IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 5, October, 2023, pp. 2121-2129
Journal homepage: http://iieta.org/journals/ts

A Comprehensive Review on Machine Learning Approaches for Enhancing Human Speech Recognition

Maha Adnan Shanshool*, Husam Ali Abdulmohsin

Computer Science Department, College of Science, University of Baghdad, Al-Jadriya, Baghdad 10070, Iraq

Corresponding Author Email: Maha.jabr2101m@sc.uobaghdad.edu.iq

https://doi.org/10.18280/ts.400529

Received: 6 April 2023
Revised: 24 July 2023
Accepted: 8 September 2023
Available online: 30 October 2023

Keywords:
speech recognition, ASR, speaker recognition

ABSTRACT

As a fundamental element of human-computer interaction, speech recognition—the ability of software systems to identify and interpret human language—has garnered immense attention in recent years. This review offers a rigorous examination of machine learning techniques deployed for optimizing speech recognition capabilities. It delves into the utilization of prominent datasets—such as Librispeech, Timit, and Voxforge—in speech recognition research and underscores their significant contributions to enhancing the accuracy of recognition systems. Furthermore, the efficacy of assorted classification techniques—including deep neural networks (DNN), convolutional neural networks (CNN), support vector machines (SVM), and random forests (RF)—is evaluated in the context of voice recognition. It is observed that Mel-Frequency Cepstral Coefficients (MFCC) often render superior discriminatory abilities in human voice recognition trials. This review stands to provide valuable insights for both researchers and professionals active in the field of speech recognition, thereby paving the way for future advancements in this domain.

1. INTRODUCTION

Speech recognition, commonly referred to as automatic speech recognition (ASR), is an area of research that tries to make it possible for computers to comprehend and interpret spoken language. Due to the rising need for intelligent voice-controlled systems and the spread of speech-enabled gadgets like smartphones, virtual assistants, and smart home appliances, it has attracted considerable interest and achieved tremendous breakthroughs in recent years. Numerous fields, including human-computer interaction, natural language processing, telecommunications, healthcare, and automotive technology, can benefit from the capacity to effectively transcribe and understand spoken language. The development of algorithms and models that can automatically translate spoken language into written text is the main objective of voice recognition, which enables computers to interpret and comprehend human speech. By offering more practical and effective ways to enter data, conduct searches, and issue commands, this technology has altered the way people engage with machines. Real-time voice recognition and interpretation capabilities of speech recognition systems make it easier for people with physical limitations to use digital services and enable hands-free communication. Speech recognition technology has developed over time, moving from straightforward isolated word recognition systems to more robust, robust models that can handle continuous speech and adapt to different speakers and languages. Traditional methods merged acoustic and language models with statistical techniques like Hidden Markov Models (HMMs). However, with the introduction of deep learning and neural networks, the industry has made major strides, which have resulted in notable advances in voice recognition performance and accuracy. The goal of this survey article is to give a thorough overview of the most recent developments and methods in the field of voice recognition. The development of speech recognition systems will be known, from the best datasets. Features extraction, classification methods to more recent advancements based on deep learning.

We can learn more about machine learning's potential to revolutionize numerous industries and enhance human-computer interaction by examining how it affects human speech recognition. Developed a method for locating a speaker in an audio stream based on the biometric characteristics of the human voice, such as pitch, loudness, and frequency, a model for unsupervised learning that can learn speech representation from a little dataset. This study made use of the Librispeech dataset, and were able to attain a word error rate of 1.8 [1].

Speech recognition is a field that tries to make it possible for computers to comprehend and interpret human speech while taking into consideration the distinctive qualities that set each person apart from the other, such as speaking style, accent, method of pronunciation, and rhythm. In this field, speaker recognition, which includes speaker verification and speaker identification, is extremely important. Determine the identity of a speaker from a known collection of voices by using speaker identification [2]. The development of speech recognition technology has allowed individuals to control their digital gadgets only with voice commands, doing away with the need for more conventional input devices like a mouse or keyboard [3]. Speech recognition technology has been shown to be a useful tool for language development, helping people to pronounce words well and express themselves more clearly. These systems have also helped to increase the accessibility of education for students who are blind [4]. Speech recognition systems frequently use machine learning techniques, particularly when training speaker utterances using datasets. By examining test utterances, these trained algorithms can

then recognize speakers [5]. Recent studies have shown the value of utilizing the signal's power spectral density as a way to improve speaker verification. A unique approach has been put forth that makes use of the signal's power spectrum density and a feature vector with reduced Mel-frequency cepstral coefficients (MFCC) [6]. In speaker recognition, deep learning techniques have also shown promising results. Deep Locally Connected Networks (LCN) and Convolutional Neural Networks (CNN) have both been studied for their efficacy in identifying speakers from text [7]. Additionally, a whole deep learning-based speech system has been created, and it has the potential to outperform existing recognition pipelines, especially in difficult situations like talks and noisy surroundings [8]. CNNs have also been used for speech recognition, directly enabling different types of voice variability through the network's structure [9]. A thorough examination into domain robustness has been carried out [10], which involved training a single model utilizing data from several application domains, sampling rates, noise conditions, and codec settings. In the fields of object and pattern recognition, datasets are essential. we will examine the most popular types of datasets, the best feature extraction strategies, and the most often used algorithms in this subject.

This paper is organized as section 1 shows the introduction, section 2 shows related works of past researchers, section 3 comparsion of speech recognition and speaker recognition, section 4 Types of speaker recognition 5 Types of ASR methods 6 classification of ASR system 7shows the state-of-the-art works, and section 8 shows the discussion and conclusion.

2. RELATED WORK

The scholarly landscape of speech recognition has been enriched by extensive research efforts. In 2015, foundational aspects of speech recognition systems, as well as several techniques for feature extraction and pattern matching, were discussed [11]. The ensuing year witnessed a comprehensive presentation of the architecture, voice parameterization, methodologies, characteristics, challenges, databases, tools, and applications associated with speech recognition [12].

Further exploration in 2017 introduced an examination of the principles steering speech recognition systems, bolstered by a detailed analysis of various feature extraction and pattern-matching techniques [13]. This year also saw the unveiling of a brief description of an ASR-based approach to speech therapy, specifically targeting patients with apraxia [14].

By 2018, there was a shift in focus towards the review of ASR error detection and correction techniques, with particular emphasis on approaches grounded in word error rate metrics [8]. In 2019, a comprehensive analysis emerged, dissecting numerous studies conducted for voice applications since 2006, the year deep learning first carved out its niche in the machine learning sphere [15].

The year 2021 marked the evolution of attention models for Transformer and recurrent neural network-based offline and streaming speech recognition architectures [16]. In the subsequent year, research pivoted towards identifying models and concepts with the potential to facilitate fully unsupervised ASR. This included unsupervised sub-word and word modeling, unsupervised speech signal segmentation, and unsupervised mapping from speech segments to text [17]. In 2022, The authors' goal is to review and compile the most recent research on Arabic Part of Speech (APoS), highlighting tagger techniques for the Arabic language that should be used to build corpora for the Arabic language [18]. These studies collectively illuminate the principles and methodologies underpinning voice recognition systems, the associated challenges, applications, error correction techniques, and advancements in the deployment of deep learning.

3. COMPARSION BETWEEN SPEECH RECOGNITION AND SPEAKER RECOGNITION

3.1 Speech recognition

Speech recognition is the process of translating spoken language into written text utilizing computers. This entails analyzing recorded speech and extracting distinguishing aspects such as spectral, temporal, and frequency characteristics, Speech recognition's major objective is to make it possible for people to communicate with computers by speaking commands, using transcription services, or using voice-activated programs, Voice assistants, dictation software, call center automation, voice-controlled systems, and transcription services all frequently make use of speech recognition technologies. Speech recognition algorithms are used to extract these features and compare them to pre-existing models to accurately recognize spoken words and phrases. Speech recognition technology enables voice input into devices. Other forms of input, such as typing, clicking, or choosing in another way, are replaced by the technology [19]. Based on their recorded speech, it seeks to identify a person.

3.2 Speaker recognition

Speaker recognition is a type of technology that uses a person's distinctive voice characteristics to identify or verify them. It emphasizes identifying and verifying the speaker's identity, speaker identification uses voice patterns, such as vocal pitch, tone, accent, and speech features, to analyze and compare people in order to identify or authenticate them. In order to identify or verify speakers, speaker recognition systems use algorithms to construct voiceprints or speaker models from speech samples. Security systems, access control, speech biometrics, forensic investigations, and speaker identification in audio recordings all use speaker recognition technologies. Speaker recognition systems place more emphasis on identifying and classifying people based on their distinctive vocal traits than they do on transcription or content comprehension. This procedure involves evaluating the speaker's speech and extracting unique biometric data such as voice parameters (e.g., frequency, pitch, duration) and prosodic variables (e.g., intonation, stress). Speaker identification algorithms construct speaker-specific models to recognize the speaker's identity and compare them with different speakers. The process of identifying human voice using artificial intelligence methods, speaker identification methods are widely used in speech authentication, security and surveillance, electronic voice eavesdropping, and identity verification [5].

4. TYPES OF SPEAKER RECOGNITION

4.1 Text-dependent

Text-dependent the procedure described by ASR makes the test utterance equivalent to the text used during the registration step [20].

The test subject is acquainted with the model already. The local lexicon has insufficient enrolment and trial phases to produce an accurate outcome. It still has a few technological and scientific obstacles to overcome. The first text-dependent speaker identification system, which was developed in the 1990s, introduced the key components of the current state of the art through the use of feature extraction, speaker modeling, and score normalization with a likelihood ratio score [20].

4.2 Text-independent

The text-independent speaker identification system identifies speakers without regard to the text that they are speaking [5].

The speaker can freely talk to the system, making text-independent speaker recognition more convenient than text-dependent speaker recognition system (SRS). To get improved accuracy, however, it needs to undergo longer training and testing sessions [20].

5. TYPES OF ASR METHODS

5.1 Open set

Open set ASR techniques are intended to handle speech recognition tasks where the system must convert spoken words into text while allowing for the potential to come across words or expressions that are not already part of the system's established vocabulary. The following are some essential traits and methods related to open set ASR:

Flexible Vocabulary: Open set ASR systems can recognize a variety of words and expressions, even those that aren't expressly listed in their vocabulary.

Out-of-Vocabulary (OOV) Handling: By utilizing contextual information, methods including language models, statistical language models, and neural language models are frequently utilized to enhance OOV word recognition.

Open set ASR systems need to be capable of handling recognition faults well because OOV terms might increase error rates. These errors are minimized using methods like confidence scoring and error correction procedures.

Benefits: Open set ASR offers more flexibility in addressing different speech recognition jobs, especially in situations where the vocabulary is dynamic or constantly changing.

Inaccuracy and vocabulary coverage are hampered by the inclusion of OOV words because the system must successfully generalize and handle unfamiliar words.

5.2 Closed set

Closed set ASR methods are intended for voice recognition jobs where the system only recognizes words or phrases that fall under the scope of a specified vocabulary. Here are some crucial traits and methods related to closed set ASR. Closed set ASR systems are limited in their ability to recognize only the words and phrases that are part of their vocabulary, which is frequently compiled from a single domain or application.

Pronunciation Modeling: Closed set ASR systems frequently use methods like using phonetic dictionaries or acoustic models trained on domain-specific data to model word pronunciations accurately in order to increase accuracy.

Closed set ASR systems can gain from language model optimization that is explicitly customized to the vocabulary and domain of interest.

Closed set ASR systems have the advantage of high recognition accuracy within the limited vocabulary, resulting in more accurate and trustworthy transcriptions.

The limitation of a predefined vocabulary makes closed set ASR less appropriate for applications needing vocabulary expansion or dynamic speech recognition tasks since it restricts the system's flexibility in handling new or unfamiliar terms. It's important to note that some ASR systems can combine the benefits of both open and closed set techniques to meet particular application requirements. The task-specific requirements, including vocabulary quantity, vocabulary variability, and the system's ability to handle new or unfamiliar terms, will determine whether open set or closed set ASR should be used.

6. CLASSIFICATION OF ASR SYSTEM

1. Based on System Architecture

A. Conventional ASR Systems:

Conventional ASR systems relate to conventional methods that were widely used prior to the emergence of deep learning. Feature extraction, acoustic modeling (using Hidden Markov Models, for example), language modeling, and decoding algorithms (using the Viterbi algorithm, for example) are frequently included in these systems.

Benefits: Traditional ASR systems are computationally effective and have a solid foundation. They have been extensively utilized and researched for many years.

Drawbacks: Conventional ASR systems may have trouble with big vocabulary sizes and complex speech patterns. In difficult circumstances, they might not function at the cutting edge.

B. ASR Systems Based on Deep Learning:

Deep neural networks (DNNs) or recurrent neural networks (RNNs) are used in deep learning-based ASR systems to directly learn complicated patterns and representations from speech input. These systems frequently include decoding components, acoustic models, and language models.

Benefits: Deep learning-based ASR systems have demonstrated considerable performance increases, particularly in noisy and large vocabulary voice recognition tasks. They are able to record complex audio correlations as well as language context.

Drawbacks: Deep learning-based ASR systems need a lot of labeled training data as well as a lot of processing power to train. When compared to standard ASR systems, they could be more difficult to implement and perfect.

2. Based on the availability of data

A. Continuous Speech Recognition with a Large Vocabulary (LVCSR):

LVCSR systems are designed to recognize continuous speech in situations with a huge vocabulary, often involving vocabularies with tens of thousands to millions of words.

Large vocabularies are handled by these systems, which also provide correct transcriptions.

Benefits: LVCSR systems are appropriate for applications requiring extensive vocabulary support and assistance with unrestricted speech recognition tasks.

LVCSR systems may need a lot of computer power and training data, which is a drawback. The handling of unfamiliar words and the management of speech changes can be more difficult. Systematic keyword spotting (KWS)

B. KWS systems concentrate on finding particular words or phrases within a given speech input. These systems are appropriate for applications such as voice-controlled assistants or command-based interfaces since they are built to swiftly recognize certain target words or phrases.

Benefits: KWS systems are effective in locating target keywords rapidly, which lowers the amount of computation needed and speeds up response times.

KWS systems' limited vocabulary coverage and potential difficulty processing speech inputs devoid of the targeted keywords are drawbacks.

3. According to the training paradigm

A. Supervised ASR: The speech signals in labeled data are aligned with the associated transcriptions to train supervised ASR systems. To reduce the discrepancy between predicted and actual transcriptions, the models are optimized.

Advantages: When trained on accurately transcribed data, supervised ASR achieves excellent accuracy. It enables accurate modeling and text-speech alignment.

Drawbacks: Supervised ASR relies largely on annotated data, the production of which can be expensive and time-consuming. In managing unseen or outside-of-domain communication, it can have restrictions.

B. Semi-Supervised ASR: In semi-supervised ASR, training involves combining a smaller amount of labeled data with a greater amount of unlabeled data. The models gain knowledge from both labeled and unlabeled data, utilizing the additional knowledge to enhance their performance. Cons: It can be difficult to choose and use unlabeled data properly. For efficient model training, a certain volume of labeled data might still be necessary.

The goal of unsupervised ASR is to train ASR systems without using any labeled data. The models do not require explicit transcription information; instead, they learn directly from the input speech sounds.

Benefits: Since unsupervised ASR doesn't require labeled data, it can be applied in situations where there is little to no annotated data. In contexts with limited resources, it can help with ASR. First of all, it can be challenging to determine where a word begins and ends. Another issue is that each phoneme's creation is influenced by the production of the phonemes around it.

7. STATE OF THE ARTWORKS SURVEY

As shown in Table 1 about the Librispeech dataset for different authors and numerous time there are authors used the Librispeech dataset with machine learning algorithms and deep learning, in study [5] Findings indicated that on the LibriSpeech dataset, MFCCT features in combination with DNN outperformed the baseline MFCC and time-domain features. In study [1] Librispeech was good but require a more detailed dataset with labels for specific individuals on their audio file.

Table 1. The literature survey on researchers that adopted the libri speech dataset in speech recognition

Ref.	Year	Classification	Acc.%
[5]	2020	DNN	99.94%
[21]	2020	AFEASI	99.05%
[22]	2020	acoustic modeling (AM), combined with neural LM rescoring	97.4%
[23]	2021	CTC	90.02%
[24]	2021	Monotonic Chunkwise Attention (MoCha)	94.21%
[1]	2022	Unsupervised Siamese NN combined with CNN	98.2%
		Average of all works	96.47%

Table 2. The literature survey on researchers that adopted the TIMIT dataset in speech recognition

Ref.	Year	Classification	Acc.%
[25]	2016	CNN	97%
[6]	2016	GMM	88.3%
		Average of all works	92.65%

Table 3. The literature survey on researchers that adopted the voxforge dataset in speech recognition in speech recognition

Ref.	Year	Classification	Acc.%
[26]	2015	PNN	94%
[27]	2018	CNN	98.8%
		Average of all works	96.4%

Table 4. The literature survey on researchers that adopted the OK Google dataset in speech recognition in speech recognition

Ref.	Year	Classification	Acc.%
[7]	2015	LCN	96.4%
		CNN	96.48%
[28]	2016	DNN	98%
		Average of all works	97.2%

The authors used the TIMIT data set, and they reached good results without the need for handcrafted features.

The authors used an OK google dataset collected from anonymized voice search logs. For improved noise robustness they perform multi-style training.

In Table 3 the authors used the Voxforge dataset, which is a free and open-source voice database where various speakers have freely provided speech data towards the creation of speech recognition software. The Voxforge database was picked because it was created with issues like channel variance, session variation, and noise robustness in mind, whereas the majority of corpora for speaker verification were not. Thus, achieving the best results. In Table 4 the authors used an OK google dataset collected from anonymized voice search logs. For improved noise robustness they perform multi-style training.

As shown in Table 5 the authors used CNN classification methods, in study [29] present convolutional neural networks (CNNs) can model raw and tonal speech signals the experimental results show that the current CNN architecture performs significantly, with an accuracy rate of 89.15% and a WER of 10.56% for continuous and broad vocabulary sentences of speech signals with various tones. And in study [27] present a speaker verification method that uses end-to-end CNNs to learn speaker discriminative data straight from the raw audio signal, results have been obtained 98.8%, in study

[30] introduced Effective speaker recognition in emotional and noisy talking situations: GMM-CNN model, results have been obtained 84.69%. The previous study [7] analyzes the effectiveness of deep Locally Connected Networks (LCN) and Convolutional Neural Networks (CNN) at recognizing speakers from text, results have been obtained is 96.48%.

Finally in terms of the neural network when CNN is used good results were obtained, but other methods that gave better results.

As shown in Table 6 the authors used the SVM classification method, in study [24] Mel-Frequency Cepstral Coefficients (MFCC), and statistical features are employed as the models' input features in Support Vector Machine (SVM) and Random Forest (RF) models, Compared to Random Forest, the Support Vector Machine has encouraging results with an accuracy of 94%. The authors [25] used a comparative analysis of various classifiers to concentrate on discovering appropriate voice signal properties. The results have shown that the best accuracy in voice pathology detection is achieved using the Support Vector Machine algorithm this approach has an accuracy of roughly 86% when classifying a voice as pathological or healthy. the authors used the output of a genetic algorithm (GA) and the inputs of a NN algorithm were merged to develop a hybrid feature selection approach. SVM, neural networks, and GMM were used for classification. In terms of SVM, the best results were 94.55% accuracy in identifying illnesses [31].

Table 5. The literature survey on researchers that adopted the CNN classification methods in speech recognition in speech recognition

REF	Year	Feature Extraction	Acc.%
[7]	2015	a small footprint global password TD-SV task. Weight, depth, multiple	96.48%
[32]	2015	MFCC	93%
[25]	2016	Spectrogram of voice data, MFCC and GMM	97%
[33]	2017	Spectrogram, MFCC, and CMVN	80.5%
[34]	2018	Frame-wise MFEC	87.3%
[27]	2018	Raw speech data	98.8%
[35]	2019	Multiple audio (wav, flac, mono, stero) Power spectrm log mel (mfsc) Mfcc Fftw	95.09%
[30]	2021	Angry, neural, slow, loud, fast Mfcc MFCC	84.68%
[29]	2022	LibROSA Fourier transformation Based on melfiltter bank	89.15%
[36]	2022	Higher-order spectral analysis HOSA COVAREP, HOSA, fused feature	85%
		Average of all works	90.69%

Table 6. The literature survey on researchers that adopted the SVM classification methods in speech recognition in speech recognition

REF	Year	Feature Extraction	Acc.%
[37]	2018	MFCC	86%
[38]	2019	MFCC, LPC	94%
[39]	2022	MFCC, RFE, MRMR, CHI-2	94%
		Average of all works	90%

As shown in Table 7 the authors used the RF classification method, in study [6] to demonstrate various audio preprocessing techniques, such as noise reduction and vocal augmentation, to enhance the audios that are now available in real scenarios. The results that the classification procedure was more accurate when a machine learning classifier was used, with (RF) classifiers achieving 97.9% accuracy. Mel-Frequency Cepstral Coefficients (MFCC) [39] and statistical features are employed as the models' input features in Support Vector Machine (SVM) and Random Forest (RF) models, the result has been obtained with RF is 83%. When the authors used RF and SVM achieving the good results but there are other methods achieve better results.

Table 7. The literature survey on researchers that adopted the RF classification methods in speech recognition in speech recognition

REF	Year	Feature Extraction	Acc.%
[6]	2016	MFCC	88.3%
[40]	2017	MFCC	97.8%
		Average of all works	93.05%

As shown in Table 8 the authors used the GMM classification method, study [41] presented Mel-Frequency To model speakers, the Gaussian mixture model-universal background model is used, and cepstral coefficients are used for feature extraction, results show 97.8%. Paper [6] presented a reduced feature vector that makes use of fresh information gleaned from the speaker's speech to carry out GMM-based text-free speaker verification applications, results show 88.3%.

As shown in Table 9, the authors used the DNN classification method. Jahangir et al. [5] introduced a novel fusion of time-based and MFCC features (MFCCT), which combines the efficiency of time-domain and MFCC features to boost the precision of text-independent speaker identification (SI) systems. The speaker identification model was created using the retrieved MFCCT features as input from a deep neural network (DNN), DNN obtained better classification results compared with five machine learning algorithms that were recently utilized in speaker recognition, results show 99.94%. Kabir et al. [20] introduced a novel end-to-end method for speaker verification that uses the same loss for training and evaluation and directly maps the utterance to a score while jointly optimizing the internal speaker representation and the speaker model, proposed an approach improved our best small footprint DNN baseline from over 3% to 2% equal error rate on our internal “Ok Google” benchmark. By looking at the results of the classification methods, it was found that DNN is the method that achieved the highest results.

Table 8. The literature survey on researchers that adopted the GMM classification methods in speech recognition in speech recognition

REF	Year	Feature Extraction	Acc.%
[41]	2021	MFCC, VQ	97.9%
[39]	2022	MFCC	83%
		Average of all works	90.45%

Table 9. The literature survey on researchers that adopted the DNN classification methods in speech recognition in speech recognition

REF	Year	Feature Extraction	Acc.%
[28]	2016	LSTM, DNN	98%
[42]	2018	MFCC	99.75%
[5]	2020	MFCC	99.94%
		The average of all works	99.23%

Table 10. State-of-the-art researcher that adopted MFCC feature extraction in Speech recognition

REF	Year	Feature Extraction	Acc.%
[25]	2016	MFCC and GMM	97%
[33]	2017	MFCC	80.5%
[43]	2017	MFCC	80.36%
[44]	2018	MFCC	83%
[34]	2018	MFEC	87.3%
[37]	2018	MFCC	86%
[44]	2019	EEG	99.38%
[30]	2021	MFCC	84.68%
[3]	2021	MFCC	88.21%
[41]	2021	MFCC, VQ	97.9%
[39]	2022	MFCC	94%
[29]	2022	MFCC, LibROSA	89.15%
[45]	2022	MFCC	98.4%
		The average of all works	89.68%

As shown in Table 10, the authors used MFCC feature extraction, Dua et al. [29] introduced CNNs, MFCCs, and LibROSA were incorporated to reveal the best system performance for recognizing uncommon input speech signals. And in study [24] the key benefit of the MFCC is that it is good at error reduction and capable of producing a robust feature when the signal is influenced by noise. It is a leading strategy and frequently used algorithm in speech feature extraction. MFCC is a technique for detecting frequencies above 1kHz that makes use of human hearing activity [41]. Looking at previous research, we found that MFCC has the best features.

8. DISCUSSION

Speech is the most fundamental, widely used, and effective type of interactivity between individuals. It might be challenging to distinguish between the phonetic content and the auditory variances in utterances made by various speakers.

From our reviewer in the tables above, we discussed different techniques for different authors and in numerous times. In Table 1 the authors used the Librispeech dataset for different authors, and numerous times there are authors used the Librispeech dataset with machine learning algorithms and deep learning. Findings indicated that on the LibriSpeech dataset, MFCCT features in combination with DNN outperformed the baseline MFCC and time-domain features, librispeech was good but require a more detailed dataset with labels for specific individuals on their audio file.

In Table 2, the authors used the TIMIT dataset, and they reached good results without the need for handcrafted features. In Table 3 the authors used the Voxforge dataset, which is a free and open-source voice database where various speakers have freely provided speech data towards the creation of speech recognition software. The Voxforge database was picked because it was created with issues like channel variance, session variation, and noise robustness in mind, whereas the majority of corpora for speaker verification were not. Thus, achieving the best results. In Table 4 the authors used an OK google dataset collected from anonymized voice search logs. For improved noise robustness they perform multi-style training. In Table 5 the authors used CNN classification methods, convolutional neural networks (CNNs) can model raw and tonal speech signals. The experimental results show that the current CNN architecture performs significantly, with an accuracy rate of 89.15% and a WER of 10.56% for continuous and broad vocabulary sentences of speech signals with various tones. A speaker verification method that uses end-to-end CNNs to learn speaker discriminative data straight from the raw audio signal, results have been obtained 98.8%. Effective speaker recognition in emotional and noisy talking situations: GMM-CNN model, results have been obtained 84.69%. Analyzes the effectiveness of deep Locally Connected Networks (LCN) and Convolutional Neural Networks (CNN) at recognizing speakers from text, results have been obtained is 96.48%. Finally in terms of the neural network when CNN is used good results were obtained, but other methods gave better results. In Table 6 the authors used the SVM classification method, Mel-Frequency Cepstral Coefficients (MFCC), statistical features are employed as the models' input features in Support Vector Machine (SVM) and Random Forest (RF) models, Compared to Random Forest, the Support Vector Machine has encouraging results with an accuracy of 94%. The authors used a comparative analysis of various classifiers to concentrate on discovering appropriate voice signal properties. The results have shown that the best accuracy in voice pathology detection is achieved using the Support Vector Machine algorithm this approach has an accuracy of roughly 86% when classifying a voice as pathological or healthy.

In Table 7 the authors used the RF classification method; to demonstrate various audio preprocessing techniques, such as noise reduction and vocal augmentation, to enhance the audios that are now available in real scenarios, the results that the classification procedure was more accurate when a machine learning classifier was used, with (RF) classifiers achieving 97.9% accuracy.

Mel-Frequency Cepstral Coefficients (MFCC) and statistical features are employed as the models' input features in Support Vector Machine (SVM) and Random Forest (RF) models, result have been obtained with RF is 83%. When the authors used RF and SVM to achieve good results but other methods that achieve better results. In Table 8 the authors used the GMM classification method, Mel-Frequency to model speakers, the Gaussian mixture model-universal background model is used, and cepstral coefficients are used for feature extraction, results show 97.8%. Reduced feature vector that makes use of fresh information gleaned from the speaker's speech to carry out GMM-based text-free speaker verification applications, results show 88.3%. From Table 9 the authors used the DNN classification method, a novel fusion of time-based and MFCC features (MFCCT), which combines the efficiency of time-domain and MFCC features to boost the precision of text-independent Speaker Identification (SI) systems.

The speaker identification model was created using the retrieved MFCCT features as input from a Deep Neural Network (DNN), DNN obtained better classification results compared with five machine learning algorithms that were recently utilized in speaker recognition, results show 99.94%. By looking at the results of the classification methods, it was found that DNN is the method that achieved the highest results. In Table 10 the authors used MFCC feature extraction CNNs, MFCCs, and LibROSA incorporated to reveal the best system performance for recognizing uncommon input speech signals. The key benefit of the MFCC is that it is good at error reduction and capable of producing a robust feature when the signal is influenced by noise.

It is a leading strategy and frequently used algorithm in

speech feature extraction. MFCC is a technique for detecting frequencies above 1kHz that makes use of human hearing activity. Looking at previous research, we found that MFCC has the best features.

9. CONCLUSION

The following important issues are highlighted in this review, which concentrates on the field of speech recognition:

Datasets: TIMIT, LibriSpeech, and Voxforge are the three most often used datasets in speech recognition.

The frequently used dataset TIMIT contains recordings of speakers from different parts of the United States, representing diverse dialects and speech patterns.

English audiobooks make up the dataset LibriSpeech, which was created especially for Large Vocabulary Continuous Speech Recognition (LVCSR) problems. A free and open-source voice database called Voxforge deals with speech recognition problems such channel variance, session variance, and noise robustness. Methods of classification: In speech recognition, DNN, CNN, RF and SVM are excellent classification techniques. DNN's performance in effectively identifying speech patterns and differentiating between speakers is demonstrated by its approximate result of 99% accuracy. CNN's efficiency in speech recognition tasks was demonstrated by the approximate result of 90.69% accuracy.

SVM demonstrated its capacity to categorize speech patterns by achieving an approximate result of 90% accuracy.

Strong Qualities: Mel-Frequency Cepstral Coefficients (MFCC) are the most effective characteristic for voice recognition, especially in distinguishing people. This paper emphasizes the value of speech recognition research datasets like TIMIT, LibriSpeech, and Voxforge. While highlighting the usefulness of classification techniques like DNN, CNN, and SVM, it points out that MFCC is the most effective feature for distinguishing speech patterns and specific people. These discoveries help the speech recognition industry and its applications continue to advance.

ACKNOWLEDGMENT

Our sincere gratitude to all researchers in the field of speech recognition for their tremendous work in this area and for assisting us in finishing this review study.

REFERENCES

[1] Arshad, S.R., Haider, S.M., Mughal, A.B. (2022). Speaker identification using speech recognition. arXiv Preprint arXiv: 2205.14649. https://doi.org/10.48550/arXiv.2205.14649

[2] Paulose, S., Mathew, D., Thomas, A. (2017). Performance evaluation of different modeling methods and classifiers with MFCC and IHC features for speaker recognition. Procedia Computer Science, 115: 55-62. https://doi.org/10.1016/j.procs.2017.09.076

[3] Mahmood, A., Utku, K. (2021). Speech recognition based on convolutional neural networks and MFCC algorithm. Advances in Artificial Intelligence Research (AAIR), 1(1): 6-12.

[4] Ren, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y. (2019). Almost unsupervised text to speech and automatic speech recognition. In Proceedings of the 36th International Conference on Machine Learning, PMLR, pp. 5410-5419.

[5] Jahangir, R., Teh, Y.W., Memon, N.A., Mujtaba, G., Zareei, M., Ishtiaq, U., Akhtar, M.Z., Ali, I. (2020). Text-independent speaker identification through feature fusion and deep neural network. IEEE Access, 8: 32187-32202. https://doi.org/10.1109/ACCESS.2020.2973541

[6] Chakroun, R., Zouari, L.B., Frikha, M., Hamida, A.B. (2016). Improving text-independent speaker recognition with GMM. In 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), IEEE, pp. 693-696. https://doi.org/10.1109/ATSIP.2016.7523169

[7] Chen, Y.H., Moreno, I.L., Sainath, T.N., Visontai, M., Alvarez, R., Parada, C. (2015). Locally-connected and convolutional neural networks for small footprint speaker recognition. Proceedings Interspeech, 2015: 1136-1140. https://doi.org/10.21437/Interspeech.2015-297

[8] Errattahi, R., El Hannani, A., Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: a review. Procedia Computer Science, 128: 32-37. https://doi.org/10.1016/j.procs.2018.03.005

[9] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., Ng, A.Y. (2014). Deep speech: scaling up end-to-end speech recognition. arXiv Preprint arXiv: 1412.5567. https://doi.org/10.48550/arXiv.1412.5567

[10] Narayanan, A., Misra, A., Sim, K.C., Pundak, G., Tripathi, A., Elfeky, M., Haghani, P., Strohman, T., Bacchiani, M. (2018). Toward domain-invariant speech recognition via large scale training. In 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, pp. 441-447. https://doi.org/10.1109/SLT.2018.8639610

[11] Saksamudre, S.K., Shrishrimal, P.P., Deshmukh, R.R. (2015). A review on different approaches for speech recognition system. International Journal of Computer Applications, 115(22): 23-28. https://doi.org/10.5120/20284-2839

[12] Karpagavalli, S., Chandra, E. (2016). A review on automatic speech recognition architecture and approaches. International Journal of Signal Processing, Image Processing and Pattern Recognition, 9(4): 393-404. https://doi.org/10.14257/ijsip.2016.9.4.34

[13] Vadwala, A.Y., Suthar, K.A., Karmakar, Y.A., Pandya, N., Patel, B. (2017). Survey paper on different speech recognition algorithm: challenges and techniques. International Journal of Computer Applications, 175(1): 31-36.

[14] Jamal, N., Shanta, S., Mahmud, F., Sha'abani, M.N.A.H. (2017). Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: a review. In AIP Conference Proceedings, AIP Publishing, 1883(1): 020028. https://doi.org/10.1063/1.5002046

[15] Nassif, A.B., Shahin, I., Attili, I., Azzeh, M., Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. IEEE Access, 7: 19143-19165. https://doi.org/10.1109/ACCESS.2019.2896880

[16] Karmakar, P., Teng, S.W., Lu, G.J. (2021). Thank you for attention: A survey on attention-based artificial neural networks for automatic speech recognition. arXiv Preprint arXiv: 2102.07259.

https://doi.org/10.48550/arXiv.2102.07259
[17] Aldarmaki, H., Ullah, A., Ram, S., Zaki, N. (2022). Unsupervised automatic speech recognition: a review. Speech Communication, 139: 76-91. https://doi.org/10.1016/j.specom.2022.02.005
[18] Mohammed, Z.K., Abdullah, N.A. (2022). Survey for Arabic part of speech tagging based on machine learning. Iraqi Journal of Science, 63(6): 2676-2685. https://doi.org/10.24996/ijs.2022.63.6.33
[19] Jebbar, M., Maizate, A., Abdelouahid, R.A. (2022). Moroccan's Arabic speech training and deploying machine learning models with teachable machine. Procedia Computer Science, 203: 801-806. https://doi.org/10.1016/j.procs.2022.07.120
[20] Kabir, M.M., Mridha, M.F., Shin, J., Jahan, I., Ohi, A.Q. (2021). A survey of speaker recognition: fundamental theories, recognition methods and opportunities. IEEE Access, 9: 79236-79263. https://doi.org/10.1109/ACCESS.2021.3084299
[21] Li, R.R., Jiang, J.Y., Liu, J.H., Hsieh, C.C., Wang, W. (2020). Automatic speaker recognition with limited data. In Proceedings of the 13th International Conference on Web Search and Data Mining, pp. 340-348. https://doi.org/10.1145/3336191.3371802
[22] Wang, Y.Q., Mohamed, A., Le, D., Liu, C.X., Xiao, A., Mahadeokar, J., Huang, H.Z., Tjandra, A., Zhang, X.H., Zhang, F., Fuegen, C., Zweig, G., Seltzer, M.L. (2020). Transformer-based acoustic modeling for hybrid speech recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 6874-6878. https://doi.org/10.1109/ICASSP40776.2020.9054345
[23] Galvez, D., Diamos, G., Ciro, J., Cerón, J.F., Achorn, K., Gopi, A., Kanter, D., Lam, M., Mazumder, M., Reddi, V.J. (2021). The people's speech: a large-scale diverse English speech recognition dataset for commercial usage. arXiv Preprint arXiv: 2111.09344. https://doi.org/10.48550/arXiv.2111.09344
[24] Kim, C., Garg, A., Gowda, D., Mun, S., Han, C. (2021). Streaming end-to-end speech recognition with jointly trained neural feature enhancement. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 6773-6777. https://doi.org/10.1109/ICASSP39728.2021.9414117
[25] Lukic, Y., Vogt, C., Dürr, O., Stadelmann, T. (2016). Speaker identification and clustering using convolutional neural networks. In 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, pp. 1-6. https://doi.org/10.1109/MLSP.2016.7738816
[26] Ahmad, K.S., Thosar, A.S., Nirmal, J.H., Pande, V.S. (2015). A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network. In 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR), IEEE, pp. 1-6. https://doi.org/10.1109/ICAPR.2015.7050669
[27] Muckenhirn, H., Doss, M.M., Marcell, S. (2018). Towards directly modeling raw speech signal for speaker verification using CNNs. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 4884-4888. https://doi.org/10.1109/ICASSP.2018.8462165
[28] Heigold, G., Moreno, I., Bengio, S., Shazeer, N. (2016). End-to-end text-dependent speaker verification. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Shanghai, China, pp. 5115-5119. https://doi.org/10.1109/ICASSP.2016.7472652
[29] Dua, S., Kumar, S.S., Albagory, Y., Ramalingam, R., Dumka, A., Singh, R., Rashid, M., Gehlot, A., Alshamrani, S.S., AlGhamdi, A.S. (2022). Developing a speech recognition system for recognizing tonal speech signals using a convolutional neural network. Applied Sciences, 12(12): 6223. https://doi.org/10.3390/app12126223
[30] Nassif, A.B., Shahin, I., Hamsa, S., Nemmour, N., Hirose, K. (2021). CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions. Applied Soft Computing, 103: 107141. https://doi.org/10.1016/j.asoc.2021.107141
[31] Abdulmohsin, H.A. (2022). Automatic health speech prediction system using support vector machine. In Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021, Singapore: Springer Nature Singapore, pp. 165-175. https://doi.org/10.1007/978-981-19-0604-6_15
[32] Palaz, D., Doss, M.M., Collobert, R. (2015). Convolutional neural networks-based continuous speech recognition using raw speech signal. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, pp. 4295-4299. https://doi.org/10.1109/ICASSP.2015.7178781
[33] Nagrani, A., Chung, J.S., Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612. https://doi.org/10.21437/Interspeech.2017-950
[34] Torfi, A., Dawson, J., Nasrabadi, N.M. (2018). Text-independent speaker verification using 3d convolutional neural networks. In 2018 IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6. https://doi.org/10.1109/ICME.2018.8486441
[35] Pratap, V., Hannun, A., Xu, Q., Cai, J., Synnaeve, G., Liptchinsky, V., Collobert, R. (2019). Wav2letter++: A fast open-source speech recognition system. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6460-6464. https://doi.org/10.1109/ICASSP.2019.8683535
[36] Miao, X., Li, Y., Wen, M., Liu, Y., Julian, I.N., Guo, H. (2022). Fusing features of speech for depression classification based on higher-order spectral analysis. Speech Communication, 143: 46-56. https://doi.org/10.1016/j.specom.2022.07.006
[37] Verde, L., De Pietro, G., Sannino, G. (2018). Voice disorder identification by using machine learning techniques. IEEE Access, 6: 16246-16255. https://doi.org/10.1109/ACCESS.2018.2816338
[38] Dharmale, G.J., Patil, D.D. (2019). Evaluation of phonetic system for speech recognition on smartphone. International Journal of Innovative Technology and Exploring Engineering, 8(10): 3354-3359. https://doi.org/10.35940/ijitee.J1215.0881019
[39] Alsaify, B.A., Arja, H.S.A., Maayah, B.Y., Al-Taweel, M.M., Alazrai, R., Daoud, M.I. (2022). Voice-based human identification using machine learning. In 2022 13th International Conference on Information and Communication Systems (ICICS), IEEE, Irbid, Jordan,

pp. 205-208. https://doi.org/10.1109/ICICS55353.2022.9811154
[40] Algabri, M., Mathkour, H., Bencherif, M.A., Alsulaiman, M., Mekhtiche, M.A. (2017). Automatic speaker recognition for mobile forensic applications. Mobile Information Systems, 2017. https://doi.org/10.1155/2017/6986391
[41] Ali, A.T., Abdullah, H.S., Fadhil, M.N. (2021). Voice recognition system using machine learning techniques. Materials Today: Proceedings, pp. 1-7. https://doi.org/10.1016/j.matpr.2021.04.075
[42] Miguel, A., Llombart, J., Ortega, A., Lleida, E. (2018). Tied hidden factors in neural networks for end-to-end speaker recognition. arXiv Preprint arXiv: 1812.11946. https://doi.org/10.48550/arXiv.1812.11946
[43] Bagavathi, S., Padma, S.I. (2017). Neural network based voiced and unvoiced classification using EGG and MFCC feature. International Research Journal of Engineering and Technology, 4(4): 1934-1937.
[44] Krishna, G., Tran, C., Yu, J.G., Tewfik, A.H. (2019). Speech recognition with no speech or with noisy speech. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Brighton, UK, pp. 1090-1094. https://doi.org/10.1109/ICASSP.2019.8683453
[45] Abdusalomov, A.B., Safarov, F., Rakhimov, M., Turaev, B., Whangbo, T.K. (2022). Improved feature parameter extraction from speech signals using machine learning algorithm. Sensors, 22(21): 8122. https://doi.org/10.3390/s22218122