93.0% respectively.

Keywords: *Age Estimation, Automatic Speech Recognition, Gender Identification, Machine Learning, MFCC, Artificial Neural Network (ANN).*

1. INTRODUCTION

The key rule of correspondence is principally to trade thoughts among companions and companions. Individuals for the most part impart and see each other through discourse. Be that as it may, this may demonstrate hard for certain individuals because of the wide assortment of dialects spoken all inclusive. These days, numerous PC applications in the territory of computational etymology, have been intended to contemplate the issue of perceiving and deciphering communicated in language [1]. The profitability of such programming applications factually improves and enhances the numerous controls, which exist inside the normal language handling field. In the most recent decade, PC researchers have given exceptional consideration to creating proficient calculations to perceive verbally expressed words in the Speech Recognition (SR) space. Progress in this area has been critical, and it is presently broadly known as Automatic Speech Recognition (ASR) innovation [2][3].

ASR has been created to perceive voices, which can improve correspondence among people. These uses of ASR can make up for the troubles which are brought about by the presence of such a wide scope of dialects on the planet [4]. Along these lines, a few strategies have been presented in the zone of ASR [5][6]. The Support Vector Machine (SVM) is one of amazing ASR procedures that has been generally utilized for discourse acknowledgment [7][8]. An ASR-based framework perceives verbally expressed words by distinguishing and investigating the info voice in a waveform, as showed in Figure 1 the structure of ASR.

Age and gender recognition is the process of identifying the age and estimates the gender information from the uttered speech. This technique enables speech recognition systems to personalize the ads according to the person’s age and gender and also it can have some uses in criminal cases since most of the proofs are of telephone speeches. Also, it can be used for processing waiting for queue music for different genders and age groups. Not all people appreciate the same type of music. Older people might like slow music whereas younger people might like rock or metal music. Another usage of this system can be to try to understand the age and gender distribution of a population in an experimental study which gives more details about the experiment.

Indirect communication, people estimate the attributes of other people, such as age and gender, simply by looking at them and listening to their voices. In social interactions, age, and gender are the most important factors Determine an appropriate manner of social interaction [19].

Since the Arabic Language is one of the most using Languages in the world and the process of listening to a low-quality recording of his/her voice allowed us to estimate certain characteristics of an unknown speaker. This paper deals with automatic gender and age estimation from speech records using a recognizer based on Gaussian Mixture Model (GMM).

This paper is organized to have an introduction in section 1, the related previous attempts in section 2 were some information is given in subsection about speech recognition in general and Arabic speech recognition. Meanwhile, a brief description of the approach is given in the methodology section (3). In section 3, subsections about dataset formulation and feature extraction are given. In section 4 the experiments, the measurements used, and the accuracy is evaluated. In section 5 the conclusion and future work are introduced. Finally, an acknowledgment part was used followed by a big set of references.

2. RELATED WORKS

Safavi et al. (2018), the recognition of the speaker's identity, gender and age group from children's speeches is studied. In that study, the performances of several classification methods including GMM-UBM, GMM-SVM and i-vector based approaches are compared. In the tests, it is observed that the rate of speaker recognition increases with increasing age, but the effect of age on gender and age group recognition is more complex. In the same study, the use of different frequency bands in speaker [20].

Qawaqneh et al. (2017), designed a classifier for gender estimation based on knowledge gained from dependencies among gender, age and pose facial attributes. Deep neural networks were used for gender and age classification from facial images and speech [21].

Shahin (2013), focused on improving emotion identification performance and accuracy based on a two-stage recognizer that is composed of a gender recognizer followed by an emotion recognizer. This work is a gender-dependent, text-independent and speaker-independent emotion recognizer. Both HMMs and SPHMMs have been used as classifiers in the two-stage architecture [22].

Dobry et al. (2011), presents a novel dimension reduction method that improves the accuracy and the efficiency of speakers age estimation systems based on speech signal. Two different gender-based age estimation approaches were implemented, the first age group (Senior, Adult, and Young) classification, and the second, accurate age estimation using regression technique [23].

Bahari and Hamme (2011), introduces a new gender detection and an age estimation approach. To create this strategy, after determining an acoustic model for all speakers of the database, Gaussian mixture weights are extracted and concatenated to create a supervector for each speaker. Then, hybrid architecture of WSNMF and GRNN is developed using the supervectors of the training data set [24].

Meinedo and Trancoso (2010), present gender detection is a very useful task for a wide range of applications. In the Spoken Language Systems lab of INESC-ID, the Gender Identification module is one of the basic components of our Voice processing system, where it is mainly used for speaker clustering, to avoid mixing speakers from different genders in the same cluster. Gender information (male or female) is also used for building gender-dependent acoustic models for speech recognition [25].

Table 1. Related Word Summary

Author	Methodology	Dataset Used	Accuracy
Safavi et al. (2018),	GMM & SVM	OGI Kids Speech corpus	85.8%
Qawaqneh et al. (2017)	Deep Neural Networks	private corpus	63.78%
Shahin (2013)	HMMs & SPHMMs	Specific Dataset	79.92 %
Dobry et al. (2011),	novel dimension reduction. GMM & SVM	LDC's Switchboard corpus	79.0%
Bahari and Hamme (2011),	WSNMF & GRNN	The N-best evaluation corpus	96.0%
Meinedo and Trancoso (2010)	GMM-UBM, MLP and SVM	Four different corpora were used	83.1%

2.1 Speech Recognition System (Srs)

Discourse acknowledgment alludes basically to the applications that can recognize spoken expressions deciphered by a mechanized machine. The fundamental point of SRS is to change over the perceived words and expressions into a comprehensible organization by the machine-programs. SRS is isolated into two sections: phoneme-based and syllable-based models. The SRS typically relies upon models utilized in language figuring out how to improve the pace of acknowledgment exactness [9][10].

Automatic Speech Recognition (ASR) empowers the PC to distinguish the expressions of an individual talking into a mouthpiece or phone. Human-Computer Interactions (HCI) is utilized, for instance, as a validation procedure for client login using a voice acknowledgment apparatus. Some ASR applications incorporate voice interface as an order acknowledgment application for PC clients, transcription and composed content remedy, intuitive correspondence, voice reaction, as a guide in learning unknown dialects, and for voice-controlled activity of machines. Also, ASR innovation can improve personal satisfaction for impaired individuals permitting them to speak with others and interface in the public arena [8]. Practically speaking, the primary standard of an ASR framework is to perceive the proper word designs for verbally expressed articulations by applying a digitized investigating procedure to enter simple sound waves [10].

2.2 Arabic Speech Recognition

One of the most established Semitic dialects on the planet is the Arabic language. It is formally the 6th most communicated in language and one of the official dialects of the United Nations. There is an official Arabic etymological structure known as Modern Standard Arabic (MSA), which is primarily utilized informal media, courts, workplaces, and by educators for instructing in schools and colleges [11]. As of late, numerous ASR frameworks have been created in the area of Arabic Speech Recognition to perceive the MSA variant of Arabic. Sadly, perceiving conventional Arabic is as yet a test due its lexical assortment and the shortage of information. Besides, the Arabic language is viewed as one of the most unpredictable dialects because of the morphological varieties of its letters [12][10].

ASR Arabic language research is still in its early stages age contrasted with the ASR previously utilized in research identified with different dialects [13][14]. Thus, we will audit the main five investigations that have been created to improve Arabic discourse acknowledgment. In [15], the creator managed consistent Arabic discourse acknowledgment, tending to the marking of Arabic discourse. Another model for Arabic discourse acknowledgment concentrated on unmistakable issues in the perceiving of conversational, regional and everyday Arabic discourse [13]. The creators detailed critical improvement as for word blunder rate as indicated by the 1997 NIST benchmark assessments. An Arabic ASR framework utilizing ANN procedures was created to improve the Arabic programmed acknowledgment process [15][16]. Another Arabic ASR dependent on Hidden Markov Models (HMM), SVM or a half and half of the two was additionally evolved [14][17]. The last zone identifies with crafted by some Arabic ASR specialists who contemplated articulation varieties to improve the presentation of Arabic ASR frameworks [18].

3. METHODOLOGY

The research methodology depends on extracting MFCC features from acoustic signals of corpora and supervised learning models. The features then passed to a set of machine learning algorithms for training and validation. Part of the acoustic signals for both males and females were used for testing. After that results of both gender and age estimation are gained and compared. Figure 1 illustrates the methodology steps.

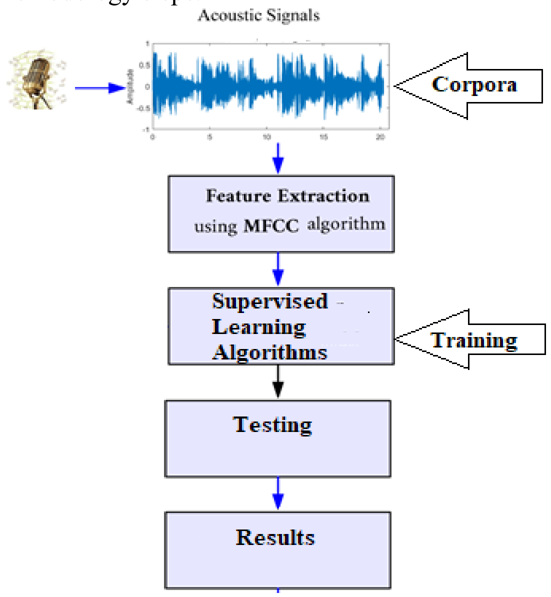


Figure 1. Steps of The Proposed Approach

In the testing phase, some external data which sometimes called outlet data (data totally outside the corpus) will be used for testing and evaluating the proposed approach. The following subsections show more details about our methodology. In the following subsections detailed information about each part of the proposed approach

3.1 Data Acquisition

The Urban Jordanian corpus was used[1]. In this corpus, 12 native speakers of Urban Jordanian Arabic were recorded (6 females, 6 males). Urban Jordanian Arabic (UJA) is spoken by people living in the major cities of Jordan more than two-thirds of the population of Jordan. The speech files contain a variety of consonants differing in place and manner of articulation as well as vowels differing in both quality and length. Moreover, males and females in these corpora were of different ages. This corpus is based upon work supported by the National Science Foundation (NSF) under Grant No. 0518969 (Acoustic and perceptual correlates of Emphasis in Arabic, Allard Jongman, P.I.). Figure 2 shows a sample of a male signal along with its full spectrum, MFCCs, and the spectrum without any weak energy.

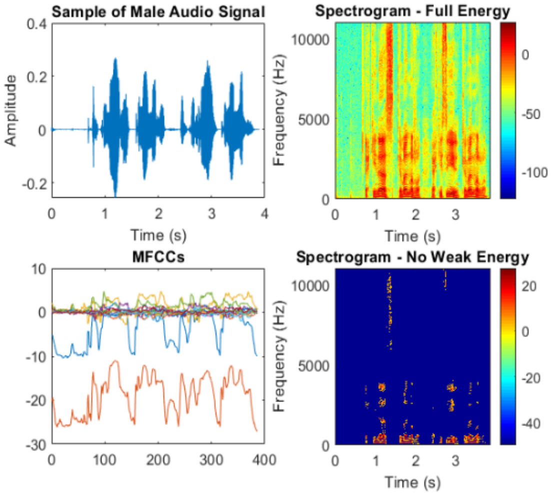


Figure 2. Sample of a Male Acoustic Signal

The audio signals of all speakers were recorded at the Hashemite University in Zarqa, Jordan using a Marantz PMD671 portable solid-state recorder and an Electro-Voice N/D767a microphone. The Sampling rate of the audio files was 22.05 kHz.

Figure 3 shows another acoustic signal of a female recording were the spectrum and MFCCs are clearly differe from each other.

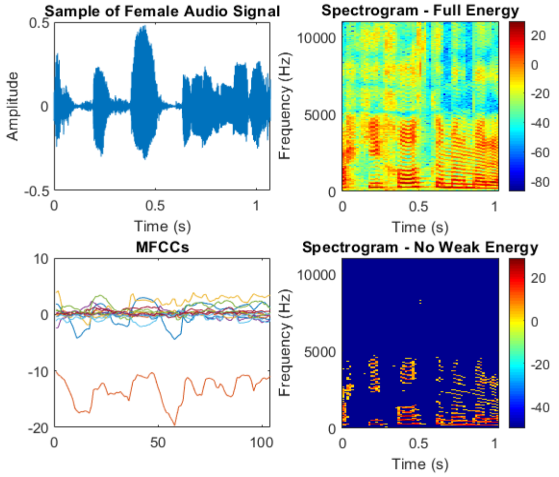


Figure 3. Sample of a Female Acoustic Signal

Additionally, we contact Hashemite University in Zarqa, and they provide us with males and females ages which covers almost all ages slots in Jordan. Moreover, some additional recordings were collected from the internet for Jordanian speakers and their speech signals were divided into segments of 15 ms frames. The gender and age of the internet recordings also pointed in the corpus.

3.2 Building The Feature Vectors

Using MATLAB tool, the MFCC features were extracted from all wave sounds. MFCC is a brief period power range that is utilized to speak to sound waves [57]. Mel frequencies depend on the basic transmission capacity of the human ear perceived as a variety with recurrence channels, which incorporates two kinds of frequencies [58]. The first at frequencies under 1 kHz, and the second logarithmic channels at frequencies higher than 1 kHz to catch phonetically significant attributes [59]. The stages engaged with MFCC extraction are shown in Figure 4.

[1] https://kuppl.ku.edu/corpus-arabic-recordings

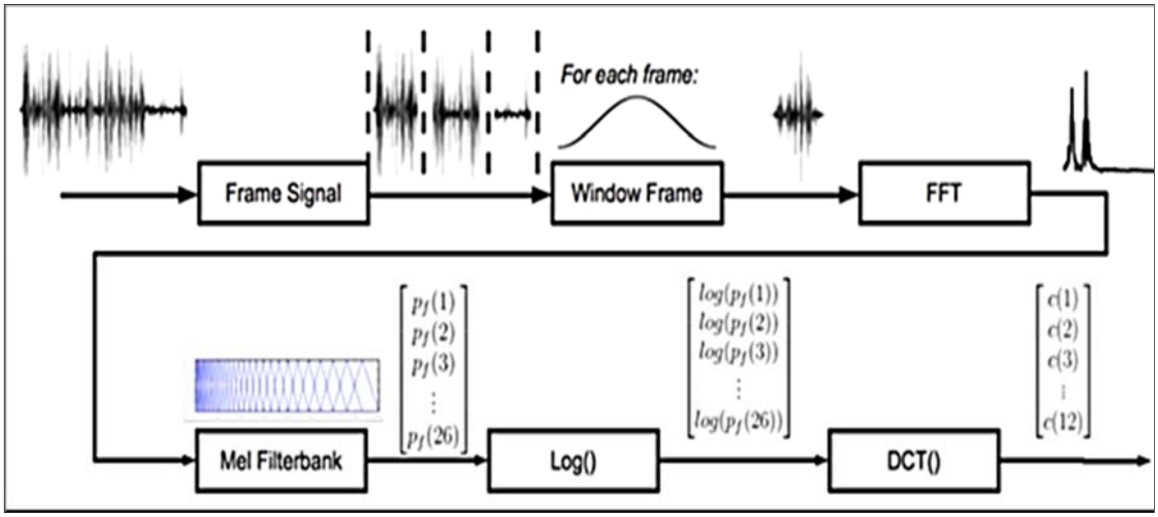

For each frame:
Frame Signal
Window Frame
FFT
Mel Filterbank
$p_f(1)$ $p_f(2)$ $p_f(3)$ ⋮ $p_f(26)$
Log()
$log(p_f(1))$ $log(p_f(2))$ $log(p_f(3))$ ⋮ $log(p_f(26))$
DCT()
c(1) c(2) c(3) ⋮ c(12)

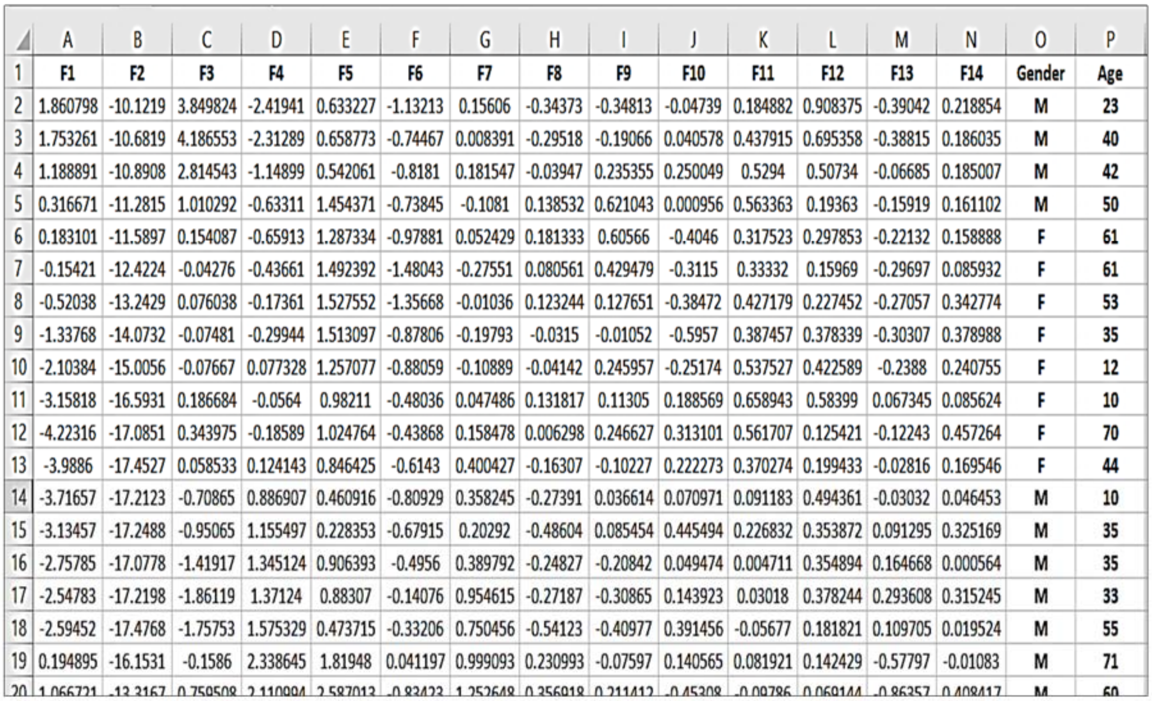
Figure 4. MFCC Feature Extraction Stages [54]

The MFCC is calculated using Eq. (1), where its implementation already provided by MATLAB tool.

$$C_i = \sum_{k=1}^{N} X_k cos\left(\frac{[\pi_i(k-0.5)]}{N}\right), \quad for\ i = 1,2, \dots, p \qquad (1)$$

Where C_i are the Cepstral coefficients, p is the order, k is the number of discrete Fourier transformations magnitude coefficients, X_k is the k^{th} order log-energy output of the filter bank, and N is the number of filters (usually 20). Thus, 13 coefficients and an energy feature were extracted, generating a vector of 14 coefficients per-frame.

The features vectors of all utterances were collected together in one CSV file with the corresponding gender and age of the speaker. Figure 5 shows part of the features, were the last two columns are the gender, and age targets.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	Gender	Age
2	1.860798	-10.1219	3.849824	-2.41941	0.633227	-1.13213	0.15606	-0.34373	-0.34813	-0.04739	0.184882	0.908375	-0.39042	0.218854	M	23
3	1.753261	-10.6819	4.186553	-2.31289	0.658773	-0.74467	0.008391	-0.29518	-0.19066	0.040578	0.437915	0.695358	-0.38815	0.186035	M	40
4	1.188891	-10.8908	2.814543	-1.14899	0.542061	-0.8181	0.181547	-0.03947	0.235355	0.250049	0.5294	0.50734	-0.06685	0.185007	M	42
5	0.316671	-11.2815	1.010292	-0.63311	1.454371	-0.73845	-0.1081	0.138532	0.621043	0.000956	0.563363	0.19363	-0.15919	0.161102	M	50
6	0.183101	-11.5897	0.154087	-0.65913	1.287334	-0.97881	0.052429	0.181333	0.60566	-0.4046	0.317523	0.297853	-0.22132	0.158888	F	61
7	-0.15421	-12.4224	-0.04276	-0.43661	1.492392	-1.48043	-0.27551	0.080561	0.429479	-0.3115	0.33332	0.15969	-0.29697	0.085932	F	61
8	-0.52038	-13.2429	0.076038	-0.17361	1.527552	-1.35668	-0.01036	0.123244	0.127651	-0.38472	0.427179	0.227452	-0.27057	0.342774	F	53
9	-1.33768	-14.0732	-0.07481	-0.29944	1.513097	-0.87806	-0.19793	-0.0315	-0.01052	-0.5957	0.387457	0.378339	-0.30307	0.378988	F	35
10	-2.10384	-15.0056	-0.07667	0.077328	1.257077	-0.88059	-0.10889	-0.04142	0.245957	-0.25174	0.537527	0.422589	-0.2388	0.240755	F	12
11	-3.15818	-16.5931	0.186684	-0.0564	0.98211	-0.48036	0.047486	0.131817	0.11305	0.188569	0.658943	0.58399	0.067345	0.085624	F	10
12	-4.22316	-17.0851	0.343975	-0.18589	1.024764	-0.43868	0.158478	0.006298	0.246627	0.313101	0.561707	0.125421	-0.12243	0.457264	F	70
13	-3.9886	-17.4527	0.058533	0.124143	0.846425	-0.6143	0.400427	-0.16307	-0.10227	0.222273	0.370274	0.199433	-0.02816	0.169546	F	44
14	-3.71657	-17.2123	-0.70865	0.886907	0.460916	-0.80929	0.358245	-0.27391	0.036614	0.070971	0.091183	0.494361	-0.03032	0.046453	M	10
15	-3.13457	-17.2488	-0.95065	1.155497	0.228353	-0.67915	0.20292	-0.48604	0.085454	0.445494	0.226832	0.353872	0.091295	0.325169	M	35
16	-2.75785	-17.0778	-1.41917	1.345124	0.906393	-0.4956	0.389792	-0.24827	-0.20842	0.049474	0.004711	0.354894	0.164668	0.000564	M	35
17	-2.54783	-17.2198	-1.86119	1.37124	0.88307	-0.14076	0.954615	-0.27187	-0.30865	0.143923	0.03018	0.378244	0.293608	0.315245	M	33
18	-2.59452	-17.4768	-1.75753	1.575329	0.473715	-0.33206	0.750456	-0.54123	-0.40977	0.391456	-0.05677	0.181821	0.109705	0.019524	M	55
19	0.194895	-16.1531	-0.1586	2.338645	1.81948	0.041197	0.999093	0.230993	-0.07597	0.140565	0.081921	0.142429	-0.57797	-0.01083	M	71
20	1.066721	-13.3167	0.759508	2.110094	2.587013	-0.83423	1.252648	0.356918	0.211412	-0.45308	-0.09786	0.069144	-0.86357	0.408417	M	60

Figure 5. Feature Vectors .CSV File

4. EXPERIMENTAL SETUP AND RESULTS

Many researchers in other languages have used classical classifiers such as Support Vector Machine (SVM) and Artificial Neural Networks (ANN), Random Forest, K-Nearest Neighbor (KNN), Naïve Bayes and Decision Tree which gave good results [26]. We decide to apply these famous machine learning algorithms and compare the results.

We have used Orange Software tool for Machine Learning (ML) and Big Data (BD) to build the models. The feature corpus is divided into two parts: the first part forming 80% from the dataset for training and the second part forms 20% for validation and testing. Figure 6 shows the building for the training phase for SVM, ANN with "Relu" activation function, KNN with k=11 and the other models. Two main experiments were done one for gender identification, and one for Age estimation using the same model.

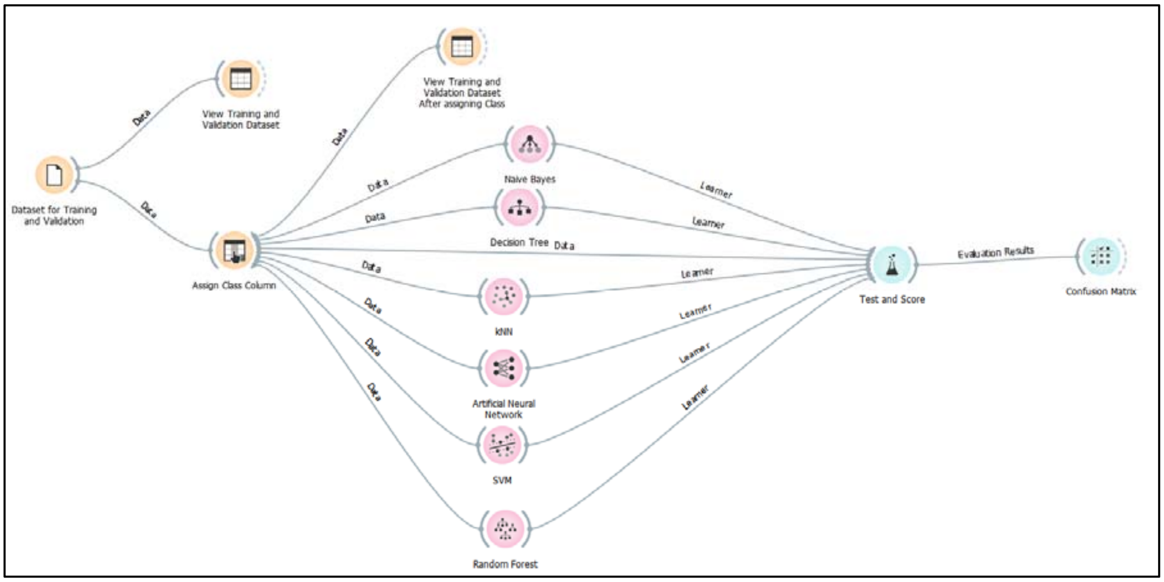


Figure 6. Model Architecture for Six Learning Algorithms Using Orange Tool

4.1 Evaluation & Analysis Metrix

Evaluation of the model is an important step for any learning model. One of the most used metrics is the accuracy score metric, it is a good metric to evaluate the model. Sometimes using the accuracy score metric alone is not enough. Therefore, some other metrics were used such as F1 score, Precision, and Recall. These measurements mainly depend on miss and hit prediction of a class. In these hits and miss classification indicators are summarized.

Table 2. Classification Indicators

The Measurement	The Meaning
TP	Number of true positives (instances correctly classified as a given class).
FP	Number of false positives (instances falsely classified as a given class).
FN	Number of incorrect classification of positives instances.
Precision (p)	Proportion of instances that are truly of a class divided by the total instances classified as that class
Recall (r)	Proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate).

The next sections explain these measurements in detail.

4.1.1. Precision

Precision was used to check the classifier's ability to return relevant instances only. Equation 2 represents this metric. Simply it is the number of correct positive results divided by the number of the positive results predicted by the algorithm[2].

[2] https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used

$$\text{Precision} = \frac{\text{TP}}{\text{T P + FP}} \qquad (2)$$

4.1.2. Recall

Recall (also known as sensitivity) is used to know the classifier's ability to identify all relevant instances. The equation used to calculate it is Equation 3. It is the number of correct positive results divided by the number of all relevant samples[2].

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \qquad (3)$$

4.1.3. F-Measure

F-Measure is used to combine Precision and Recall into one measurement metric. It uses the harmonic means to combine them. Equation 4 is used to calculate this measure[2].

$$\text{F1} - \text{Measure} = 2 * \frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \qquad (4)$$

4.1.4. Accuracy

Accuracy is the most popular used performance measure. It is known as the ratio of correctly predicted observation of the total observations. Accuracy would give us a good indicator and strong evaluation only when the dataset is balanced. Since our data is not balanced, we used the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) counts to measure the accuracy. Equation 5 represents the definition of accuracy and Equation 6 is the accuracy based on previous counts when the dataset is not balanced[2].

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \qquad (5)$$

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP + TN + FP + FN}} \qquad (6)$$

4.2 Gender Identification Experiment

It is conceivable to recognize the sex or gender of a speaker with a precision of practically 100% by tuning in to his voice by human ears. This paper tests result to demonstrate that it is conceivable to recognize the sex naturally with results near emotional sex estimation by human audience members. After running the model and by targeting the gender as the target class, the results are shown in Table 3.

Table 3. Gender Experimental Results of Classification

ML-Algorithm	Accuracy	Precision	Recall	F-Measure
ANN	0.965	0.920	0.911	0.92
SVM	0.985	0.979	0.989	0.991
Random Forest	0.541	0.538	0.542	0.541
Decision Tree	0.459	0.459	0.46	0.459
Naïve Bayes	0.438	0.408	0.423	0.438
KNN	0.343	0.341	0.344	0.343

It clear from Table 3, ANN and SVM showed superiority over other ML-Algorithms especially SVM where the accuracy reaches 98.5%. a pictorial view of the gender experimental result is shown in Figure 7.

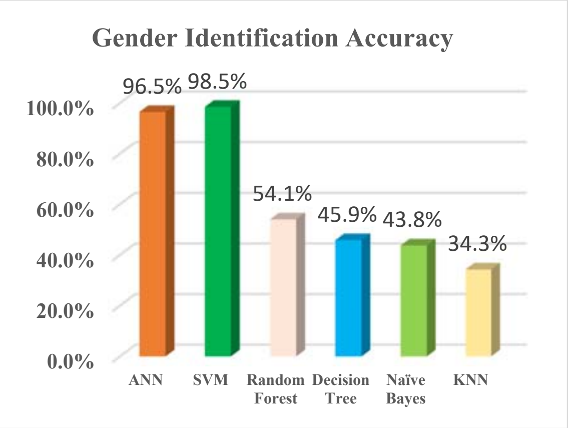


Figure 7. Gender Identification Accuracy

4.3 Age Estimation Experiment

Age estimation is more troublesome than sexual orientation distinguishing proof. Exact age estimation is incomprehensible even by human audience members. The estimation is constantly upset by certain deviation between the genuine speaker age and the assessed age.

When utilizing entire sentences in the event of average speakers, the error is for the most part not more prominent than 10 years. When utilizing short words if there should arise an occurrence of atypical speakers, the mix-up can be as long as 50 years [12].

The most predominant voice signs in age estimation are contributed general, a female discourse has a higher pitch (120 - 200 Hz) than male discourse (60 - 120 Hz) yet the vocal power (commotion), jitter and gleam (unpleasantness), the formant frequencies and the ghastly extension (voice quality), the length and pausation are likewise significant. In this paper, just otherworldly envelope spoke to by MFCC coefficients is utilized.

Based on the spectrum power and MFCC features and since the age of the speakers are known, supervised learning solved the issue of age estimation. The accuracy gained in this experiment summarized in Table 4.

Table 4. Age Experimental Results of Classification

ML-Algorithm	Accuracy	Precision	Recall	F-Measure
ANN	0.667	0.666	0.667	0.663
SVM	0.632	0.629	0.636	0.632
Random Forest	0.93	0.938	0.942	0.941
Decision Tree	0.959	0.959	0.96	0.959
Naïve Bayes	0.838	0.808	0.823	0.838
KNN	0.743	0.741	0.744	0.743

It is clear that Random Forest and Decision Tree shows superiority over all others. It is clear that SVM and ANN were the worst accuracies while they were the best in gender identification. A pictorial view of these results is shown in Figure 8

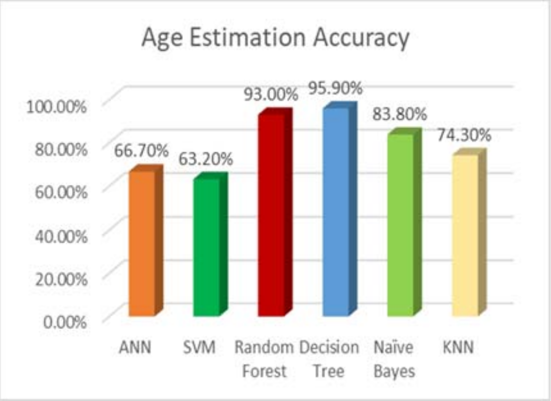


Figure 8. Age Estimation Accuracy

For the sake of comparison, both age and gender accuracies achieved are set together in one graph shown in Figure 9.

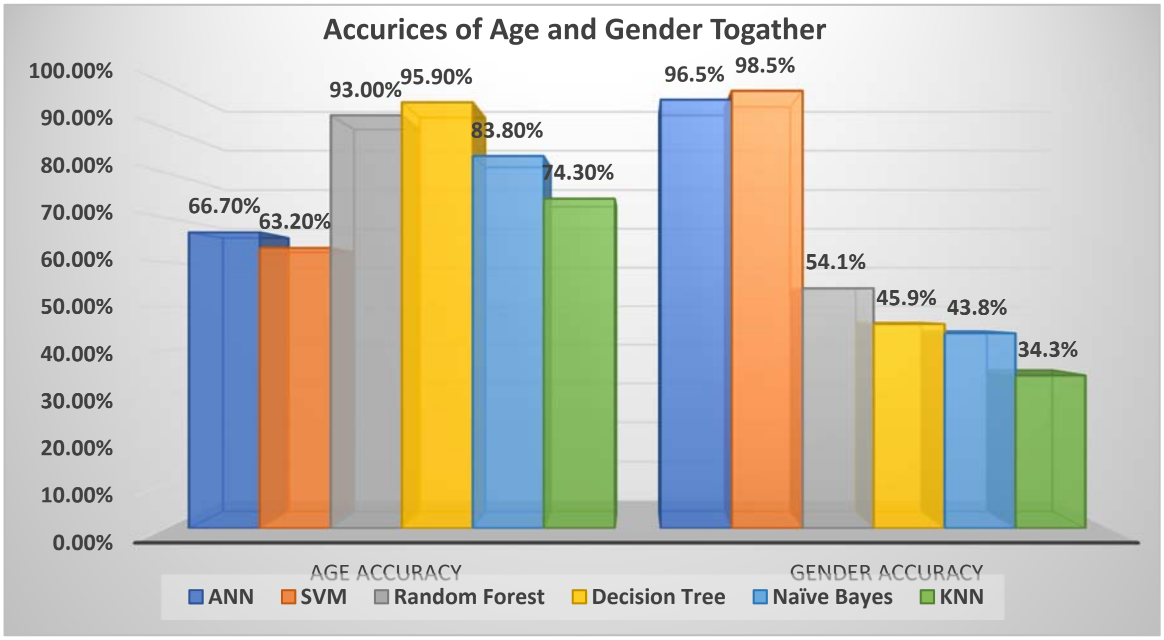


Figure 9. Accuracy of Both Gender and Age in Comparison

5. DISCUSSION AND ANALYSIS

For complex problems of classification is a challenging problem, especially when the data distribution is not linear, and the number of classes is large. In this context, SVM have demonstrated superior performance [2]. However, SVM was originally designed for binary classification and its extension for multi-class classification is still an on-going research issue [3].

Based on the results gained from experiments we notice that SVM and ANN Achieve high accuracy and good generalization. Since SVM is a binary classifier it shows superiority in gender classification.

Meanwhile, in problems where a rule is to be generated or a specific decision has to be taken, the Decision Tree (DT) is useful. In DT algorithm a lot of preparing models are separated into littler and littler subsets while simultaneously a related choice tree gets steadily created. Toward the finish of the learning procedure, a choice tree covering the preparation set is returned. The key thought is to utilize a choice tree to segment the information space into bunch (or thick) districts and unfilled (or inadequate) locales [21].

In DT Classification another model is characterized by submitting it to a progression of tests that decide the class name of the model. These tests are sorted out in a progressive structure called a choice tree. Choice Trees follow Divide-and-Conquer Algorithm [21].

A Random Forest (RF) is a group of n choice trees. Every choice tree in the timberland is prepared on various subsets of the preparation set, produced from the first marked information by stowing [8]. Irregular Forest uses randomized element determination while the tree is growing.

Referring to results in a gender experiment, DT and RF did not show good results since the problem is binary. In age estimation, DT and RF succeeded in achieving high accuracy compared to other ML-algorithms.

6. CONCLUSIONS

This examination expects to explore the degree to which audience members can pass judgment on some obscure speakers' acoustic attributes. Also, we attempted to assess the connection of the ordered ages with different acoustic qualities: pitch, open remainder, phantom tilt, music to-clamor proportion, shine, jitter and MFCC. In this paper, we propose a relative estimation technique for abstract age by utilizing speaker's acoustical attributes and their sequential age.

In this paper, we presented gender Identification and Age estimation model based on machine learning. The importance of this research relay on its use in the fields of Automatic Speech Recognition (ASR) and Artificial Intelligence. For example, in the field of robotics, it is important to identify human sex and age for emotion recognition and robot interaction. In this paper, we targeted Arabic speakers by identifying their genders and estimating their ages. Experiments were conducted using six famous learning algorithms such as NB-Tree (Decision Tree), Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Naïve Bayes (NB). We focused on the accuracy with some important classification measurements. Automatic gender identification and age estimation system is proposed based on extracting MFCC features from Arabic speech. The MFCC features extraction algorithm was performed on the Arabic Audio corpus and feature vectors are formulated. The feature vectors matrix is fed to the machine learning algorithms through a model build by Orange tool (a tool for Machine Learning and Big Data). Our target was to determine whether the speech sample is male or female and assign the approximate age slice for it. Two experiments were done, the first one targeted gender identification, and the second one targeted the age estimation. Results are extracted, Analyzed, and compared through ML-Algorithms. SVM and ANN were superior in gender determination with accuracies 98.5% and 96.5% respectively, while Decision Tree (NB-Tree), and Random Forest (RF) were superior in age estimation with accuracies 95.9%, and 93.0% respectively.

7. FUTURE WORK

As a future work, it is worthy to build a huge corpus for Arabic language specifically for age estimation and gender identification. By this corpus we can precisely estimate the age and decrease the error in the estimation to be as much small as possible. Meanwhile, by building a huge corpus that include all types of Arabic speakers we may achieve 100% of Arabic gender recognition.

ACKNOWLEDGMENT

Authors would like to acknowledge the National Science Foundation since the Arabic corpus was created under a Grant No. 0518969 (Acoustic and perceptual correlates of Emphasis in Arabic, Allard Jongman, P.I.). Moreover, we thank Hashemite University in Zarqa, Jordan where the recordings are hosted.

REFERENCES:

[1] Y. Kato, “Speech recognition system,” *J. Acoust. Soc. Am.*, vol. 107, no. 5, p. 2326, 2000.

[2] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993.

[3] T. K. Das and K. M. O. Nahar, “A Voice Identification System using Hidden Markov Model,” *Indian J. Sci. Technol.*, vol. 9, no. 4, 2016.

[4] J. M. Baker *et al.*, “Developments and directions in speech recognition and understanding, Part 1 [DSP Education],” *IEEE Signal Process. Mag.*, vol. 26, no. 3, 2009.

[5] J. Chong, E. Gonina, D. Kolossa, S. Zeiler, and K. Keutzer, “An Automatic Speech

Recognition Application Framework for Highly Parallel Implementations on the GPU," 2012.

[6] K. Nahar, H. Al-Muhtaseb, W. Al-Khatib, M. Elshafei, and M. Alghamdi, "Arabic phonemes transcription using data driven approach," *Int. Arab J. Inf. Technol.*, vol. 12, no. 3, pp. 237–245, 2015.

[7] Y.-H. Shao and N.-Y. Deng, "A coordinate descent margin based-twin support vector machine for classification," *Neural networks*, vol. 25, pp. 114–121, 2012.

[8] K. Nahar, M. Abu Shquier, W. G. Al-Khatib, H. Al-Muhtaseb, and M. Elshafei, "Arabic phonemes recognition using hybrid LVQ/HMM model for continuous speech recognition," *Int. J. Speech Technol.*, vol. 19, no. 3, pp. 495–508, 2016.

[9] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, "An Architecture for a Generic Dialogue Shell," *Nat. Lang. Eng.*, vol. 6, no. 3–4, pp. 213–228, 2000.

[10] K. M. O. Nahar, W. G. Al-Khatib, M. Elshafei, H. Al-Muhtaseb, and M. M. Alghamdi, "Data-driven Arabic phoneme recognition using varying number of HMM states," in *2013 1st International Conference on Communications, Signal Processing and Their Applications, ICCSPA 2013*, 2013, pp. 1–6.

[11] M. Abdul-Mageed, M. Diab, and S. Kübler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 20–37, 2014.

[12] F. Diehl, M. J. F. Gales, M. Tomalin, and P. C. Woodland, "Morphological decomposition in Arabic ASR systems," *Comput. Speech Lang.*, vol. 26, no. 4, pp. 229–243, 2012.

[13] K. Kirchhoff *et al.*, "Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins summer workshop," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 2003, vol. 1, pp. I--I.

[14] E. Zarrouk, Y. Ben Ayed, and F. Gargouri, "Hybrid continuous speech recognition systems by HMM, MLP and SVM: A comparative study," Int. J. Speech Technol., vol. 17, no. 3, pp. 223–233, 2014.

[15] A. Al-Otaibi, "Speech Processing," Br. Libr. Assoc. with UMI, 1988.

[16] M. M. El Choubassi, H. E. El Khoury, C. E. J. Alagha, J. A. Skaf, and M. A. Al-Alaoui, "Arabic speech recognition using recurrent neural networks," in Signal Processing and Information Technology, 2003. ISSPIT 2003. Proceedings of the 3rd IEEE International Symposium on, 2003, pp. 543–547.

[17] A. Ali et al., "Automatic dialect detection in arabic broadcast speech," arXiv Prepr. arXiv1509.06928, 2015.

[18] F. Biadsy, N. Habash, and J. Hirschberg, "Improving the Arabic pronunciation dictionary for phone and word recognition with linguistically-based pronunciation rules," in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

[19] Yousuke Arai, and Yutaka Matsui, "The relationship of behavior and affection, to that of the comparison standard which occurs within hierarchical-treatment of those who are of close age," Japanese Journal of Interpersonal and Social Psychology , vol. 31, no. 3, pp. 23-28, 2003.

[20] Safavi, S., Russell, M., & Jančovič, P. (2018). Automatic speaker, age-group and gender identification from children's speech. Computer Speech & Language, 50, 141-156.

[21] Qawaqneh, Z., Mallouh, A. A., & Barkana, B. D. (2017). Age and gender classification from speech and face images by jointly fine-tuned deep neural networks. Expert Systems with Applications, 85, 76-86.

[22] Shahin, I. M. A. (2013). Gender-dependent emotion recognition based on HMMs and SPHMMs. International Journal of Speech Technology, 16(2), 133-141.

[23] Dobry, G., Hecht, R. M., Avigal, M., & Zigel, Y. (2011). Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal. IEEE Transactions on Audio, Speech, and Language Processing, 19(7), 1975-1985.

[24] Bahari, M. H., & Van Hamme, H. (2011, September). Speaker age estimation and gender detection based on supervised non-negative matrix factorization. In 2011 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS) (pp. 1-6). IEEE.