Automatic Arabic Speech Recognition- A Comparative Study

Ali Alanazy [*], Mohammed Alatawi[*]

[*] Department of Information Technology, Faculty of Computers and Information Technology, University of Tabuk, The Kingdom of Saudi Arabia

DOI: 10.29322/IJSRP.11.12.2021.p12064
http://dx.doi.org/10.29322/IJSRP.11.12.2021.p12064

Abstract- The cutting-edge studies on Automatic Speech Recognition approach have reported exceptional accuracy rates that are even comparable to human transcribers – posing a question if machine has reached human performance. Automatic Speech Recognition can be used as a biometric authentication technique, which is essential in ciphering many applications used. In light of the Arabic language, only few studies have proposed to assess the effectiveness of using Automatic Speech Recognition in Arabic language; therefore, this study aims to implement Arabic speaker recognition using three different algorithms, including (i) Dynamic Time Warping (DTW), (ii) Gaussian mixture model (GMM), and (iii) Support Vector Machine (SVM). To measure the effectiveness of these algorithm in recognizing the Arabic speech, two datasets are used to train and test them, which are: (i) speech agent archive, and (ii) Arabic speech corpus. The results reveled that the DTW outperforms the GMM and SVM in terms of accuracy, precision, recall and f-measure, as it achieves 95.7%, 96%, and 95%, and 96%, respectively.

Index Terms- Arabic Speech Recognition, DTW, GMM, SVM.

I. Introduction

The rapid advancement of Automatic Speech Recognition (ASR) stems from the progress of machine and deep learning [1-8], which have simultaneously enhanced its performance to be on par with human performance [19]. The conventional ASR approach involves modular design that varies based on language modelling, acoustic modelling, and pronunciation lexicon trained in isolation. Meanwhile, the present end-to-end (E2E) modelling are trained for converting acoustic aspects to text transcriptions and this optimizes all end task fragments [20]. As ASR performance gets closer to HSR, some researchers have benchmarked the performance exhibited by cutting-edge ASR against professional transcribers [21]. For instance, E2E ASR was exceptional in its simple speech recognition task, such as reading the newspaper [21], while Microsoft [22] discovered that the ASR had attained the level of professional transcriber in intricate task, such as conversational speech. Nonetheless, the IBM [21] disclosed that HSR still demonstrated better performance for conversational speech. Notably, these studies assessed the context of the English language, while disregarding morphologically intricate languages, such as the Arabic language. Being the largest Semitic language, the Arabic language has plenty of derivations and affixations that generate a massive number of word forms. Besides, about 400 million Arabic native speakers use Dialectal Arabic (DA) for their daily communication. As the DA has no standard orthographic rule [23] the varied Arabic dialects are regarded as multiple languages. Nevertheless, the Arabs view dialects as deterioration from the classical Arabic, as almost similar Arabic letters are applied in DA. Hence, objectively comparing the variations in DA may arrive at the conclusion that DA is linked historically, but not synchronically. This unique attribute makes the Arabic language a great option to identify speech recognition obstacles.

The three obstacles faced while building ASR for the Arabic language are: (1) As a consonantal language, most texts in the Arabic language are non-discretized. Thus, it is difficult to locate the vowels that may result in varied meaning. (2) The DA has limited labelled data, in which the Arabic written language differs from the spoken one due to non-standard orthographic rule. (3) The morphologically intricate Arabic language with a wide range of derivations and affixations increases the rate of out-of-vocabulary and makes it difficult to construct a comprehensive language model.

However, this article aims to aims to implement Arabic speaker recognition using three different algorithms, including (i) Dynamic Time Warping (DTW), (ii) Gaussian mixture model (GMM), and (iii) Support Vector Machine (SVM).

The rest of this article is organized as follows: Section 2 discusses the prior studies. Section 3 presents the methodological steps followed in this article in detail. The results and discussion are presented in Section 4, and finally, the article is concluded in Section 5.

II. Related works

A hybrid framework of Support Vector Machine (SVM)-Dynamic Time Warping (DTW) algorithm was proposed by Ismail et al. [10] to improve the process of speech recognition. This machine learning system can control smart devices via speech commands at 97% accuracy. As a result, the proposed system could successfully aid those elderly and patients to control IoT devices compatible with the proposed system using ASR. Furthermore, the proposed system has the flexibility and scalability for adapting to available smart IoT devices, thus offering better device management for patients. Essentially, the proposed system can be effectively integrated with medical institution system to facilitate both patients and the elderly.

By using the SVM as the classifier; Alonso et al., [11], Luengo et al., [12], and Cao et al., [13] extracted spectral, prosody, and pitch features from the database of Berlin Emotional Speech

(BES). Alonso et al., [11] reported 94.9% of emotion recognition accuracy based on five emotions (sad, angry, bored, happy, & emotion recognition accuracy in light of seven emotions (sad, angry, indifferent, bored, happy, loath, & scared). On the other hand, Cao et al., [13] revealed a score of 82.1% for emotion recognition accuracy with seven emotions, which are: boredom, anger, sadness, disgust, happiness, fear, and indifferent. By using the prosody features in SVM classifier, Wang et al., [14] found 88.8% of emotion recognition accuracy based on six emotions; anger, happiness, indifferent, sadness, anxiety, and boredom.

The Gaussian Mixture Model (GMM) has also been deployed for emotion recognition in numerous studies [15], [16]. For example, Cheng and Duan [15] grouped five emotions (surprise, indifferent, sad, happy, & angry) by using GMM. In the study, 60 fundamental attributes were combined to yield the feature vector. Next, the Principal Component Analysis (PCA) was deployed to extract the features and sent to the enhanced GMM for both classification and recognition. As a result, the selected features were viable to recognise emotions [15]. In another study, the Gaussian Mixture Vector Autoregressive Model (GMVAM) was proposed by El Ayadi et al., [16] to recognise emotions. When examined on BES dataset, the proposed statistical classifier exhibited 76% of emotion recognition accuracy based on six emotions (sad, indifferent, bored, angry, happy, & scared) [16].

Tashev et al., [22] combined feature extractor (low-level) based on GMM with NN to function as a feature extractor at high level. The proposed approach, which is a combination of classic statistical method with NN-based solutions for recognition of emotions, was assessed using Mandarin database with four emotions (angry, indifferent, sad, & happy). The proposed GMM-DNN resulted in 48.0% and 41.5% for weighted and un-weighted emotion recognition accuracy rates, respectively [22].

In another study, Kanisha et al., (2018) constructed a speech recognition model to enhance extraction of features, thus proposing the improved SVM (ISVM). The Gaussian filter was applied for denoising speech signal of input. The five feature extraction methods deployed were discrete wavelet transform (DWT), peak values, tri-spectral features, Mel frequency cepstral coefficient (MFCC), as well as the varied value between standard and input signals. Later, scaling of features involved the linear identical scaling (LIS) technique with similar scaling factors for all feature sets at training and testing stages. Hence, an ISVM was built with the best fitness validation for the training stage. The two stages of the ISVM are: (i) linear dual classifier (identify similar and varied class features concurrently) and (ii) cross fitness validation (CFV) (hinder over-fitting issue). As a result, the approach yielded 98.2% accuracy rate.

III. PROPOSED METHODOLOGY

This section discusses the methodological steps followed in this article to achieve the previously mentioned research objectives. in details, the proposed methodology consists of seven main stages, including: (i) Stage 1: dataset gathering, (ii) Stage 2: dataset preprocessing, (iii) Stage 3: read the audio files and their labels, (iv) stage 4: convert waveforms to spectrograms, (v) stage 5: convert labels to integer IDs, (vi) stage 6: build and train the model, and (vii) stage 7: evaluate the model (testing). Figure 1 depicts the methodological steps followed in this article.

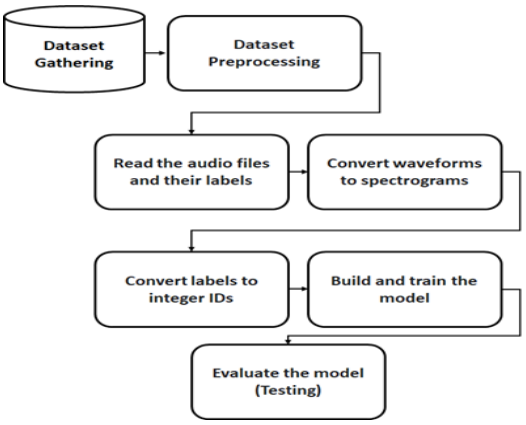


Figure 1: Research Methodology

Stage 1: Dataset gathering:

Speech accent dataset is used in this research to train and test the models. This dataset is created to uniformly offer a large dataset of speech accents from different language backgrounds. All participants, native and non-native read the same paragraph and their speech are neatly recorded.

The speech accent dataset allows you to assess the linguistic backgrounds of every speaker to evaluate which features are main predictors of every accent. This dataset pretends that accents are systematic speech rather than solely wrong speech.

Speech accent dataset has (2140) speech records, every speech is recorded from a different speaker reading the same reading paragraph. Speakers are originated from (177) different countries with more than (214) different languages. Each speaker is speaking the same phrase. However, the speech accent dataset includes the followings:

(i) Test read file (reading-passage.txt), which is the text all speakers read.

(ii) Demographic information (speakers_all.csv), which holds the demographic features on each speaker

(iii) Voice files (recording.zip), which contains .mp3 files that denotes each speech for each speaker.

In addition, to ensure robustness in training the models, another Arabic dataset is used, which is called Arabic Speech Corpus. This speech corpus was recorded in specialized studio in south Levantine Arabic. It is characterized by its high quality and natural voice.

Those two datasets (i.e., corpuses) are automatically aligned together to be used a large corpus for training speech recognition models efficiently.

Stage 2: Dataset Preprocessing:

To be in line with the main goal of this article, the record for Arabic speaker is extracted to be labeled with (1 denotes Arabic speaker) and all other voices are labeled with (1 denotes non-Arabic speaker). Whilst every WAV file consists of time-series data along with several samples per time unit (i.e., second).

Further, every sample exemplifies the amplitude of the voice signal at that particular point of time.

As the value of amplitude ranging for each WAV voice file ranging from -32,768 to +32,767. In addition, the shape of the In this stage, decoded tensors are created for each waveform and the related labels. As:

•Every WAV file consists of time-series data with number of samples per time unit (herein, seconds).

•Every voice sample denotes the amplitude of the voice signal at timeframe.

•The value of amplitude is ranging from (-32,768 to +32,767) in a 16-bit system.

•The average sample rate is 16kHz for all recorded speeches.

The form of the tensor viewed by using the following class:

where value of channels is to 1 for mono or 2 for stereo.

Then, defining the function to preprocess the raw WAV audio files into audio tensors:

After that, a function to create labels using their root directories for every file:

Once labels are created, helper function is used to group them all together as follows:

Herein, the input is the WAV filename, and the output is a tuple consisting of the audio (i.e., speech voice) and label tensors that can be used in supervised learning.

The training set and testing set are built using the following script in order to extract the audio-label pairs, respectively:

Stage 4: Convert waveforms to spectrograms

In this stage, the waveforms of all voices included in the dataset are presented in the time-series domain. After that, transforming the signal of waveforms with respect to time-series domain into signals of the time-frequency domain through the use of Fourier transform, which converts all waveforms into spectrograms respectively, this the change of frequency is represented with respect to time and also can be presented as a 2-dimensional image. At the end, the spectrogram images are used to train the model.

IV. Experimental Results and Findings

A. Evaluation Metrics

To measure the effectiveness of the Dynamic Time Warping (DTW), (ii) Gaussian mixture model (GMM), and (iii) Support Vector Machine (SVM), the confusion matrix is used to derive the following metrics:

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 1: Confusion matrix

$$\text{Accuracy}=(TPs+TNs)/(TPs+FPs+FNs+TNs) \quad \text{Eq. (1)}$$

$$\text{Recall}= (TP)/(TP+FN) \quad \text{Eq. (2)}$$

$$\text{Precision}=(TP)/(TP+FP) \quad \text{Eq. (3)}$$

$$\text{F1-measure}=(2*\text{precision}*\text{recall})/(\text{precision}+\text{recall}) \quad \text{Eq. (4)}$$

Where TPs denotes true positives that indicates the number of samples that correctly precited as Arabic speeches, FPs denotes false positives that indicates the number of speeches that wrongly predicated as Arabic speeches, FNs denotes false negatives that indicates the number of speeches that wrongly predicated as English speeches, and TNs denotes true negatives that indicates the number of speeches that correctly predicted as English speeches.

voice tensor fetched by using

Stage 3: Read the audio files and their labels

A utility function is used to convert waveforms to spectrograms according to the following conditions:

•The waveforms have to be with the same length; therefore, they'll have the same dimensions when converting them to spectrograms. This process can be achieved by using (using tf.zeros)., which is zero-padding technique, to deal the voice records that are shorter than 1 second.

•Call tf.signal.stft to conduct Fourier transform and set the length of frame and its parameters.

•Then, an array with complicated numbers is produced by Fourier transform, this array represents magnitude.

The following code is used to display the spectrogram.

Stage 5: Convert labels to integer IDs

Herein, the following script is used to transform the waveform of all files into spectrograms as well as converting the labels to integer IDs:

Stage 6: Build and train the model

To extract the features the following libraires is used:

```
import os
import pandas as pd
import numpy as np
import librosa
```

then from pyAudioAnalysis import ShortTermFeatures.

Stage 7: Evaluate the model (Testing)

To measure the effectiveness of each model, all models were trained on the same dataset, and tested using the same testing dataset to ensure fairness between them.

The confusion matrix (refer to Figure 5.1) is used to generate the evaluation metrics and to determine how well each model did recognize each voice the test dataset.

B. Experimental Results

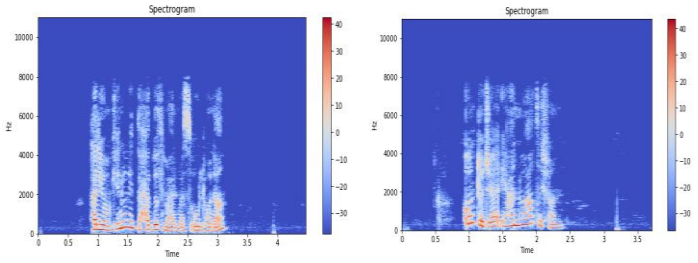
The used dataset contains mp3 files with Arabic and English speeches. As a preparation stage, all Arabic voice files and renamed into (Arabic no) since no is a sequential number started from 1, and the same thing is applied into English files with name (English no). After that, each Arabic voice file is labeled with 1 (Arabic class), and all English voice files are labelled with 0 (English class). Table 1 presents sample of dataset after preparation stage.

Table 1: Sample of Dataset

File name	Sentence	Label
Arabic 1	هل يمكني التحدث مع المسؤول هنا؟	1
Arabic 2	كنت جائعا وعطشا	1
English1	Play something by Louisiana Blues	0
English2	Find the schedule for Grand Canyon Trail.	0

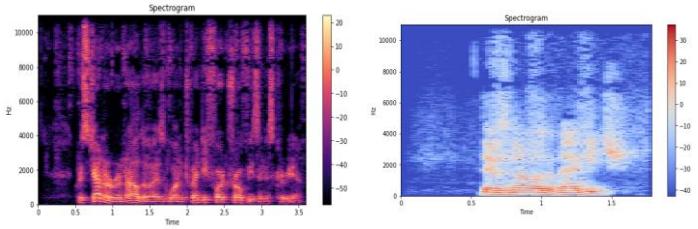
Then, as known voice data analysis is about analyzing and recognizing voice signals acquired through digital voice devices, with several applications in the enterprises. At first, before features from voice files. A spectrogram for each voice is presented to visually represent the signal strength with respect to the time at different frequencies present in a specific waveform. As shown in Figure 3and 4, which depict the spectrogram for Arabic and English samples, as represents as a heat map, which means image with the visual intensities presented by different colors or brightness levels.

The spectrogram is extracted using "librosa.display.specshow" in Python, and ".stft()" is used to transform the data into Fourier transform. Thus, the amplitude of each frequency at every point of time is known.



a- Spectrogram for Arabic 1 b- Spectrogram for Arabic 1

Figure 3: Spectrogram for Arabic Samples



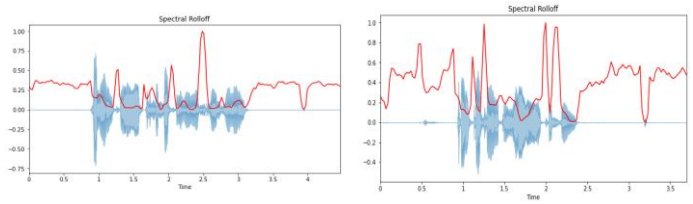
a- Spectrogram for English 1 b- Spectrogram for English 1

Figure 4: Spectrogram for English Speech

Each voice signal contains several features. However, to extract the features that are related to the problem, in needed, the spectral features are generated by transforming the time-based signal into the frequency domain using the Fourier Transform. Spectral roll-off, as shown in Figure 5 - 8, is used a measure of the shape of each voice signal. It indicted the degree of the frequency of signals at which high frequencies descend to zero.

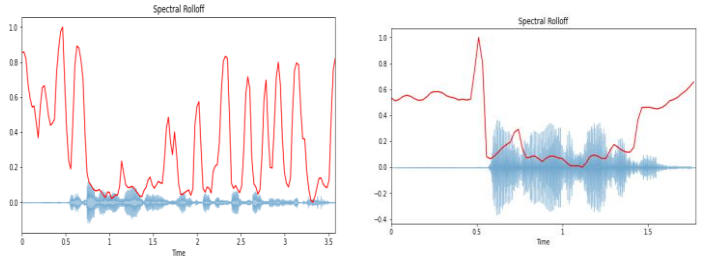
To get it, the fraction of bins is calculated in the power spectrum such that the 85% of its power is at lower signal frequencies, by suing "librosa.feature.spectral_rolloff" function.

getting begin with the voice data analysis and extract the significant



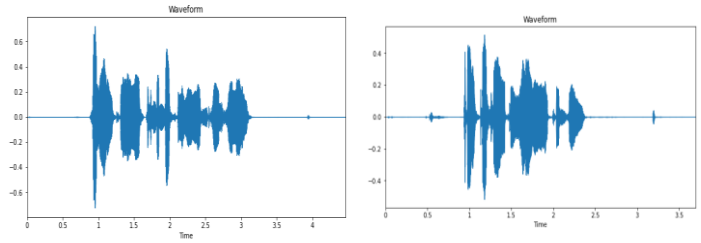
a- Spectral Rolloff for Arabic 1 b- Spectral Rolloff for Arabic 2

Figure 5: Spectral Rolloff for Arabic Speech



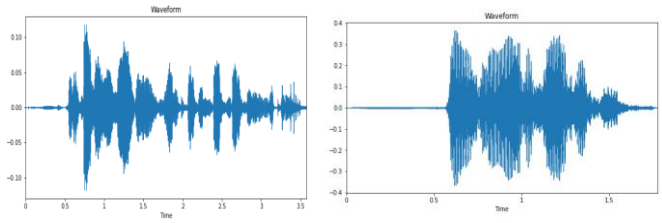
a- Spectral Rolloff for Arabic 1 b- Spectral Rolloff for Arabic 2

Figure 6: Spectral Rolloff for English Speech



a- Waveform Rolloff for Arabic 1 b- Waveform Rolloff for Arabic 2

Figure 7: Waveform Rolloff for Arabic Speech



a- Waveform Rolloff for English 1 b- Waveform Rolloff for English 2

Figure 8: Waveform Rolloff for English Speech

However, using testing dataset that consists of 7016 English voice files and 7016 Arabic voice files, the three algorithms is tested to demonstrate their performance. And the experimental results, as shown below in Table 2, and Figure 9, reveal that the SVM achieves 81.7% accuracy rate, 81% precision rate, 82.2% recall rate, and 81.6% f-measure. Whilst GMM outperforms only the SVM as it achieves 91.7% accuracy rate, 91.7% precision rate, 91.5% recall rate, and 91.6% f-measure. On the other hand, DTW outperforms SVM and GMM as it achieves 95.7%

accuracy rate, 96% precision rate, 95.4% recall rate, and 95.7% f-measure.

Table 2: Experimental Results

	Accuracy	Precision	Recall	F-measure
SVM	0.817	0.810757	0.822222	0.816449
DTW	0.957	0.96	0.954274	0.957129
GMM	0.917	0.917505	0.915663	0.916583

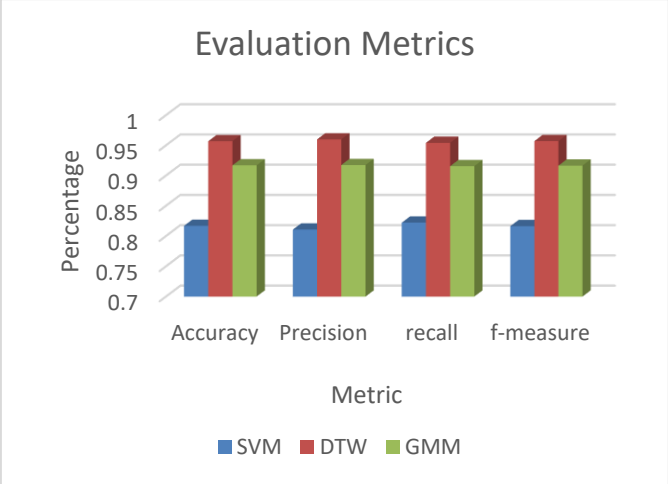


Figure 4.9: Experimental Results

V. Conclusion and Future Work

Despite the successful contributions of this study that resulting from the previously defined objectives to compare between three well-known Arabic recognition algorithms, and the results shows that the DTW outperforms the GMM and SVM in terms of accuracy, precision, recall and f-measure, as it achieves 95.7%, 96%, and 95%, and 96%, respectively. However, there is still a huge margin to improve the work this field. The followings are brief recommendations that can be explored and studied for future research directions in automatic Arabic speech recognition, such as studying the semantic relationships in Arabic voices, by utilizing knowledge base mechanisms such as WordNet, to improve ASR performance; hybridizing different models of machine learning and deep learning to provide a superior performance in Arabic language; automatic extracting the Arabic phonemes using data-driven models and clustering techniques; and finally, investigating the influence of variations of word pronunciation in ASR field in Arabic language.

References

[1] Wayman, J., Jain, A., Maltoni, D., & Maio, D, "An introduction to biometric authentication systems,", *In Biometric Systems.* Springer, London, 2005, pp. 1-20.

[2] Mahfouz, A., Mahmoud, T. M., & Eldin, A. S, "A survey on behavioral biometric authentication on smartphones," *Journal of information security and applications*, 37, 2017, pp. 28-37.

[3] Peralta, D., Galar, M., Triguero, I., Paternain, D., García, S., Barrenechea, E., & Herrera, F., "A survey on fingerprint minutiae-based local matching for verification and identification: Taxonomy and experimental evaluation," *Information Sciences*, 315, 2015, pp. 67-87.

[4] Alsaadi, I. M, "Physiological biometric authentication systems, advantages, disadvantages and future development: A review," *International Journal of Scientific & Technology Research*, 4(12), 2015, pp. 285-289.

[5] Tolba, A. S., El-Baz, A. H., & El-Harby, A, "Face recognition: A literature review," *International Journal of Signal Processing*, 2(2), 2006, pp. 88-103.

[6] Hamidi, M., Satori, H., Laaidi, N., & Satori, K,"Conception of speaker recognition methods: A review," *In 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 2020 pp. 1-6. IEEE.

[7] Hourri, S., Nikolov, N. S., & Kharroubi, J, "Convolutional neural network vectors for speaker recognition," *International Journal of Speech Technology*, 24(2), 2021, pp. 389-400.

[8] Hourri, S., Nikolov, N. S., & Kharroubi, J, "Convolutional neural network vectors for speaker recognition, " *International Journal of Speech Technology*, 24(2), 2021, pp. 389-400.

[9] Carroll, C, "*Curating Curious Collections: An Interdisciplinary Perspective*, " Predatory Pub Quarterly, 16, 1996, pp.3-134.

[10] Ismail, A., Abdlerazek, S., & El-Henawy, I, "Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping," *Sustainability*, 12(6), 2021, pp. 2403.

[11] Alonso, J. B., Cabrera, J., Medina, M., & Travieso, C. M, "New approach in quantification of emotional intensity from the speech signal: emotional temperature," *Expert Systems with Applications*, 42(24), 2015, pp. 9554-9564.

[12] Luengo, I., Navas, E., & Hernáez, I, "Feature analysis and evaluation for automatic emotion identification in speech," IEEE Transactions on Multimedia, 12(6), 2010, pp. 490-501.

[13] Cao, H., Verma, R., & Nenkova, A, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer speech & language*, 29(1), 2015, pp.186-202.

[14] Wang, K., An, N., Li, B. N., Zhang, Y., & Li, L, "Speech emotion recognition using Fourier parameters," *IEEE Transactions on affective computing*, 6(1), 2015, pp. 69-75.

[15] Cheng, X., & Duan, Q, "Speech emotion recognition using gaussian mixture model," *In The 2nd international conference on computer application and system modeling*, 2012, pp. 1222-1225.

[16] El Ayadi, M. M., Kamel, M. S., & Karray, F, "Speech emotion recognition using Gaussian mixture vector autoregressive models," *In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 4, 2017, pp. IV-957, IEEE.

[17] Tashev, I. J., Wang, Z. Q., & Godin, K, "Speech emotion recognition based on gaussian mixture models and deep neural networks," *In 2017 information theory and applications workshop (ITA)*, 2017, pp. 1-4, IEEE.

[18] Kanisha, B., Lokesh, S., Kumar, P. M., Parthasarathy, P., & Babu, G, "Speech recognition with improved support vector machine using dual classifiers and cross fitness validation," *Personal and Ubiquitous Computing*, 22(5), 2018, pp. 1083-1091.

[19] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., & Zhu, Z. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In International conference on machine learning (pp. 173-182). PMLR.

[20] Graves, A., & Jaitly, N, "Towards end-to-end speech recognition with recurrent neural networks," *In International conference on machine learning*, 2014, pp. 1764-1772. PMLR.

[21] Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., & Zweig, G, "*Achieving human parity in conversational speech recognition*," arXiv preprint arXiv:1610.05256, 2016.

[22] Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., & Hall, P, "*English conversational telephone speech recognition by humans and machines*," arXiv preprint arXiv:1703.02136, 2017.

[23] Michalowski, P.,"*The lives of the Sumerian language*," In: Margins of Writing, Origins of Cultures. 2006, pp. 159–184.

Authors

First Author – Ali Alanazy, Master candidate student of Information Security, Department of Information Technology, Tabuk University, KSA. and email address unf.t.ali@gmail.com

Second Author – Mohammed Alatawi, and email address Alatawimn@ut.edu.sa.

Correspondence Author – Mohammed Alatawi, email address, Alatawimn@ut.edu.sa