Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

A Holistic Review of Automatic Speech Recognition Systems for Real-time Implementation

Bhosale Rajkumar S.

Amrutvahini College of Engineering, Sangamne,

Department of Information Technology,

Maharashtra, India, bhos_raj@rediffmail.com

Panhalkar Archana R.

Amrutvahini College of Engineering, Sangamne,

Department of Information Technology,archana10bhosale@rediffmail.com

Article Info
Page Number: 12341-12359
Publication Issue:
Vol. 71 No. 4 (2022)

Abstract— Once considered a science-fiction idea, Automatic Speech Recognition (ASR) has now become a crucial part of information and communication technology, despite being challenged by various execution-related issues. Over time, ASR has undergone significant changes, evolving from merely responding to sounds to effectively recognizing natural language. With advancements made by researchers, ASR technology and spoken language systems continue to improve. However, there are still challenges in developing flexible solutions that can satisfy user needs in certain situations. In this article, we examine recent techniques related to ASR systems, including pre-processing techniques, classification techniques, feature extraction techniques, and speech recognition techniques.Here some of the recent techniques related to ASR system based on a few factors which are having different application and techniques used for classification. The different application are applicable to the speaker depedent and speaker independent as training part few of them applications are use for real time application and few of them are work on offline basis. One critical aspect of ASR is feature extraction, which involves identifying relevant acoustic features from speech signals to support accurate recognition. To enhance ASR performance, researchers have explored various techniques, including Support Vector Machine (SVM) and Convolutional Neural Network (CNN) for classification. These methods have been used to improve the accuracy of speech recognition and have been applied in numerous applications.

Article History
Article Received: 25 January 2022
Revised: 30 February 2022
Accepted: 15 March 2022

Keywords-Feature Extraction, Preprocessing, Neural Network, Real time application, Automatic Speech Recognition, feature extraction, , classification.

I. INTRODUCTION

Researchers in the speech recognition field are focusing on several key areas of investigation to advance the development of spoken language systems in the speech recognition area, there are several key areas of research for the current development of spoken language systems. [1]. The current focus of research in this field includes a range of critical areas such as automatic

speech recognition, robust speech recognition, and spontaneous speech recognition, among others. These key areas are automatic speech recognition, robust speech recognition, spontaneous speech, etc. [2][3]. Many consumer and industrial applications require fast and lightweight real-time speech recognition with limited vocabularies, such as hands-free control for portable music players, car audio systems, cordless phones and domestic appliances.

To enable continuous real-time speech recognition for mobile, embedded, and hands-free speech applications, a reliable and flexible speech interface is necessary. Applications like speech-controlled GPS navigation systems and speech-controlled song selection for portable music players and car stereos demonstrate the importance of this interface. Furthermore, there is a growing need for advanced natural language applications such as handheld speech-to-speech translation. Mobile, embedded and hands-free speech applications fundamentally require continuous real-time speech recognition. Many current applications, such as speech control of GPS navigation systems and speech-controlled song selection for portable music players and car stereos also require a reliable and flexible speech interface. Finally, sophisticated natural language applications such as handheld speech-to-speech translation [4] require fast and lightweight speech recognition. ASR technology has advanced rapidly in the past decade. While many ASR applications employ powerful computers to handle complex recognition algorithms, there is clearly a demand for an effective solution for embedded systems like portable communication devices and various low-cost consumer electronic systems [5][6][7].

The predominant method used in speech recognition systems today is pattern-matching, with the classifier commonly based on hidden Markov models (HMM). Current speech recognition systems use a pattern-matching approach. The classifier is commonly based on hidden Markov models (HMM) [8]. Considerable progress has been made in Hidden Markov Model (HMM)-based speech recognition technologies, resulting in high recognition performance. As speech input interfaces become more commonplace in mobile devices, it is necessary to balance recognition accuracy, miniaturization, and low-power consumption. HIDDEN Markov model (HMM)-based speech recognition technologies have developed considerably and can now obtain a high recognition performance. The development of speech input interfaces embedded in mobile terminals requires recognition accuracy, miniaturization, and low-power consumption [9]. Previous research on custom hardware described the implementation of the HMM algorithm using application-specific integrated circuits (ASICs) [10] and field-programmable gate arrays (FPGAs) [11], [12]. In the preliminary case study report, a real-time, continuous speech recognition system is discussed that leverages MFCC feature input and Hidden Markov Models (HMM) for optimal performance. A real-time continuous speech recognition system using MFCC feature input and Hidden Markov Models (HMM) is reported in the preliminary case study report in [13]. The effectiveness of real-time speech recognition systems for psychology experiments using MFCC is discussed in [14]. Although such systems perform adequately in quiet environments, they experience significant performance degradation in noisy surroundings. Real-time speech recognition for psychology experiments using MFCC is reported in [14]. The speech recognition systems with MFCC-

derived cestrum work well in clean environments, but speech recognition performance is severely degraded in noisy environments [15].

II. TYPICAL CONVENTIONAL SPEECH RECOGNITION SYSTEM

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

In ASR system three main approaches are followed which are:

- Acoustic Phonetic Approach
- Pattern Recognition Approach
- Statistics-based Approach
- Artificial Intelligence Approach

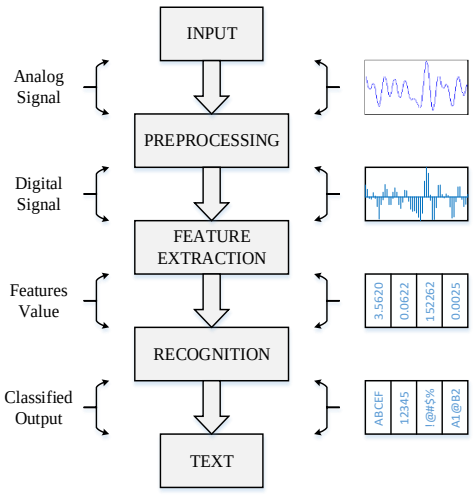

Figure 1. Standard automatic voice recognition system model

Automatic Speech Recognition is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program. The most common process involved in the ASR system is given in fig 1.

A. Pattern Recognition Techniques for Speech Processing

The pattern recognition approach requires no explicit knowledge of speech. This approach has two steps – namely, training of speech patterns based on some generic spectral parameter set and recognition of patterns via pattern comparison. The popular pattern recognition techniques include template matching, Hidden Markov Model [27]. Yang Liu et al. [28] described the ICSI-SRI-UW structural metadata extraction (MDE) system that automatically detects structural events, including sentence boundaries, disfluencies and filler words. A metadata detection system that combines information from different types of textual knowledge sources with information from a prosodic classifier. Kuldeep Kumar and R. K. Aggarwal [29] proposed a Hidden Markov Model Toolkit (HTK) to develop the system. They aimed to build a speech recognition system for the Hindi language.

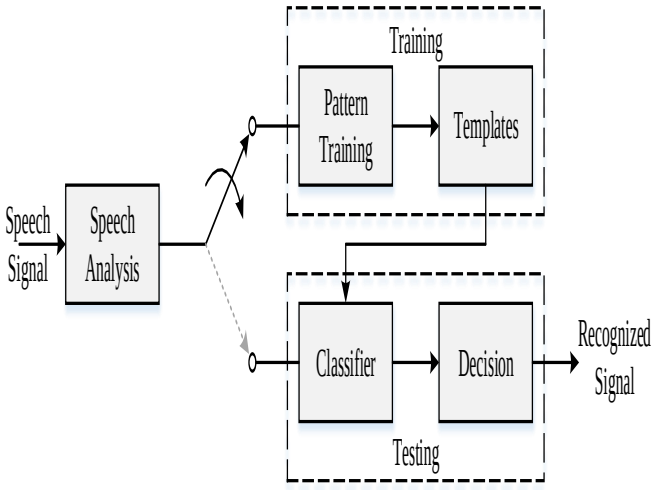

Figure 2. Block representation of speech recognition system based on pattern recognition

In his study, Ghulam Muhammad [30] proposed an innovative method for feature extraction, called interlaced derivative pattern (IDP), intended for automatic speech recognition (ASR) systems designed to be deployed on cloud servers. This technique employs Mel filter bank coefficients in relation to different neighborhood directions of the speech signal. The IDP-based ASR system demonstrates satisfactory performance even when the speech is transmitted through smartphones. The IDP exploits the relative Mel filter bank coefficients along with different neighborhood directions from the speech signal. The IDP-based ASR system performs reasonably well even when the speech is transmitted via smartphones. David Rybach et al. [31] announce the public availability of the RWTH Aachen University speech recognition toolkit. This toolkit includes state-of-the-art speech recognition technology for acoustic model training and decoding.

They published the toolkit under an open-source license, called “RWTH ASR License”. Chao Weng et al. [32] proposed a model training methodology based on the non-uniform Minimum Classification Error (MCE) approach. The main idea is to adapt the fundamental MCE criteria to reflect the cost-sensitive notion that errors on keywords are much more significant than errors on non-keywords in an automatic speech recognition task.

Philippe Dreuw and his team [33] propose an approach that begins with a large vocabulary speech recognition system in order to leverage the knowledge gained from ASR research. They concentrate on feature and model combination methods employed in ASR, as well as the use of pronunciation and language models (LM) in sign language recognition. These techniques can be applied to any sign language recognition system. They focus on feature and model combination techniques applied in ASR, and the usage of pronunciation and language models (LM) in sign language. These techniques can be used for all kinds of sign language recognition systems. Maria Schuster et al. [34] Proposed a new method, that was applied to recordings of a standard test to evaluate articulation disorders (psycholinguistic analysis of speech disorders of children PLAKSS) of 31 children at the age of 10. 1 ± 3. 8 years.

Mark D. Skowronski and John G. Harris [35] proposed the Echo State Network (ESN). The predictive ESN classifier combined an ESN with a competitive state machine framework. The

predictive ESN classifier group readout filters temporally into states, with filters in each state competing during training and testing using a winner-take-all scheme. Those classifiers made it an attractive classification engine for automatic speech recognition. Jonathan Dalby and Diane Kewley Port [36] described, provisionally named Pronto, which uses automatic speech recognition (ASR) for training pronunciation of second languages in adult learners. Pronto grows out of work in the Indiana Speech Training Aid (ISTRA) research program.

B. Various Approaches based on the Methods of Statistics

KartikAudhkhasi and colleagues [37] have emphasized the importance of having diverse or complementary ASR systems to achieve a reduction in word error rate (WER) when using the ROVER algorithm for fusion. They have provided theoretical evidence to explain the connection between ASR system diversity and ROVER performance, which is often observed in practice. Meanwhile, Diego Giuliani and team [38] have studied speaker adaptive acoustic modeling and introduced a new approach for speaker normalization, along with a commonly used method for vocal tract length normalization. Automatic speech recognition (ASR) systems is crucial for achieving a reduction in word error rate (WER) upon fusion using the ROVER algorithm. KartikAudhkhasi et al. [37] have presented theoretical proof explaining this often-observed link between ASR system diversity and ROVER performance. Diego Giuliani et al. [38] have investigated speaker adaptive acoustic modeling by using a novel method for speaker normalization and a well-known vocal tract length normalization method.

At run time, apply statistical processes to search through the space of all possible solutions, and pick the statistically most likely one. Naoki Hirayama et al. [39] have presented an automatic speech recognition (ASR) system that accepts a mixture of various kinds of dialects. The system recognized dialect utterances based on the statistical simulation of vocabulary transformation and combinations of several dialect models. Ángel de la Torre et al. [40] have described a method of compensating for nonlinear distortions in speech representation caused by noise. The described method was based on the histogram equalization method often used in digital image processing.

C. Different appraches for Artificial Intelligence system

Artificial Intelligence Approach in speech recognition can be the knowledge-based approach, which aims to automate the recognition process by emulating how humans use their intelligence to visualize, analyze and make decisions based on acoustic features. This approach often employs expert systems. In a study by Nelson Morgan [41], a series of research works conducted in the past decade on speech recognition utilizing discriminatively trained, feed-forward networks were reviewed, and it was concluded that while deep processing structures could potentially enhance the genre, careful selection of features and structure incorporation, such as layer width, were crucial for optimal results. Expert system is used widely in this approach. Nelson Morgan [41] has reviewed a line of research carried out over the last decade in speech recognition assisted by discriminatively trained, feed-forward networks and concluded that while the deep processing structures could provide improvements for this genre, the choice of features and the structure with which they are incorporated, including layer width.

Kun Han et al. [42] proposed to perform speech dereverberation using supervised learning, and the supervised approach was then extended to address both dereverberation and denoising. Deep neural networks were trained to directly learn spectral mapping from the magnitude spectrogram of corrupted speech to that of clean speech. The proposed approach substantially attenuated the distortion caused by reverberation, as well as background noise, and is conceptually simple. Heike Adel et al. [43] have presented their latest investigations on different features for factored language models for Code-Switching speech and their effect on automatic speech recognition (ASR) performance. They focused on syntactic and semantic features that have been extracted from Code-Switching text data and integrated them into factored language models.

Zakir Ali et al. [44] have developed Pashto Spoken Digits Database for Automatic Speech Recognition using Mel frequency cepstral coefficients (MFCC) for features extraction and Linear Discriminate Analysis (LDA) for features classification. Mel frequency cepstral coefficients were used to extract speech features. The features of speech were classified by the K nearest neighbor classifier and compared its accuracy with linear discriminate analysis. The experimental results were evaluated, and the overall average recognition exactitude of 76. 8 % was obtained. Arun Narayanan and DeLiang Wang [45] have presented a supervised speech separation system and automatic speech recognition (ASR) performance in realistic noise conditions has been improved. The system performed separation via ratio time-frequency masking; the ideal ratio mask (IRM) was estimated using DNNs (Deep Neural Networks).

Fred Richardson et al. [46] have presented the application of a single DNN (Deep Neural Network) for both SR (Speaker Recognition) and LR (Language Recognition) using the 2013 Domain Adaptation Challenge speaker recognition (DAC13) and the NIST 2011 language recognition valuation (LRE11) benchmarks.

D. Analysis based on different multi topic speech added

The study of noise robustness in speech recognition is a critical area of research. While humans can handle noise, ASR systems suffer from decreased performance in noisy environments. To tackle this challenge, researchers have investigated the application of psychoacoustic models to improve ASR system robustness. One such study by Peng Dai et al. [47] focused on implementing psychoacoustic models additively, which is a departure from the traditional subtractive approach used to remove noise. Another approach was proposed by Vikas Joshi et al. [48], who presented a novel sub-band-based Histogram Equalization (HEQ) framework for robust speech recognition. Their approach involved frequency band-specific equalization to compensate for noise distortion on individual frequency bands. These novel approaches show promising results for improving the robustness of ASR systems in noisy environments. Noise robustness has long been one of the most important goals in speech recognition. While the performance of automatic speech recognition (ASR) deteriorates in noisy situations, the human auditory system is relatively adept at handling noise. To mimic this adeptness, Peng Dai et al. [47] have studied and applied psychoacoustic models in speech recognition as a means to improve the robustness of ASR systems. Psychoacoustic models

were usually implemented in a subtractive manner to remove noise, hence presenting a novel algorithm that implements psychoacoustic models additively. Vikas Joshi et al. [48] have described a novel framework for sub-band based Histogram Equalization (HEQ) applied to robust speech recognition. They proposed a frequency band-specific equalization to compensate for the noise distortion on the individual frequency bands.

J. D. Echeverry-Correa et al. [49] have presented an efficient speech recognition approach for the multi-topic speech by combining information retrieval techniques and topic-based language modeling. Information retrieval-based techniques, such as topic identification using Latent Semantic Analysis, were used to identify the topic in a recognized transcription of an audio segment.

Andre Coy and Jon Barker [50] have presented a speech recognition system that couples these processes by using a combination of primitive and schema-driven processes: first, a set of coherent spectro-temporal fragments was generated by primitive segmentation techniques; then, a decoder based on statistical ASR techniques performs a simultaneous search for the correct background/foreground segmentation and word sequence hypothesis. J. Baker et al. [51] have focused on historically significant developments in the ASR area, including several major research efforts that were guided by different funding agencies, and suggest general areas in which to focus research.

III. Performance Analysis Based On Dataset, Preprocessing And Classification Techniques

Performance analysis can be conducted on different datasets, including clean and noisy datasets, to evaluate the performance of ASR systems in various scenarios. The analysis can also be conducted by comparing the performance of different ASR systems, including different approaches and models, to identify the best performing system for a given scenario. The ASR is one of the advantageous techniques used for the man-machine interface, in this paper we have reviewed some of so far research methodologies based on a different approach. Then in this section, the performance of various research methodologies is compared. Table 1 shows the performance of the various methodologies.

The performance analysis of some of the past research work in the field of the automatic speech recognition system is given in table 1. The analysis is made based on the different stages of ASR and corresponding techniques used in the stages, moreover, the performance measures like SNR, WER and Accuracy are considered in the past work and also analyzed.

Based on the performance analysis it is been observed that most of the speech recognition systems are not working for real time application. So, the research implementation in point view of real time input data and output data application is must.

Today's system gives more accuracy and efficiency but most of the application is speaker dependent such system gives more accuracy but beyond this researcher having opportunity to enter in the speaker independent system to be develops. As well if such system is real time application system then it will be applicable to many more application.

Below table also mentioned different classification and feature extraction methods classification methods are mostly prefer from start of the speech recognition are support vector machine, hidden markov model, Deep Neural Network, Deep Neural Network, and many more. The hidden markov model stayed and plays role in speech recognition for many years. Parallel to this support vector machine also become important part of speech recognition.

The MFCC, LPC are feature extraction techniques are been use currently to get more accuracy and efficiency.

TABLE I. Performance Analysis headings.

Author	**Techniques**			**Performance**			**Implementation**
	Pre-processing	**Features**	**Classifier**	**SNR**	**WER**	**Accuracy**	**Real Time**
Zhen-Hua Ling et al. [16]	Vocoder Analysis	MFCC & log-filterbank features	Deep Neural Network	□	□	□	□
Jesper Jensen and Zheng-Hua Tan [17]	-	MFCC & delta- and acceleration features	Short-Time Fourier Transform	□	□	□	□
Umit H. Yapanel and John H. L. Hansen [18]	Perceptual-MVDR	MFCC & Mel-scaled filterbank	Hidden Markov Model	□	□	□	□
Rongfeng Su et al. [19]	-	Gaussian mean, variance and mean transform	Hidden Markov Model	□	□	□	□
Javier Gonzalez-Dominguez et al. [20]	Multi Recognizer Module	Filter bank Energy	Deep Neural Network	□	□	□	□
MarijnHuijbregts et al. [21]	speech activity detection	twelve MFCC & one energy	Hidden Markov Model	□	□	□	□
Aitzol Ezeiza et al. [22]	Fractal Dimension	MFCC	Hidden Markov Model	□	□	□	□
Volker Leutnant et al. [23]	-	BAYESIAN feature	Stochastic Observation Models	□	□	□	□
TheologosAthanaselis et al. [24]	Knowledge Infrastructure	Text highlighting, line spacing,	Hidden Markov Model	□	□	□	□

	Component	adjusting fonts, segmenting words into its syllables					
Frank Wessel and Hermann Ney [25]	-	Acoustic feature	Hidden Markov Model	□	□	□	□
Octavian Cheng et al. [26]	Gaussian Mixture Model Accelerator	MFCC	Adaptive Pruning Algorithm	□	□	□	□
Yang Liu et al. [28]	-	prosodic and textual features	Hidden Markov Models	□	□	□	□
Kuldeep Kumar and R. K. Aggarwal [29]	-	MFCC	Hidden Markov Model Toolkit	□	□	□	□
Ghulam Muhammad [30]	pre-emphasized	MFCC, interlaced derivative pattern	hidden Markov model	□	□	□	□
David Rybach et al. [31]	-	Acoustic features	Hidden Markov Models	□	□	□	□
Chao Weng and B. H. F Juang [32]	-	MFCC	Hidden Markov Models	□	□	□	□
Philippe Dreuw et al. [33]	-	Composite Features; intensity & motion for visual	-	□	□	□	□
Maria Schuster et al. [34]	-	acoustic features, mean, standard deviation	-	□	□	□	□
Mark D. Skowronski and John G. Harris [35]	-	log energy of each frame and the first 12 cepstral coefficients	predictive echo state network	□	□	□	□

Jonathan Dalby and Diane Kewley-Port [36]	-	Phonetic Features	Hidden Markov Models	□	□	□	□
KartikAudhkhasi et al. [37]	-	MFCC	Deep Neural Network	□	□	□	□
Diego Giuliani et al. [38]	Speaker Normalization	MFCC & log-energy	Hidden Markov Model	□	□	□	□
Naoki Hirayama et al. [39]	Maximization Of Recognition Likelihoods	Acoustic And Linguistic Features	Dialect Language Model	□	□	□	□
Ángel de la Torre et al. [40]	Histogram Equalization	MFCC	Hidden Markov Model	□	□	□	□
Nelson Morgan [41]	Spectral Estimation	MFCC	Deep Belief Networks	□	□	□	□
Kun Han et al. [42]	Denoising,	MFCC features	Deep neural network	□	□	□	□
Heike Adel et al. [43]	-	open class words, Brown word clusters, POS tags	Recurrent Neural Network	□	□	□	□
Zakir Ali et al. [44]	split the audio of digits	MFCC	K nearest neighbor	□	□	□	□
Arun Narayanan et al. [45]	Denoising	MFCC, ratio masking	Deep Neural Network	□	□	□	□
Fred Richardson et al. [46]	-	MFCC, Bottleneck features, shifted delta cepstra	Deep Neural Network	□	□	□	□
Maria Schuster et al. [47]	acoustic analysis	Acoustic features	Hidden Markov Models	□	□	□	□
EnginAvci and ZuhtuHakan Akpolat [48]	Data acquisition Filtering White de-	wavelet packet decomposition	adaptive network based fuzzy inference	□	□	□	□

	noising		system				
Mark Gales and Steve Young [49]	Framing and Windowing	MFCC	Hidden Markov Model	□	□	□	□
Vikas Joshi et al. [50]	histogram analysis	MFCC	Deep Neural Network	□	□	□	□
J. D. Echeverry-Correa et al. [51]	Data Pre-processing	topic identification and dynamic language model adaptation	Generalized Vector Model	□	□	□	□
Andre Coy´ and Jon Barker [52]	Bregman's Auditory Scene Analysis	MFCC, DNE, SNE, SSP, SPD	Statistical ASR Technique	□	□	□	□
J. Baker et al. [53]	-	Acoustic features	-	□	□	□	□
Mark Gales and Steve Young et al. [54]	-	MFCC	Hidden Markov Model	□	□	□	□
J.Zhao,X.Mao,and L. Chen [55]	-	-	CNN and DBN with 4 LFLBs and with LSTM	□	□	□	□
AbhinavGarg, Gowtham P. Vadisetti, DhananjayaGowda, Sichen Jin [56]	-	-	Weighted finite state transducer (WFST)	□	□	□	□
J. Jorge, A. Giménez, J. A. Silvestre-Cerdà, J. Civera, A. Sanchis and A. Juan [57]	-	-	Deep Bidirectional Long-Short Term Memory (BLSTM)	□	□	□	□

IV. DATABASE DECRIPTION

EPPS Spanish Database: The EPPS Spanish database consists of speech data from 160 speakers (80 male and 80 female) from five different regions of Spain. The database includes a total of 20,000 utterances, with each speaker recording 125 utterances. The speech was recorded in a quiet environment and digitized at 16 kHz with 16-bit resolution.

AURORA-4 database: The AURORA-4 database is a noise-corrupted speech database consisting of 1000 utterances each of training, development, and evaluation sets. The training set is a mixture of clean and noisy speech, while the development and evaluation sets contain only noisy speech. The database includes four types of noises: car noise, babble noise, white noise, and factory noise.

TIMIT database: The TIMIT database consists of speech data from 630 speakers from eight different dialect regions of the United States. The database includes a total of 6300 utterances, with each speaker recording 10 utterances. The speech was recorded in a quiet environment and digitized at 16 kHz with 16-bit resolution. The TIMIT database is widely used as a benchmark for speech recognition systems. The TIMIT database is widely used for training and testing automatic speech recognition systems. It consists of 6300 phonetically balanced sentences spoken by 630 speakers, with each speaker contributing 10 sentences. The dataset is split into a training set of 4620 utterances and a test set of 1680 utterances, with each set containing equal numbers of male and female speakers. The speech samples are recorded at a sampling rate of 16 kHz and are quantized with 16-bit resolution. To simulate the effect of telephone channels, a version of the TIMIT database called NTIMIT was created, in which the speech data was transmitted through different telephone channels and re-digitized, resulting in a reduced bandwidth of 0.3-3.4 kHz.

SPEECON database: The SPEECON database consists of speech data from 120 speakers (60 male and 60 female) from seven different European countries. The database includes a total of 14,400 utterances, with each speaker recording 120 utterances. The speech was recorded in a quiet environment and digitized at 16 kHz with 16-bit resolution.

Databases such as SI84 database, EPPS Spanish Database, AURORA-4 database, TIMIT database, SPEECON database and MASS EFFECT 3 video game were used in the referred speech recognition systems as benchmark datasets. The description of the benchmark databases were given below.

SI84 database: The SI84 data (7077 utterances, or 15 hours of speech from 84 speakers) are used during the training phase. The training material is separated into a 6877-sentence training set and a 200-sentence validation set. The testing phase uses the Nov92 evaluation data, which contains 330 utterances from 8 speakers. The number of context-independent phonemes is 40.

EPPS Spanish Database: The EPPS Spanish Database (European Parliament Plenary Sessions), training set consists of 21127 sentences grouped in 1802 speaker turns. Development set consists of 2402 sentences grouped in 106 speaker turns.

AURORA-4 database: The AURORA-4 continuous speech recognition corpus, derived from the WallStreet Journal (WSJ0) corpus, has 14 test sets grouped into the following 4 groups: (a)Test set A - clean/multi-style speech in training and clean speech in test, same channel (set 1), (b) Test set B -clean/multi-style speech in training and noisy speech in test, same channel (sets 2-7), (c) Test set C - clean/multi-style speech in training and clean speech in test, different channel (set 8), and (d) Test set D -clean/multi-style speech in training and noisy speech in test, different channel (sets 9-14). The number inside the brackets represents the test set number defined in the AURORA-4 corpus.

MASS EFFECT 3 video game: The MASS EFFECT 3 video game consists of 20,000 speech recordings, around 500 roles, around 50 speakers, and around 20 hours of speech from professional actors. A subset of 4,000 speech recordings was used for the annotation of speech classes. Each speech recording was recorded in professional conditions and encoded into a 48 kHz-16 bits format. The duration of speech recording varies from 0.1s to 15s. Speech recordings shorter than 1s were removed from the speech database. It should be noted that the MASS EFFECT 3 dataset is not a standard benchmark dataset for speech recognition systems, but rather a specific dataset used for speech recognition in the context of a video game. Therefore, the performance analysis of speech recognition systems on this dataset may not be directly comparable to performance on other standard benchmark datasets. However, it can still provide valuable insights into the performance of speech recognition systems in real-world applications.

V. PERFORMANCE AND RESULT COMPARISON MATRIX FOR AUTOMATIC SPEECH RECOGNITION

The ASR system has variable results based on the front end, training and testing approach. Every testing methodology gives its prominent result to show that ASR system best among the other biomedical system. In this paper, we try to justify the result of recognition for various methodologies and feature extraction-based system which shows substantially noted outcomes for every recognition system.

The speech recognition system result varies according to languages for recognition like English, Marathi, Japanese, and French and so on.

The performance of an ASR system can vary depending on the language being recognized. Some languages are easier to recognize due to their phonetic structure and the availability of large training datasets, while others may be more challenging. Additionally, the performance of an ASR system can also be affected by the specific dialect or accent being used in the speech input. It is important to carefully evaluate the performance of an ASR system across different languages, dialects, and accents to ensure its generalizability and reliability in real-world applications.

VI. FUTURE DIRECTION

The ASR system is one of the essential technologies in the modern world, because of its wide application in many industries for man-machine interface. Hence the ASR system is still in research to find out a better technology so that the accuracy of the system has to enhance. In this paper we have reviewed some of the recent past research methodologies for the ASR system, from this review we suggest that research in ASR systems based on artificial intelligence techniques is engorged. So a system with better pre-processing, feature extraction and classification technique is a motivated research area in the future. The better pre-processing technique has to perform denoising, word separation, etc. In feature extraction, feature parameter selection is important, so the most possible feature has to extract and to reduce the complexity the feature section process has to penetrate using optimization techniques. Then finally the classification technique is the most important process, in recognition, so a novel or hybrid classifier for ASR system is engorged in the future.

It is true that the ASR system plays a crucial role in many industries, and research in this field is ongoing to improve its accuracy and performance. Artificial intelligence techniques, such as deep learning, have shown promising results in recent years and are being extensively used in ASR systems. Improving pre-processing techniques to handle noise and speech segmentation is also an active area of research.

Feature extraction is an important step in ASR systems, and selecting the appropriate features is critical for achieving high accuracy. Optimization techniques can be used to simplify this process and select the most relevant features. Finally, the classification technique used in ASR systems is essential for achieving high recognition accuracy, and research in developing novel or hybrid classifiers can further improve the performance of ASR systems.

In summary, future research in ASR systems should focus on developing better pre-processing techniques, optimizing feature selection, and developing novel or hybrid classifiers to achieve higher accuracy and performance.

VII. CONCLUSION

The review presented here provides an overview of recent techniques related to ASR based on different approaches, including the Acoustic Phonetic Approach, Pattern Recognition Approach, Statistics-based approach, and Artificial Intelligence Approach.

We also compared recent techniques of ASR based on pre-processing techniques, classification techniques, feature extraction techniques, and performance measures. We found that the performance of ASR systems is influenced by various factors, including the type of feature extraction techniques, the pre-processing techniques used, and the classification techniques employed.

The review also discussed several benchmark datasets used in the literature, including the SI84 database, EPPS Spanish Database, AURORA-4 database, TIMIT database, SPEECON database and MASS EFFECT 3 video game. The description of these datasets is important to understand the performance of ASR systems.

Finally, we discussed the future scope of ASR systems and suggested that research on ASR systems based on artificial intelligence techniques is engorged. We proposed that a system with better pre-processing, feature extraction, and classification techniques is a motivated research area in the future. We also suggested that a novel or hybrid classifier for ASR systems can enhance the performance of ASR systems in the future.

It is important to note that the automatic speech recognition system is an ongoing research area with a lot of potential for future advancements. By analyzing the current state of the art techniques and benchmark datasets, this review provides a foundation for future research directions in ASR. Here presented a detailed review of recent techniques related to the automatic speech recognition system. Initially, we presented the analysis of ASR technique based on different approaches like the Acoustic Phonetic Approach, Pattern Recognition Approach, Statistics-based approach and Artificial Intelligence Approach. Then a comparison of recent techniques of ASR is presented based on the factors such as pre-processing techniques, classification techniques, feature extraction techniques and performance measures. The benchmark datasets used in the literature and their details are presented subsequently we presented the future scope of the ASR by analyzing the literature used in this review.

REFERENCES

1. M Siafarikas, T. Ganchev and N. Fakotakis, "Wavelet packet based speaker verification", In ODYSSEY04-The Speaker and Language Recognition Workshop, pp. 257-264, 2004.
2. E. Avci and Z.H Akpolat, "Speech recognition using a wavelet packet adaptive network based fuzzy inference system", Expert Systems with Applications, Vol. 31, No. 3, pp. 495-503, 2006.
3. J. Manikandan, B. Venkataramani, K. Girish, H. Karthic and V. Siddharth, "Hardware implementation of real-time speech recognition system using TMS320C6713 DSP", In Proceedings of IEEE International Conference on VLSI Design, pp. 250-255, 2011.
4. Waibel, A. Badran, A.W. Black, R. Frederking, D. Gates, A. Lavie, L. Levin, K. Lenzo, L. Mayfield Tomokiyo, J. Reichert, T. Schultz, D. Wallace, M. Woszczyna and J. Zhang, "Speechalator: Two-way speech-to-speech translation in your hand", In Proceedings of Conference of North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol. 4, pp. 29-30, 2003.
5. Yuan, T. Lee, P.C. Ching and Y. Zhu, "Speech recognition on DSP: issues on computational efficiency and performance analysis", Microprocessors and Microsystems, Vol. 30, No. 3, pp. 155-164, 2006
6. Delaney, N. Jayant, M. Hans, T. Simunic and A. Acquaviva, "A low-power, fixed-point, front-end feature extraction for a distributed speech recognition system", In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, pp. I-793 - I-796, 2002.
7. W. Han, K.W. Hon, C.F. Chan, T. Lee, C.S. Choy, K.P. Pun and P.C. Ching, "An HMM-based speech recognition IC", In Proceedings of IEEE International Symposium on Circuits and Systems, Vol. 2, pp. 744-747, 2003.
8. Nadeu, D. Macho and J. Hernando, "Time and frequency filtering of filter-bank energies for robust HMM speech recognition", Speech Communication, Vol. 34, No. 1, pp. 93-114, 2001

9. S. Yoshizawa, N. Wada, N. Hayasaka and Y. Miyanaga, "Scalable architecture for word HMM-based speech recognition and VLSI implementation in complete system", IEEE Transactions on Circuits and Systems I: Regular Papers, Vol. 53, No. 1, pp. 70-77, 2006.
10. W. Han, K.W. Hon, C.F. Chan, T. Lee, C.S. Choy, K.P. Pun and P.C. Ching, "An HMM-based speech recognition IC", In Proceedings of IEEE International Symposium on Circuits and Systems, ISCAS'03, Vol. 2, pp. 744-747, 2003.
11. S.J. Melnikoff, S.F. Quigley and M.J. Russell, "Implementing a simple continuous speech recognition system on an FPGA", In Proceedings of IEEE Annual Symposium on Field-Programmable Custom Computing Machines, pp. 275-276, 2002
12. J.J. Rodríguez-Andina, R.D.R. Fagundes and D.B. Júnior, "A FPGA-based Viterbi algorithm implementation for speech recognition systems", In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 1217-1220, 2001.
13. Huggins-Daines, M. Kumar, A. Chan, A.W. Black, M. Ravishankar and A.I. Rudnicky, "Pocket sphinx: A free, real-time continuous speech recognition system for hand-held devices", In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, pp. 185-188, 2006.
14. C. Donkin, S.D. Brown and A. Heathcote, "Choice Key: A real-time speech recognition program for psychology experiments with a small response set", Behavior research methods, Vol. 41, No. 1, pp. 154-162, 2009.
15. V. Tyagi and C. Wellekens, "On desensitizing the Mel-Cepstrum to spurious spectral components for Robust Speech Recognition", In Proceedings of IEEE International Conference on ICASSP, Vol. 1, pp. 529-532, 2005.
16. Zhen-Hua Ling, Shi-yin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen Meng, and Li Deng, "Deep Learning for Acoustic Modeling in Parametric Speech Generation", IEEE Signal Processing Magazine, Vol. 32, No. 3, pp. 35-52, 2015.
17. Jesper Jensen and Zheng-Hua Tan, "Minimum Mean-Square Error Estimation of Mel-Frequency Cepstral Features-A Theoretically Consistent Approach", IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 23, No. 1, pp. 186-197, 2015.
18. Umit H. Yapanel and John H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition", Speech Communication, Vol. 50, pp. 142–152, 2008
19. Rongfeng Su, Xunying Liu and Lan Wang, "Automatic Complexity Control of Generalized Variable Parameter HMMs for Noise Robust Speech Recognition", IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 23, no, 1, pp. 102-114, 2015.
20. Javier Gonzalez-Dominguez, David Eustis, Ignacio Lopez-Moreno, Andrew Senior, Françoise Beaufays, and Pedro J. Moreno, "A Real-Time End-to-End Multilingual Speech Recognition Architecture", IEEE Journal of Selected Topics In Signal Processing, Vol. 9, No. 4, pp. 749-759, 2015.
21. MarijnHuijbregts, Roeland Ordelman, and Franciska de Jong, "Annotation of Heterogeneous Multimedia Content Using Automatic Speech Recognition", Semantic Multimedia, Lecture Notes in Computer Science, Vol. 4816, pp. 78-90, 2007
22. AitzolEzeiza, KarmeleLópez de Ipiña, Carmen Hernández and Nora Barroso, "Enhancing the Feature Extraction Process for Automatic Speech Recognition with Fractal Dimensions", Cognitive Computation, Vol. 5, No. 4, pp. 545-550, 2013
23. Volker Leutnant, Alexander Krueger, and Reinhold Haeb-Umbach, "A New Observation Model in the Logarithmic Mel Power Spectral Domain for the Automatic Recognition of

Noisy Reverberant Speech", IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 22, No. 1, pp. 95-104, 2014.

24. TheologosAthanaselis, SteliosBakamidis, IoannisDologlou, Evmorfia N. Argyriou and AntonisSymvonis, "Making assistive reading tools user friendly: a new platform for Greek dyslexic students empowered by automatic speech recognition", Multimedia tools and applications, Vol. 68, No. 3, pp. 681-699, 2014.
25. Frank Wessel and Hermann Ney, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition", IEEE Transactions on Speech and Audio Processing, Vol. 13, No. 1, pp. 23-31, 2005.
26. Octavian Cheng, Waleed Abdulla, and ZoranSalcic, "Hardware–Software Codesign of Automatic Speech Recognition System for Embedded Real-Time Applications", IEEE Transactions on Industrial Electronics, Vol. 58, No. 3, pp. 850-859, 2011
27. Bishnu S. Atal, and Lawrence R. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced Classification with Application to Speech Recognition", in proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing, Pennsylvania, Vol. 24, No. 3, pp. 201-212, 1976.
28. Yang Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 5, pp. 1526-1540, 2006.
29. Kuldeep Kumar and R.K. Aggarwal "Hindi speech recognition system using HTK", International Journal of Computing and Business Research, Vol. 2, pp. 2229-6166,2011.
30. Ghulam Muhammad, "Automatic speech recognition using interlaced derivative pattern for cloud based healthcare system", Cluster Computing, Vol. 18, No. 2, pp. pp. 795-802,2015.
31. David Rybach, Christian Gollan, Georg Heigold, Bjorn Hoffmeister, Jonas Loof, Ralf Schluter and Hermann Ney , "The RWTH Aachen University Open Source Speech Recognition System", In Interspeech, pp. 2111-2114, 2009.
32. Chao Weng and B.H.F Juang, "Discriminative Training Using Non-Uniform Criteria for Keyword Spotting on Spontaneous Speech", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 2, pp. 300-312, 2015.
33. Philippe Dreuw, David Rybach, Thomas Deselaers, MortezaZahedi, and Hermann Ney, "Speech Recognition Techniques for a Sign Language Recognition System", Hand, Vol. 60, pp. 80-83, 2007.
34. Maria Schuster, Andreas Maier, TinoHaderlein, EmekaNkenke, Ulrike Wohlleben, Frank Rosanowski, Ulrich Eysholdt and ElmarNoth, "Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition", International Journal of Pediatric Otorhinolaryngology, Vol. 70, No. 10, pp. 1741-1747,2006.
35. Mark D. Skowronski and John G. Harris, "Automatic speech recognition using a predictive echo state network classifier", Neural Networks , Vol. 20, No. 3, pp. 414-423, 2007.
36. Jonathan Dalby and Diane Kewley-Port, "Explicit Pronunciation Training Using Automatic Speech Recognition Technology", CALICO Journal, Vol. 16, No. 3, pp. 425-445, 1999.
37. KartikAudhkhasi, Andreas M. Zavou, Panayiotis G. Georgiou, and Shrikanth S. Narayanan, "Theoretical Analysis of Diversity in an Ensemble of Automatic Speech Recognition Systems", IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 22, No. 3, pp. 711-726, 2014.

38. Diego Giuliani, MatteoGerosa and Fabio Brugnara, "Improved automatic speech recognition through speaker normalization", Computer Speech and Language, Vol. 20, pp. 107–123, 2006.
39. Umbarkar, A. M., Sherie, N. P., Agrawal, S. A., Kharche, P. P., & Dhabliya, D. (2021). Robust design of optimal location analysis for piezoelectric sensor in a cantilever beam. Materials Today: Proceedings, doi:10.1016/j.matpr.2020.12.1058
40. Vadivu, N. S., Gupta, G., Naveed, Q. N., Rasheed, T., Singh, S. K., & Dhabliya, D. (2022). Correlation-based mutual information model for analysis of lung cancer CT image. BioMed Research International, 2022, 6451770. doi:10.1155/2022/6451770
41. Veeraiah, D., Mohanty, R., Kundu, S., Dhabliya, D., Tiwari, M., Jamal, S. S., & Halifa, A. (2022). Detection of malicious cloud bandwidth consumption in cloud computing using machine learning techniques. Computational Intelligence and Neuroscience, 2022 doi:10.1155/2022/4003403
42. Naoki Hirayama, Koichiro Yoshino, KatsutoshiItoyama, Shinsuke Mori, and Hiroshi G. Okuno, "Automatic Speech Recognition for Mixed Dialect Utterances by Mixing Dialect Language Models", Journal OF IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 2, pp. 373 - 382, 2015.
43. Ángel de la Torre, Antonio M. Peinado, José C. Segura, José L. Pérez-Córdoba, Ma Carmen Benítez, and Antonio J. Rubio, "Histogram Equalization of Speech Representation for Robust Speech Recognition", IEEE Transactions on Speech and Audio Processing, Vol. 13, No. 3, pp. 355-366, 2005.
44. Nelson Morgan, "Deep and Wide: Multiple Layers in Automatic Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 1, pp. 7-13, 2012.
45. Kun Han, Yuxuan Wang, De Liang Wang, William S. Woods, Ivo Merks and Tao Zhang, "Learning Spectral Mapping for Speech Dereverberation and Denoising", IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 23, No. 6, pp. 982-992, 2015.
46. Heike Adel, Ngoc Thang Vu, Katrin Kirchhoff, Dominic Telaar and Tanja Schultz, "Syntactic and Semantic Features For Code-Switching Factored Language Models", IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 23, No. 3, pp. 431-440 , 2015.
47. Zakir Ali, ArbabWaseem Abbas, T.M. Thasleema , BurhanUddin, TanzeelaRaaz, SahibzadaAbdur and RehmanAbid, "Database development and automatic speech recognition of isolated Pashto spoken digits using MFCC and K-NN", International Journal of Speech Technology, Vol. 18, No. 2, pp. 271-275, 2015
48. 45.Arun Narayanan and DeLiang Wang, "Improving Robustness of Deep Neural Network Acoustic Models via Speech Separation and Joint Adaptive Training", IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 23, No. 1, pp. 92-101, 2015.
49. Fred Richardson, Douglas Reynolds, and NajimDehak, "Deep Neural Network Approaches to Speaker and Language Recognition", IEEE Signal Processing Letters, Vol. 22, No. 10, pp. 1671-1675, 2015.
50. Peng Dai, Frank Rudzicz, IngYann Soon, Alex Mihailidis, Huijun Ding, "2D Psychoacoustic modeling of equivalent masking for automatic speech recognition", Signal Processing, Vol. 115, pp. 9-19, 2015.

51. Vikas Joshi, RaghvendraBilgi, S. Umesh, Luz Garcia and Carmen Benitez, "Sub-band based histogram equalization in cepstral domain for speech recognition", Speech Communication, Vol. 69, pp. 46-65, 2015.
52. J..D. Echeverry-Correa, J. Ferreiros-López, A. Coucheiro-Limeres, R. Córdoba and J.M. Montero, "Topic identification techniques applied to dynamic language model adaptation for automatic speech recognition", Expert Systems with Applications, Vol. 42, pp. 101–112, 2015.
53. Andre Coy´ and Jon Barker, "An automatic speech recognition system based on the scene analysis account of auditory perception", Speech Communication, Vol. 49, No. 5, pp. 384-401, 2007
54. J. Baker, Li Deng, J. Glass, S. Khudanpur, Chin-hui Lee, N. Morgan and D. O'Shaughnessy, "Developments and directions in speech recognition and understanding, Part 1 [DSP Education]", IEEE Signal Processing Magazine, Vol. 26, No. 3, pp. 75-80, 2009.
55. Mark Gales and Steve Young, "The Application of Hidden Markov Models in Speech Recognition", Signal Processing, Vol. 1, No. 3, pp. 195–304, 2007.
56. Baker, Li Deng, J. Glass, S. Khudanpur, Chin-hui Lee, N. Morgan and D. O'Shaughnessy, "Developments and directions in speech recognition and understanding, Part 1 [DSP Education]", IEEE Signal Processing Magazine, Vol. 26, No. 3, pp. 75-80, 2009.
57. Mark Gales and Steve Young, "The Application of Hidden Markov Models in Speech Recognition", Signal Processing, Vol. 1, No. 3, pp. 195–304, 2007.
58. J.Zhao,X.Mao,and L. Chen, ‘‘Speech emotion recognition using deep1D & 2D CNN LSTM networks,’’ Biomed. Signal Process. Control,vol. 47, pp. 312–323, Jan. 2019.
59. A.Garg, G. P. Vadisetti, D. Gowda, S. Jin, A. Jayasimha, Y. Han, J. Kim, J. Park, K .Kim, S. Kim, Y. yoon Lee, K. Min, and C. Kim, “Streaming On-Device End-to-End ASR System for Privacy-Sensitive Voice-Typing, ”in Proc. Inter speech 2020, 2020, pp. 3371–3375.
60. J. Jorge, A. Giménez, J. A. Silvestre-Cerdà, J. Civera, A. Sanchis and A. Juan, "Live Streaming Speech Recognition Using Deep Bidirectional LSTM Acoustic Models and Interpolated Language Models," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 148-161, 2022, doi: 10.1109/TASLP.2021.3133216.