Al-Rafidain Journal of Computer Sciences and Mathematics (RJCM)

www.csmj.mosuljournals.com

Unsupervised and Semi-Supervised Speech Recognition System: A Review

Mazin Ali Fadhel 1, Yusra Faisal Mohammed 2

Integrity Commission, Nineveh Investigation Directorate, Mosul, Iraq [1]
Department ofComputer Science, College of computer science and mathematics, Mosul University, Mosul, Iraq [2]
**Corresponding author. Email: mazin.20csp74@student.uomosul.edu.iq* [1,]

Article information

Article history:
Received: 1/9/2022
Accepted: 20/11/2022
Available online:

Abstract

Voice is a behavioral biometric that may reveal a person's age, gender, ethnicity, and emotional state. Speaker recognition is the method of identifying individuals through their sounds. Despite the fact that over the last eight decades, academics have already been focusing on speaker identification, technological advancements like the Internet of Things (IoT), smart homes, voice assistants, smart gadgets and humanoids have made their use popular in modern society. This study offers a thorough analysis of the speaker identification literature. It looks at recent developments as well as problems in this area of study. This study looks into feature extraction, classifiers, and the structure of the speaker recognition system. Also covered is how speaker recognition is used in apps. The objective is to increase researchers' understanding of speaker identification by machine learning since recent research has shown that it is easy to deceive machine learning into producing an incorrect prediction.

Keywords:
Gaussian mixture models (GMM), Automatic speech recognition (ASR), Hidden Markov models (HMM), Machine learning (ML), Multiple layer perceptron (MLP), Radial basis probabilistic neural network (RBPNN), Support vector machine (SVM), Learning vector quantization (LVQ), Mel frequency cepstral coefficients (MFCCs), Dynamic Time Wrapping (DTW), Mixture Invariant Training (MixIT).

Correspondence:
Author: Mazin Ali Fadhel
Email: mazin.20csp74@student.uomosul.edu.iq

1. INTRODUCTION

The twentieth century has seen a fast increase in scientific study as well as suitable advances in the field of documentation, science, research, and development facts. This, in turn, provides a thorough grasp of the typical testing methodologies employed. Voice recognition technology is becoming one of the most popular areas of scientific research. Speech recognition is a method of analyzing the contents of the speaker's speech, and each speech recognition system employs several algorithms to convert sound waves into useful data that the system interprets and processes, and then this system produces output in the form of text to be used in accordance with the requirements. [1].

The increased concern in security has resulted in an increase in the usage of biometrics, other distinguishing characteristics, in addition to the face, are the retina, voice, and iris. As seen in Fig. 1, biometrics are categorized into two types: physiological and behavioral [1,2].

Unlike the latter, the former includes the voice, keystroke, signature, face, fingerprint, and iris. A voice is any sound people use to express their thoughts, ideas, opinions, etc. However, voices are clearly defined as any sound generated by vocal folds vibrating while air is pushed through the lungs. [3].

Voice is the most natural mode of communication. The speaker's race, age, gender, and emotions are revealed. Speaker recognition research has greatly improved over the last 80 years because of advances in hardware, design, algorithms, and technology. [4].

The field of digital signal processing known as automatic speech recognition (ASR) focuses on speech recognition. Since many decades ago, automatic speech recognition has been a significant and active research area. Multiple speech recognition is one possible implementation of ASR, in which case we attempt to extract a speaker's voice from a disorganized speech stream. Blind source separation, or BSS, is the initial step in learning about many amplifiers. BSS is eliminating the sources from the mixture without being aware of the sources beforehand. DUET and Independent Component Analysis (ICA) are two BSS approaches that have both benefits and drawbacks. Following BSS, characteristics such as strength, pitch, vocal tract composition, modality, etc. are retrieved from speech signals [5]. Following the features' extraction, a number of methods for speech modeling and classification have been created and put to use. HMM, GMM, SVM, and vector quantization are a few of the techniques. All of these techniques are utilized to develop a particular model of the speaker via training [6,7]. The goal of this study is to offer pertinent research that may open the door to unattended and semi-supervised ASR. Training may be supervised or unsupervised, or it may be both.

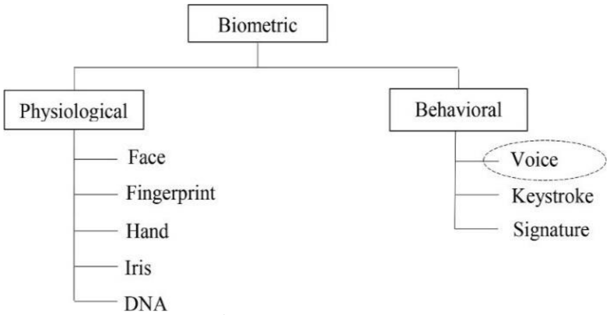


Figure 1. Types of biometrics: Physiological and Behavioural.

2. Summary of Previous Studies

(In 2003) A technique to combine active and unsupervised learning for automated speech recognition is described by scholars Giuseppe Riccardi and Dilek Hakkani-T ür in their work "active and unsupervised learning for automatic voice recognition" (ASR). In order to optimize performance using transcribed and untranscribed data, it is important to reduce the amount of human supervision for acoustic training and language models. The results of the trials demonstrate a 75% reduction in the quantity of labeled data needed for a particular level of word accuracy when active and unsupervised learning are combined [2].

(In 2010) In their work, "Prosodic Feature Based Text Dependent Speaker Recognition Using Machine Learning Algorithms," researchers Sunil Agrawal, Shruti A.K., and C. Rama Krishna made their findings available. It is suggested to use prosodic features-based text-dependent speaker identification, where prosodic characteristics are obtained by linear predictive coding. For characterizing a speaker's voice, formants are useful criteria. The formants' related amplitudes, fundamental frequencies, speech utterance lengths, and windowed section energies are merged. Machine learning (ML) algorithms are given this feature vector for recognition. Four machine learning (ML) algorithms—MLP, RBFN, C4.5 decision tree, and BayesNet—have been compared for performance. All of these ML algorithms operate similarly to the C4.5 decision tree. MLP performs better when identifying gender, and experimental findings indicate that RBFN improves with population size [12].

(In 2011) Moaz Abdulfattah Ahmad and Rasheed M. ElAwady submitted their article, "Phonetic Recognition of Arabic Alphabet letters using Neural Networks." an approach for utilizing artificial neural networks to recognize Arabic letters spoken by any speaker Understanding Arabic speech at this level is crucial (continuous words). To recognize the letters of the Arabic alphabet used by independent speakers, a recommended identification technique has been developed. This technique based on a phonetical analysis of individual Arabic script characters. To generate recognizable binary codes for each letter, the Principal Component Analysis (PCA) approach uses multilayer perceptron and feed-forward back propagation neural networks. It obtains the primary elements of an audio signal. A big dataset was used to get a detection accuracy of over 96 percent [4].

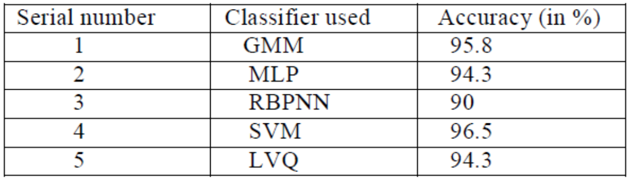
(In 2013) Deep Neural Networks (DNNs) can be trained in low-resource settings to serve as data-driven feature front-ends for continuous speech recognition with a large vocabulary, according to a new method proposed by researchers Samuel Thomas1, Michael L. Seltzer2, Kenneth Church3, and Hynek Hermansky in their study "Deep Neural Network Features and Semi-supervised Training for Low Resource Speech Recognition" (LVCSR). We combine transcribed multilingual data and semi-supervised training to build the suggested feature front-ends in order to get around the dearth of training material for acoustic modeling in these situations. In a low-resource LVCSR scenario with just an hour's worth of in-domain training data, the trials demonstrate that the suggested features provide an absolute improvement of 16%. Three-quarters of these gains are accounted for by DNN-based features, while the remaining one-fourth is accounted for by semi-supervised training [1].

(In 2014) Amer M. Elkourd, a researcher, created an original Arabic isolated word speaker-dependent identification system using feature extraction and classification techniques. The system is created using a laptop with a G62 Core I3/2.26 GHz CPU and Matlab. In a quiet room with 5 different speakers, 40 Arabic words were captured using a laptop microphone. Each sentence will be repeated eight times by each speaker. Five of them are utilized in training, while the other four are employed in the testing phase. First identifying the beginning and end of each syllable and removing pauses using an endpoint identification approach based on energy and zero crossing rates, then using a discrete wavelet transform to eliminate noise from the data. The system recognizes the speaker before merely loading the user's reference model to speed up the system and reduce execution time. There was five different techniques: Gaussian Mixture Model (GMM) with MFCC, pairwise Euclidean distance with MFCC, Dynamic Time Warping (DTW) with Formants features, MFCC+DTW, and Itakura distance with LPCF (LPC). Was recognized by 57 percent, 87.9 percent, 90.9 percent, 83.23 percent, and 85.23 percent of people, respectively. To improve the system's accuracy, the experimented with various pairings of these 5 techniques. The best combination is discovered to be MFCC | Euclidean + Formant | DTW + MFCC | DTW + LPC | Itakura, which has a high accuracy of

94.39 percent but a lengthy calculation time of 2.9 seconds. the investigation several sub combinations of this hybrid in order to shorten calculation time and discovers that the first combination, MFCC | Euclidean + LPC | Itakura, provides the greatest performance. The system reduces average calculation time to 1.56 seconds and boosts system accuracy to 94.56 percent by only combining Formant | DTW + MFCC | DTW techniques when the two methods do not concur [19].

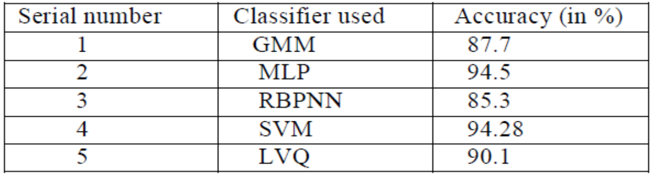
(In 2015) The researchers Amber Singh and R.S. Anand examined the Utilizing five models; speech recognition accuracy of test patterns is unknown in their study "Speech recognition using supervised and unsupervised learning techniques." As the number of classes for categorizing GMM, MLP, SVM, LVQ, and RBPNN are utilized, and their ability to recognize speech on five isolated digits is examined. Unknown test patterns are increased from three to five. Three isolated words with values between 0 and 2 are utilized for training and testing in the first experiment. The second experiment uses five solitary words with numbers ranging from 0 to 4. Using unidentified test patterns, the classification accuracy of each classifier is ultimately determined, and conclusions are drawn. The SVM classifier was determined to have the best accuracy in the first experiment, whereas the RBPNN classifier had the lowest accuracy. 96.5 percent and 90 percent of the accuracy, respectively. In the second trial, the MLP classifier was determined to have the best accuracy while the RBPNN classifier had the lowest accuracy. [15].

Table 1. Speech recognition accuracy for experiment one



Serial number	Classifier used	Accuracy (in %)
1	GMM	95.8
2	MLP	94.3
3	RBPNN	90
4	SVM	96.5
5	LVQ	94.3

Table 2. Speech recognition accuracy for experiment two



Serial number	Classifier used	Accuracy (in %)
1	GMM	87.7
2	MLP	94.5
3	RBPNN	85.3
4	SVM	94.28
5	LVQ	90.1

(In 2016) saw the publication of research results by academics M. K. Ahirwal, P. Lodha, N. D. Londhe, an IEEE Senior Member, and others in the journal "Machine Learning Paradigms for Speech Recognition of an Indian Dialect." Artificial Neural Network (ANN) and Support Vector Machine, two fundamental machine learning methods, provide the foundation on which ASR generates concepts for the unique and strategically significant Indian dialect "Chhattisgarhi" (SVM). An SVM and traditional feed-forward ANN were used to analyze a dataset containing 50 unique words from 15 speakers. While ANN surpasses HMM at speaker independent speech synthesis, SVM performs better than the other two classifiers. For terms with ten, twenty, and fifty letters, the accuracy was 90.60, 87.27, and 82.49 percent, respectively [3].

(In 2017) In their paper "Speaker Independent Arabic Speech Identification Using Support Vector Machine," the researchers Shady Y. EL-Mashed, Mohammed I. Sharway, and Hala H. Zayed discussed issues related to the recognition of speaker independent Arabic speech using SVM. The recommended model is used to generate the connected Arabic digits using neural networks as an example (number). Furthermore, the technology may be used in any other area. This was done through:

- initially building a corpus of 1000 numbers made up of 10,000 digits recorded in a noisy environment by 20 speakers with a range of characteristics, including gender, age, physical condition, and so on.
- The second is that each recorded number has been divided into 10 separate digits. The characteristics of these digits were then extracted using the Mel Frequency Cepstral Coefficients (MFCC) method and input into neural networks for recognition. When we used the Support Vector Machine (SVM), the system's performance was over 94 percent [13].

(In 2018) The research paper entitled Speaker Recognition Using Deep Belief Networks to CCIS Proceedings was given by Adrish Banerjee, Akash Dubey, Abhishek Menon, Shubham Nanda, and Gora Chand Nandi. When attempting to identify speakers, make use of the short-term spectral qualities that you have picked up from the DBN and improved using MFCC features. On the ELSDSR dataset, it was able to detect objects with an accuracy of 0.95 utilizing these characteristics, as compared to 0.90 when using single MFCC features. This was due to the fact that these characteristics included a more comprehensive collection of information. [14].

(In 2019) In their study "End-to-End Speech Recognition Sequence Training with Reinforcement Learning," the researchers Andros Tjandra, Sakriani Sakti, (member, ieee), and Satoshi Nakamura presented a policy gradient reinforcement learning method as a potential approach for optimizing the end-to-end ASR model. The model that was used to train the recommended approach has the following benefits:

1. Using a free sampling technique and its own sample as input, the model repeats the inference phase.
2. make the model more effective by introducing a reward function linked to the ASR evaluation metric (e.g., negative Levenshtein distance).

Their experiment's findings show that their suggested approach considerably enhances model performance when compared to a model trained just using instructor forcing and the maximum likelihood objective function [9].

(In 2019) The researchers Dongwei Jiang, Xiaoning

Lei, Wubo Li, Ne Luo, Yuxuan Hu, Wei Zou, and Xiangang Li presented an unsupervised pre-training method known as Masked Predictive Coding in their study "Improving Transformer-Based Speech Recognition Using Unsupervised Pre-Training" to address the issue of rising popularity of Speech recognition technologies in a variety of industrial applications. It is also expensive to gather the quantity of transcribed data needed to construct a decent voice recognition system. Research at HKUST demonstrates that they can surpass the best end-to-end model by more than 0.2 percent absolute CER and obtain CER 23.3 percent using the same training data. With additional pretraining data, they can lower the CER to 21.0 percent, which is a decrease of 11.8 percent over the baseline [8].

(In 2019) In their study, "Pre-Training In Deep Reinforcement Learning For Automatic Speech Recognition," researchers Thejan Rajapakshe, Rajib Rana, Siddique Latif, Sara Khalifa, and Bjorn W. Schuller explore how deep RL pre-training can be used to reduce training time and improve performance in speech recognition, a common HCI application. Get much better results in less time on a publicly accessible dataset for spoken command recognition [11].

(In 2019) Researchers Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli investigated unsupervised pre-training for speech recognition using raw audio representations in their paper "Wav2vec: Unsupervised Pre-Training For Speech Recognition." A noisy contrastive binary classification task is used as the training set for a straightforward multi-layer convolutional neural network. When only a few hours of transcribed data are available, the WSJ experiments reduce the WER of a powerful character-based log-mel filterbank by up to 36%. According to the nov92 test set, their technique achieves a WER of 2.43% [18].

(In 2020) "Hybrid Features Extraction and Machine Learning Based Arabic Speaker Classification," a paper by Saeed Mian Qaisar and M. Akbar, was published. A hybrid model-based method for Arabic speaker recognition is suggested. The objective is to locate a very precise and efficient escape route. It may be done by skillfully combining powerful classification techniques with hybrid features extraction techniques. The Mel-Frequency Cepstral Coefficients (MFCC) and Perceptive Linear Prediction Coding Coefficients (PLPCC) are extracted from the improved Arabic speech (MFCCs). The speaker is then identified using the k- Nearest Neighbor (KNN) classifier. The categorization of Arabic speakers by the approach is 90.8 percent accurate [10].

(In 2020) In their work, "Robust Hybrid Features Based Text Independent Speaker Identification System Over Noisy Additive Channel," researchers Drs. Hesham Adnan Alabbasi, Fadhel Sahib Hasan, and Ali Muayad Jalil reported their findings. Gammatone Frequency Cepstral Coefficients (GFCC) and Power Normalized Cepstral Coefficients (PNCC) are two strong characteristics that work together to strengthen the speaker identification system's resistance to different forms of noise. The Universal Background Model Gaussian Mixture Model (UBM-GMM) is used as a feature matching and a classifier to determine claim speakers. The assessment results show that the proposed hybrid feature outperforms conventional features in a variety of noise conditions and signal-to-noise ratios [20].

(In 2021) The "Unsupervised Speech Recognition" paper's authors, Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli, provided further information. Wav2vec is sometimes known as wav2vec-U. Unsupervised learning is the process of developing speech recognition algorithms without labeled input. Self-supervised speech representations are employed in adversarial training to segment unlabeled audio, and phoneme mapping is discovered from these representations. The success of their technique depends on the use of suitable representations. Wav2vec-U decreases the phoneme error rate on the TIMIT benchmark from 26.1 to 11.3 percent when compared to the current best unsupervised work. Wav2vec-U has a test-other word error rate of 5.9 percent according to the more comprehensive English Librispeech benchmark. Despite having just 960 hours of labeled training data 2 years ago, the system can now compete with some of the best systems that have been recorded. Test nine more languages, including those with limited resources like Kyrgyz, Swahili, and Tatar [17].

(In 2022) The authors of the study entitled " Unsupervised Speech Enhancement with Speech Recognition Embedding and Disentanglement Losses" are Viet Anh Trinh and Sebastian Braun, they presented a proposal in their research to solve two issues, propose an unsupervised loss function. First, when combining clean speech and noisy corpora to create synthetic datasets, domain mismatches occur. Second, ASR performance is harmed by speech augmentation. The function is created by combining speech recognition embedding and disentanglement loss with the MixIT loss function. The findings reveal that the proposed function significantly outperforms a baseline built in a supervised manner on the noisy Vox-Celeb dataset in terms of speech enhancement performance. Full unsupervised training cannot outperform the baseline, but when supervised and unsupervised training are combined, the system can achieve the same speech quality and ASR performance from the best supervised baseline [35].

3. Background

3.1. Automatic Speech Recognition (ASR):

A technique called Automatic Speech Recognition (ASR) is used to translate what people say into text. It may be used to carry out important activities including command recognition, dictation, translation, and security control (verify the identity of the person to allow access to services such as banking by telephone) [24]. ASR might help people with disabilities engage with society since it allows writing on software programs easier and faster than using a keyboard. Additionally, it might be used to wirelessly turn on and off household lights and appliances. There are two categories of ASR. Discrete word and continuous speech recognition systems are the two main categories of speech recognition systems, with speaker dependent and speaker independent subcategories within each category [25]. In automatic speech recognition systems, there are two phases [19]:

- A training stage in which the computer system picks up on the reference patterns that stand in for the different speech sounds.
- The stage at which an unknown speech signal is identified using stored reference patterns.

A speech recognition system's block diagram is shown in Figure (2). It includes:

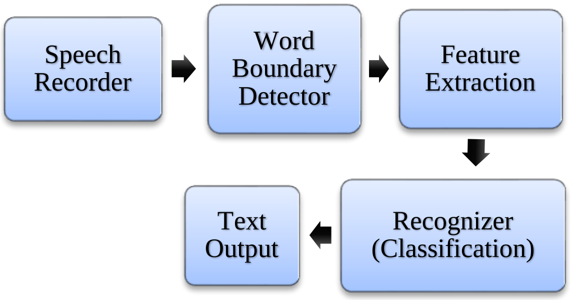


Figure 2 Speech recognition system block diagram

Upon the detection of an input signal, it is converted into an acoustic vector of fixed size. For feature extraction, speech input must first undergo pre-processing. In this component, a variety of standard operations, including pre-emphasis, noise reduction, framing, endpoint identification, and normalization, may be used [21]. A collection of data is therefore taken from the pre-processed signal using the feature extraction component. To discriminate across classes, extracted characteristics should be able to. There are several other feature extraction methods, such as Linear Predictive Coding (LPC), MFCC, and DWT [22]. Utilizing these deduced properties, the classification component efficiently classifies the incoming voice signals. The joint probability distributions over the supplied data may be allocated to the class labels during the classification stage. Currently, two methods are being used: the HMM and the Gaussian Mixture Model (GMM) [20]. Consider the data that is represented across the whole system, including the language model and audio model, to illustrate the data component first. The language model is an exact copy of the individual's fluid informational flow [24].

3.2. ASR Challenges:

One of the problems with ASR, even when it is supervised, is that natural speech sounds different depending on the speaker and the environment. Different speakers' words sound different, and it can be hard to tell these differences apart from the meaning. Even within a single speaker, there can be differences in speed, volume, affect, etc. Also, ASR models are often trained on clean speech data, but they are often tested on noisy speech data from real-time situations. Noise comes from things like background sounds and distorted signals from the input device.

Speaker-independent models could be trained with data from more than one speaker, but this usually makes ASR less accurate and requires more data for training to get good results. The same goes for background noises and other changes in the environment. Modern ASR systems, on the other hand, are speaker-adaptive. This means that they take into account the differences between speakers by using I-Vectors [26] and X-Vectors [27], which are low-dimensional vectors that represent speaker-specific features. Also, different augmentation techniques can be used to add more examples to the training data that reflect the expected differences in test conditions.

For example, volume and speed perturbation are used to show how different utterances are from each other. In the same way, noise-augmentation is used to add different environmental conditions to the training data [28]. All of these strategies are used together to train robust multi-condition ASR systems that can handle changes from many different sources.

3.3. Machine Learning:

The majority of people in today's society make use of machine learning (ML) and artificial intelligence (AI) without even giving it a second thought because ML and AI are so pervasive and valuable in today's society. Automatic Speech Recognition (ASR) software is one of the crucial areas in which these cutting-edge technologies have improved dramatically, almost to the point where they are on par with human capabilities. This makes ASR software one of the most important areas in which these advancements have occurred [30].

Recognition of intricate patterns in speech, handwriting, facial features, and other areas has become better over time. The creation of computer programs that enable computers to practice acquiring the aforementioned skills gave rise to the field of machine learning. Mitchell stated that a software program is considered to train from experiences E if its performance surpasses expectations, as measured by P, and improves with experience E when discussing a class of tasks in T and a performance metric P [29]. Some fundamental words with regard to machine learning include the following:

- A case study from the data.
- A feature is a collection of traits that acts as a vector or linear array's entry.
- The corresponding class or category of the item must be mentioned on labels.
- Training data are used to develop the machine learning algorithm during the learning phase.
- Test results: specifics that show how well the

instructional approach generalized.

Learning may be divided into four primary techniques based on how the computer obtains information to react appropriately, as shown in the following sections:

- **Supervised Learning:**

A labeled data collection with predetermined output classes or answers is used to train the computer in supervised learning. Theoretically, it should be feasible to develop a hypothesis that will work well on the test data if the training data is large enough. A simple illustration of supervised learning is the curve-fitting problem. A collection of input data is used to train the system to produce the curved surface that resembles the training data the most. The computer must interpolate the new data over the curved surface successfully when put to the test. Perceptron's, multilayer perceptron's, and limited MLPs are all part of this family of neural networks, which use either the delta learning rule or the perceptron learning rule [31].

- **Unsupervised Learning:**

The idea behind unsupervised learning is that the computer will automatically spot patterns in the input data that has not been labeled. The goal may be summarized as finding patterns in data set to group the train data into useful clusters or separate it into smaller subgroups. In this area, taxonomic difficulties are addressed, such as developing effective methods for classifying data into useful clusters. Kohonen networks (self-organizing maps), Hopfield network and Hebb (Hebbian learning), and networks' adaptive resonance theory (ART) are a few examples (competitive learning). An easy-to-use network that has been taught to generate the input is the automatic encoder, which utilizes the target output as input. The network is trained to reproduce the input using the unsupervised learning approach of gradient descent back propagation. An interconnected deep network is built using auto encoders. Pre-training employing unsupervised learning can be utilized to provide the deep network with improved starting weights and bias values [32].

- **Semi-supervised Learning:**

Labeled and unlabeled data are used in semi-supervised learning to train the system. Frequently, a large amount of unlabeled data is paired with a tiny percentage of labeled data. When obtaining labeled data is too expensive, this kind of learning approach is often utilized [33].

- **Reinforcement Learning:**

Compared to supervised learning, reinforcement learning involves a sequence of decisions and uses a smaller training set. The training set is the algorithm's dynamic environment. In order to get results via trial and error, this style of learning relies on the degree of incentives. The agent, which is the system or learner, the environment, which is what the agent interacts with, and action, which is the agent's reaction as a result of interacting, make up the three key elements of this technique of learning. Through feedback in the form of incentives and penalties, the agent or system learns. As it learns over a certain period of time, reinforcement learning will adopt behaviors that provide the highest rewards [34].

Table 3. A summary of previous work that are related to this study.

Reference (year)	***Feature extraction***	***Classification Method***	***Dataset***	***Accuracy***
[19] 2014	• (MFCC) (LPC)	Gaussian Mixture Model (GMM)	The dataset utilized in this system includes 40 Arabic words captured with a laptop microphone in a peaceful area with 5 distinct speakers. Each word will be recited 8 times by each speaker. Five of them are utilized in training, while the others are employed in testing.	• 85.23% | MFCC | • 57% | DTW + Formant | • 87% | GMM + MFCC | • 90% | DTW + MFCC | • 83% | Itakura | The optimal set-up has a calculation time of 2.9 seconds and a precision of 94.39 percent: MFCC | Euclidean + Formant | DTW + MFCC | DTW + LPC | Itakura.
[15] 2015	(MFCCs)	GMM, SVM, MLP, RBPNN, and LVQ	We employ three distinct speakers' isolated words from the first experiment. In the second experiment, five different people's solitary words are used.	SVM classifiers were shown to be the most accurate in the first experiment, whereas RBPNN classifiers were found to be the least accurate. 95.5% and 90%, respectively, of the accuracy. The MLP classifier's accuracy in the second experiment was found to be the greatest, while the RBPNN classifier's accuracy was found to be the lowest.
[3] 2016	(MFCCs)	(ANN), (SVM) and (HMM)	the dataset contains a maximum of 50 words with different speakers.	While HMM performs lower than other two classifiers, SVM and ANN perform better. has an accuracy of 90.60, 87.27, and 82.49 percent for word lengths of ten, twenty, and fifty, respectively.
[13] 2017	(MFCCs)	Support Vector Machine (SVM)	Arabic digits (number)	nearly 94%
[14] 2018	Deep generative models (DBNs) with (MFCCs)	Gaussian Mixture Model – Universal Background	ELSDSR dataset	achieved a recognition accuracy of 95% as compared to 90% when using standalone MFCC features

<table>
<tr><td></td><td></td><td>Model framework developed by Reynolds</td><td></td><td></td></tr>
<tr><td>[11]
2019</td><td>(MFCCs)</td><td>Combination model of CNN and LSTM RNNs</td><td>audio corpus of 105,829 utterances containing 30 command keywords spoken by 2,618 speakers.</td><td><table><tr><th>Classes</th><th>Improvement (%)</th></tr><tr><td>2</td><td>19.6</td></tr><tr><td>20</td><td>60.8</td></tr><tr><td>30</td><td>52.4</td></tr></table></td></tr>
<tr><td>[10]
2020</td><td>(PLPCC) and (MFCCs).</td><td>k-Nearest Neighbor (KNN) classifier</td><td>Arabic speaker categorization</td><td>90.8 percent</td></tr>
<tr><td>[35]
2022</td><td>Spectrogram</td><td>Unsupervised mixture invariant training (MixIT).</td><td>using 500 hours of LibriVox as a clean training dataset for the baseline supervised models. Use the VoxCeleb2 dataset as unsupervised training data for noisy speech.</td><td><table><tr><th>Method</th><th>Dataset</th><th>WER low SNR</th><th>WER high SNR</th><th>WER meeting</th></tr><tr><td>MixIT</td><td>Noisy</td><td>31.61</td><td>5.74</td><td>17.16</td></tr><tr><td>MixIT + (Emb)</td><td>Noisy</td><td>31.09</td><td>6.52</td><td>19.65</td></tr><tr><td>MixIT + (Dis)</td><td>Noisy</td><td>32.51</td><td>6.74</td><td>20.51</td></tr><tr><td>MixIT + Emb + Dis</td><td>Noisy</td><td>33.58</td><td>6.44</td><td>21.13</td></tr></table>Other experiences can be reviewed in the research</td></tr>
<tr><td>[9]
2019</td><td>Mel-Filterbank</td><td>SEQUENCE-TO-SEQUENCE ASR</td><td>Wall Street Journal dataset (WSJ)</td><td>Earned up to 6.10 percent CER</td></tr>
<tr><td>[18]
2019</td><td>log-mel filterbank</td><td>wav2letter++ toolkit</td><td>TIMIT</td><td>the WSJ experiments reduce the WER of a strong character-based log-mel filterbank baseline by up to 36%. On the nov92 test set, their method achieves 2.43 percent WER</td></tr>
<tr><td>[12]
2010</td><td>LPC analysis technique</td><td>MLP, RBFN, C4.5 and BayesNet</td><td>Each individual is represented by fifteen utterances of the numbers three, seven, and eight (five utterances for each digit).</td><td>Among these ML techniques, C4.5 decision tree performance is comparable. According to experimental findings, RBFN performs better as population size grows, while MLP performs better at gender identification.</td></tr>
<tr><td>[4]
2011</td><td>(PCA)</td><td>multilayer perceptron and feed-forward back propagation neural networks</td><td>Arabic Alphabet letters spoken by any speaker</td><td>96 percent</td></tr>
<tr><td>[20]
2020</td><td>• (PNCC)
(GFCC)</td><td>Universal Background Model Gaussian Mixture Model (UBM-GMM)</td><td>630 speakers are included in the (TIMIT) dataset, and each speaker has 10 utterances.</td><td>---</td></tr>
<tr><td>[2]
2003</td><td>----</td><td>combining active and unsupervised learning</td><td>two separate data sets:
- The first is derived from human-human talks and includes replies to the first request (8K utterances and 300K word tokens).
- The second is derived from human-machine dialogs (28K utterances and 318Kword tokens), and it comprises of users' replies to all system prompts.</td><td>75%</td></tr>
<tr><td>[1]
2013</td><td>----</td><td>DNN (Deep Neural Network)</td><td>multilingual data</td><td>With just one hour of training data, the absolute improvement in the LVCSR context was 16 percent.
Even though most of these gains are attributable to DNN-based features</td></tr>
<tr><td>[8]
2019</td><td>----</td><td>- Masked Predictive Coding</td><td>For pre-training, Mandarin datasets were gathered from Open SLR and the Linguistic Data Consortium (LDC).</td><td>Using the same training data, trials achieve a CER of 23.3 percent, which is higher than the absolute CER of the best final model. They can reduce the CER to 21.0 percent with more pretraining data, an 11.8 percent reduction from the baseline.</td></tr>
<tr><td>[17]
2021</td><td>---</td><td>wav2vec</td><td>TIMIT, Librispeech, MLS, ALFFA and CommonVoice corpora</td><td>on the TIMIT, reduce the phoneme error rate from 26.1 to 11.3. On the Librispeech scores a word error rate of 5.9 on test-other</td></tr>
</table>

4. Conclusion

By using clustering and classification methods, this review article quickly introduces the reader to speech recognition and machine learning (unsupervised and semi-supervised). Given how deeply buried one model is in the other, ASR and ML search techniques have been enhancing one another in recent years. This research reveals that the extraction of speech features is a common use of MFCC technology. The best techniques are HMM, GMM, and SVM. SVM beats the other two classifiers in every modeling methodology.

5. References

[1] IEEE Signal Processing Society, 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing: proceedings: May 26-31, 2013, Vancouver Convention Center, Vancouver, British Columbia, Canada.

[2] Riccardi, G., & Hakkani-Tür, D. Z. (2003). Active and unsupervised learning for automatic speech recognition. 8th European Conference on Speech Communication and Technology (Eurospeech 2003).

[3] Adhiparasakthi Engineering College. Department of Electronics and Communication Engineering, Institute of Electrical and Electronics Engineers. Madras Section, and Institute of Electrical and Electronics Engineers, ICCSP-2016 : 4th-6th April, 2016.

[4] M. A. Ahmad and R. M. el Awady, "Phonetic Recognition of Arabic Alphabet letters using Neural Networks," 2011.

[5] Yang W, Pilozzi A, Huang X. An Overview of ICA/BSS-Based Application to Alzheimer's Brain Signal Processing. Biomedicines. 2021; 9(4):386. https://doi.org/10.3390/biomedicines9040386.

[6] Sumita Nainan and Vaishali Kulkarni. 2021. Enhancement in speaker recognition for optimized speech features using GMM, SVM and 1-D CNN. Int. J. Speech Technol. 24, 4 (Dec 2021), 809–822. https://doi.org/10.1007/s10772-020-09771-2.

[7] Peng, Pingan & He, Zhengxiang & Wang, Liguan. (2019). Automatic Classification of Microseismic Signals Based on MFCC and GMM-HMM in Underground Mines. Shock and Vibration. 2019. 1-9. 10.1155/2019/5803184.

[8] Jiang, D., Lei, X., Li, W., Luo, N., Hu, Y., Zou, W., & Li, X. (2019). Improving Transformer-based Speech Recognition Using Unsupervised Pre-training. doi:10.48550/ARXIV.1910.09932.

[9] A. Tjandra, S. Sakti, and S. Nakamura "End-to-End Speech Recognition Sequence Training with Reinforcement Learning," IEEE Access, vol. 7, pp. 79758–79769, 2019, doi: 10.1109/ACCESS.2019.2922617.

[10] S. M. Qaisar and M. Akbar, "Hybrid features extraction and machine learning based arabic speaker classification," Oct. 2020. doi: 10.1109/ICCIS49240.2020.9257699.

[11] Rajapakshe, T., Rana, R., Latif, S., Khalifa, S., & Schuller, B. W. (2019). Pre-training in Deep Reinforcement Learning for Automatic Speech Recognition. doi:10.48550/ARXIV.1910.11256.

[12] Agrawal, S., Shruti, & Challa, R. (01 2010). PROSODIC FEATURE BASED TEXT DEPENDENT SPEAKER RECOGNITION USING MACHINE LEARNING ALGORITHMS. International Journal of Engineering Science and Technology.

[13] El-mashad, S., Sharway, M., & Zayed, H. (10 2011). SPEAKER INDEPENDENT ARABIC SPEECH RECOGNITION USING SUPPORT VECTOR MACHINE.

[14] A. Banerjee, A., Dubey, A., Menon, A., Nanda, S., & Nandi, G. C. (2018). Speaker Recognition using Deep Belief Networks. doi:10.48550/ARXIV.1805.08865.

[15] A. Singh and R. S. Anand, "Speech Recognition Using Supervised and Unsupervised Learning Techniques," in Proceedings - 2015 International Conference on Computational Intelligence and Communication Networks, CICN 2015, Aug. 2016, pp. 691–696. doi: 10.1109/CICN.2015.320.

[16] Aldarmaki, H., Ullah, A., Ram, S., & Zaki, N. (2022). Unsupervised Automatic Speech Recognition: A review. Speech Communication, 139, 76–91. doi:10.1016/j.specom.2022.02.005.

[17] Baevski, A., Hsu, W.-N., Conneau, A., & Auli, M. (2021). Unsupervised Speech Recognition. doi:10.48550/ARXIV.2105.11084.

[18] Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised Pre-training for Speech Recognition. doi:10.48550/ARXIV.1904.05862.

[19] Kourd, A., & El kourd, K. (01 2016). Arabic Isolated Word Speaker Dependent Recognition System. British Journal of Mathematics & Computer Science, 14, 1–15. doi:10.9734/BJMCS/2016/23034.

[20] A. Muayad Jalil, F. Sahib Hasan, and H. Adnan Alabbasi, "ROBUST HYBRID FEATURES BASED TEXT INDEPENDENT SPEAKER IDENTIFICATION SYSTEM OVER NOISY ADDITIVE CHANNEL," Journal of Engineering and Sustainable Development, vol. 24, no. 04, pp. 56–70, Jul. 2020, doi: 10.31272/jeasd.24.4.7.

[21] J. McKechnie, B. Ahmed, R. Gutierrez-Osuna, P. Monroe, P. McCabe, and K. J. Ballard, "Automated speech analysis tools for children's speech production: A systematic literature review," International Journal of Speech-Language Pathology, vol. 20, no. 6, pp. 583–598, Oct. 2018, doi: 10.1080/17549507.2018.1477991.

[22] Mustafa, M., Rosdi, F., Salim, S. S., & Mughal, M. U. (05 2015). Exploring the Influence of General and Specific Factors on the Recognition Accuracy of an ASR System for Dysarthric Speaker. Expert Systems

[23] with Applications, 42, 3924–3932. doi: 10.1016/j.eswa.2015.01.033.

[24] M. Cutajar, E. Gatt, I. Grech, O. Casha, and J. Micallef, "Comparative study of automatic speech recognition techniques," IET Signal Processing, vol. 7, no. 1, pp. 25–46, Feb. 2013, doi: 10.1049/iet-spr.2012.0151.

[25] D. Yu and L. Deng, Automatic Speech Recognition. London: Springer London, 2015. doi: 10.1007/978-1-4471-5779-3.

[26] Wang D, Wang X, Lv S. An Overview of End-to-End Automatic Speech Recognition. Symmetry. 2019; 11(8):1018. https://doi.org/10.3390/sym11081018.

[27] Saon, G., Soltau, H., Nahamoo, D., Picheny, M., 2013. Speaker adaptation of neural network acoustic models using i-vectors. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. pp. 55–59. http://dx.doi.org/10.1109/ ASRU.2013.6707705.

[28] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5329–5333. doi:10.1109/ICASSP.2018.8461375.

[29] Ko, Tom, Peddinti, Vijayaditya, Povey, Daniel, Khudanpur, S., 2015. Audio augmentation for speech recognition. In: INTERSPEECH.

[30] J. Padmanabhan and M. J. J. Premkumar, "Machine learning in automatic speech recognition: A survey," IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India), vol. 32, no. 4. Taylor and Francis Ltd., pp. 240–251, 2015. doi: 10.1080/02564602.2015.1010611.

[31] Jayashree Padmanabhan & Melvin Jose Johnson Premkumar (2015) Machine Learning in Automatic Speech Recognition: A Survey, IETE Technical Review, 32:4, 240-251, DOI: 10.1080/02564602.2015.1010611.

[32] Learned-Miller, E. G. (2014). Introduction to supervised learning. I: Department of Computer Science, University of Massachusetts, 3.

[33] Karhunen, J., Raiko, T., & Cho, K. (2015). Unsupervised deep learning: A short review. Advances in independent component analysis and learning machines, 125-142.

[34] Hajighorbani, M., Hashemi, S. R., Minaei-Bidgoli, B., & Safari, S. (2016, May). A review of some semi-supervised learning methods. In IEEE-2016, first international conference on new research achievements in electrical and computer engineering (pp. 1-10).

[35] Li, Y. (2017). Deep Reinforcement Learning: An Overview. doi:10.48550/ARXIV.1701.07274.

[36] Trinh, V. A., & Braun, S. (2022, May). Unsupervised Speech Enhancement with speech recognition embedding and disentanglement losses. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp.391-395). IEEE. doi:10.48550/ARXIV.2111.08678.

نظام التعرف على الكلام غير الخاضع للإشراف وشبه الإشراف
ورقة مراجعة

يسرى فيصل محمد

قسم علوم الحاسوب، كلية علوم
الحاسوب والرياضيات، جامعة
الموصل، الموصل، العراق

yusrafaisalcs@uomosul.edu.iq

مازن علي فاضل

هيئة النزاهة، مديرية تحقيق نينوى
نينوى، الموصل، العراق

mazin.20csp74@student.uomosul.edu.iq

تاريخ الاستلام:1/9/2022 **تاريخ القبول:20/11/2022**

الملخص

الصوت هو مقياس بيولوجي سلوكي قد يكشف عن عمر الشخص وجنسه وعرقه وحالته العاطفية. التعرف على المتحدث هو طريقة التعرف على الأفراد من خلال أصواتهم. على الرغم من حقيقة أنه على مدار العقود الثمانية الماضية، كان الأكاديميون يركزون بالفعل على تحديد المتحدثين، فإن التطورات التكنولوجية مثل إنترنت الأشياء (IoT)، والمنازل الذكية، والمساعدين الصوتيين، والأدوات الذكية، والإنسان جعلت استخدامها شائعًا في المجتمع الحديث. تقدم هذه

الدراسة تحليلاً شاملاً لمراجعات تحديد المتحدث. يبحث في التطورات الأخيرة وكذلك المشاكل في هذا المجال من الدراسة. تبحث هذه الدراسة في استخراج الميزات والمصنفات وهيكل نظام التعرف على المتحدث. يتم أيضًا تناول كيفية استخدام التعرف على المتحدث في التطبيقات. الهدف هو زيادة فهم الباحثين للتعرف على المتحدث من خلال التعلم الآلي حيث أظهرت الأبحاث الحديثة أنه من السهل خداع التعلم الآلي في إنتاج تنبؤ غير صحيح.

الكلمات المفتاحية : نماذج الخليط الغاوسي (GMM) ، التعرف التلقائي على الكلام (ASR) ، نماذج ماركوف المخفية (HMM) ، التعلم الآلي (ML) ، الإدراك متعدد الطبقات (MLP) ، الشبكة العصبية الاحتمالية ذات الأساس الشعاعي (RBPNN) ، آلة ناقلات الدعم (SVM) ، تعلُّم تكميم المتجهات (LVQ) ، معاملات cepstral ذات تردد التردد (MFCCs) ، التفاف الوقت الديناميكي (DTW) ، التدريب المتغير للخليط (MixIT).