IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, January 2026
Impact Factor: 7.67

An Analysis of Current Methods and Architectures in Automatic Speech Recognition

Ram Chandra Singh[1] and Dr. Sher Jung[2]
[1]Research Scholar, Department of Computer Science and Engineering
[2]Professor, Department of Computer Science and Engineering
Sunrise University, Alwar, Rajasthan

Abstract: *Speech is our most natural form of communication. voice recognition converts voice into words using a computer software. Speech recognition apps make using speech as an input modality easy and effective. Native language speech recognition interfaces will allow illiterate/semi-literate individuals to utilize technology more without a keyboard or stylus. Speech recognition and its applications were studied for over 30 years. Many devices use automated voice recognition to communicate between humans and machines. Reverberation and even mild ambient noise degrade speech recognition. Robustness to noise, reverberation, and transducer properties is an unresolved challenge that keeps voice recognition research active. This article examines automated speech recognition architecture, speech parameterization, techniques, characteristics, challenges, databases, tools, and applications.*

Keywords: Automatic Speech Recognition, Acoustic Model, Language Model.

I. INTRODUCTION

One of the most fascinating signal processing research topics is voice processing, the most natural form of human communication.voice processing examines voice signals and their processing. Signals are frequently treated digitally, hence speech processing is a specific instance of digital signal processing. This unique field covers several technology and applications. Speech processing is largely helpful in daily life. Speech processing includes Speech Coding, Text-to-Speech Synthesis, Speech Recognition, Speaker Recognition and Verification, Speech Enhancement, Speech Segmentation and Labeling (Transcription), Language Identification, Prosody, Attitude and Emotion Recognition, Audio-Visual Signal Processing, and Spoken Dialog Systems.

One of the main study topics in voice processing is Automatic voice Recognition (ASR). It uses a computer software algorithm to transform a vocal signal to text.

Probability Theory of Speech Recognition

The primary goal of an ASR system is to hypothesize the most likely discrete symbol sequence out of all valid sequences in the language L, from the given acoustic input O .As stated above, the input is treated as a set of discrete observations, such that:

$$O = o_1, o_2, o_3, \dots, o_t \quad (1)$$

Similarly, the symbol sequence to be recognized is defined as:

$$W = w_1, w_2, w_3, \dots, w_n \quad (2)$$

The fundamental ASR system goal can then be expressed as:

$$\hat{W} = \operatorname{argmax} P(W|O) \quad for\ W \in \mathcal{L} \quad (3)$$

This equation implies that for a given sequence W and acoustic input sequence O, the probability P(W|O) needs to be determined. Bayes' theorem can be applied to this probability to arrive at the following equation:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, January 2026
Impact Factor: 7.67

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \quad (4)$$

The quantities on the right-hand side of the equation are easier to compute than P(W|O). P(W) is defined as the prior probability for the sequence itself. This is calculated by using the prior knowledge of occurrences of the sequence W. Since the P(O) is the same for each candidate sentence W, thus equation 4 can be simplified as

$$\hat{W} = \text{argmax}\frac{P(O|W)P(W)}{P(O)} = \text{argmax}\, P(O|W)P(W) \qquad for\ W \in \mathcal{L} \quad (5)$$

The probability P(O|W), which is the likelihood of the acoustic input O, given the sequence W, is defined as the observation likelihood, can be called as acoustic score. This quantity can be determined using the Hidden Markov Model.

Speech Recognition Architecture

Figure 1 shows the main components of a voice recognition system: acoustic front-end, model, lexicon, language model, and decoder. Acoustic front-end converts voice signal into recognition-relevant characteristics. A microphone's audio waveform is turned into a series of fixed-size acoustic vectors during feature extraction. Word/phone model parameters are obtained using training data acoustic vectors. The decoder searches all potential word sequences for the most probable to create. Language models determine P (W) and acoustic models P (O|W) indicate probability. Automatic speech recognition systems collect speech parameters from acoustic voice signals for each word or sub-word unit. The speech characteristics' fluctuation over time creates a pattern that defines the word or sub-word. Operators read all application vocabulary throughout training. When recognizing a word, its pattern is compared to the stored patterns and the best match is chosen. This method is called pattern recognition.

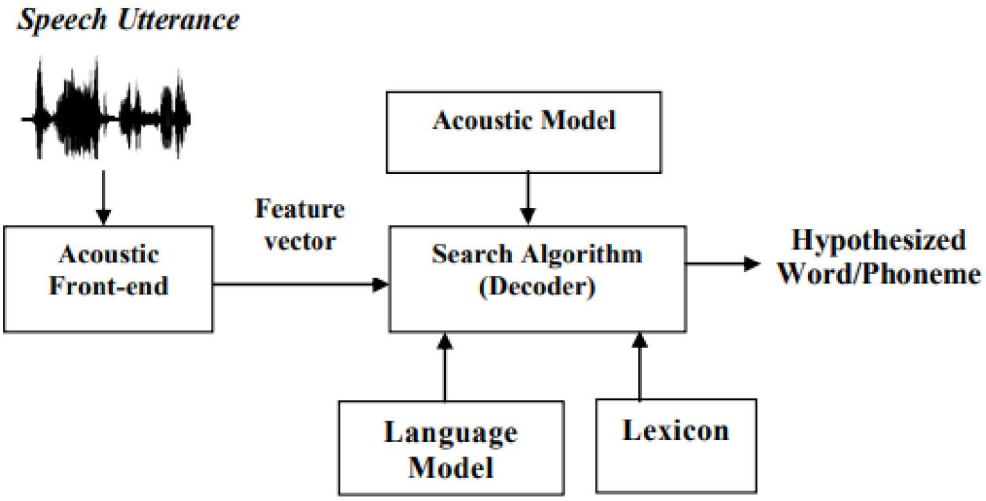

Figure 1. Speech Recognition Architecture

Acoustic Front-end

Acoustic front-end contains signal processing and feature extraction. The feature extraction process in voice recognition is to construct a compact series of feature vectors that reflect the input signal [4].

The feature extraction process normally has three steps. Acoustic front end (speech analysis) is the initial step. It conducts spectra temporal analysis of the data and creates raw characteristics of brief speech interval power spectrum envelopes. An enhanced static-dynamic feature vector is compiled in the second step. The last step (which is not usually present) compacts and robustizes these expanded feature vectors for the recognizer. No single feature is best for every application, but the features should allow an automatic system to distinguish between similar speech sounds, allow the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, January 2026 Impact Factor: 7.67

automatic creation of acoustic models for these sounds without a lot of training data, and exhibit statistics that are mostly invariant across speakers and spea Mechanisms for reducing each audio signal segment's information into a limited number of characteristics, or features, are needed to locate statistically meaningful information in incoming data.

These attributes should characterize each segment so comparable segments may be grouped by comparison. There are many fascinating and unique methods to define speech signal characteristics. PCA, LDA, ICA, LPC, Cepstral Analysis, Mel-Frequency Scale Analysis, Filter-Bank Analysis, MFCC, Kernal Based Feature Extraction, Dynamic Feature Extraction, Wavelet-based features, Spectral Subtraction, and CMS are feature extraction methods. Many auditory-based feature extraction methods are used in noise-robust speech recognition, including ZCPA, ALSD, PMVDR, PNCC, IIF, amplitude modulation spectrogram, Gammatone frequency cepstral coefficients, SPARK, and Gabor filter bank features. Many feature representations are used, but the MFCC feature set is the most used. Below are the stages of MFCC feature extraction, and figure shows the procedure.

2. **Pre-emphasis –** This stage is used to amplify energy in the high-frequencies of the input speech signal. This allows information in these regions to be more recognizable during HMM model training and recognition.

Windowing – This stage slices the input signal into discrete time segments.

This is done by using a window of N milliseconds wide and at offsets of M milliseconds long. A Hamming window is commonly used to prevent edge effects associated with the sharp changes in a rectangular window.

Discrete Fourier Transform – DFT is applied to the windowed speech signal, resulting in the magnitude and phase representation of the signal.

Mel Filter Bank – While the resulting spectrum of the DFT contains informationin each frequency, human hearing is less sensitive at frequencies above 1000 Hz. This concept also has a direct effect on performance of ASR systems; therefore, the spectrum is warped using a logarithmic Mel scale. A Mel frequency can be computed using equation 6. In order to create this effect on the DFT spectrum, a bank of filters known as triangular filters is constructed with filters distributed equally below 1000 Hz and spaced logarithmically above 1000 Hz. The output of filtering the DFT signal by each Mel filter is known as the Mel spectrum.

$$mel(f) = 1127 \ln(1 + \frac{f}{700}) \quad (6)$$

Log –Taking logarithm of this provides Mel spectrum co-efficients.

DCT –The final step in obtaining MFCC is performing discrete cosine transformon the Mel spectrum co-efficients. The output of DCT is Mel-cepstral coefficients of 13th order.

Delta MFCC Features – In order to capture the changes in speech from frameto-frame, the first and second derivative of the MFCC coefficients are also calculated and used .

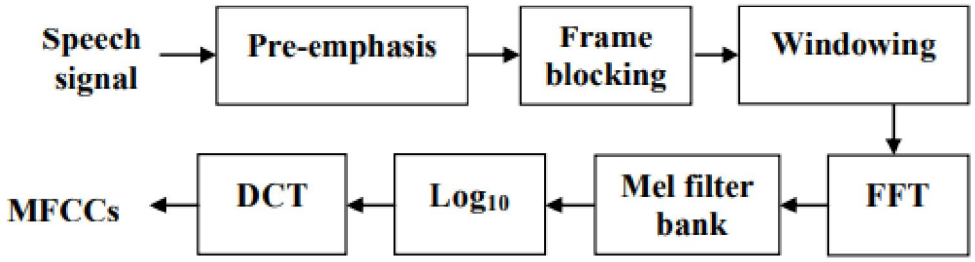

Figure 2. MFCC Feature Extraction

Acoustic Model

Acoustic model is a key knowledge source for automated speech recognition systems, representing phonetic unit acoustic properties. Selecting basic modeling units is crucial to developing an acoustic model. When the speech's target

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, January 2026 Impact Factor: 7.67

language is known, numerous sub-word units may be employed for acoustic modeling. Different acoustic modeling units may greatly impact speech recognition ability.

Acoustic speech modeling involves statistically representing speech waveform feature vector sequences. Hidden Markov Model (HMM) is a popular statistical model for acoustic modeling. Segmental, supersegmental, neural network, maximum entropy, and (hidden) conditional random fields are further auditory models. Acoustic models are files with statistical representations of each word sound.

Each statistical representation has a phoneme label. Acoustic models are developed by applying specific training methods to construct statistical representations for each phoneme in a language from a huge voice corpus. Every phoneme has an HMM. After listening for user noises, the speech decoder finds an HMM in the acoustic model. Each word w is broken down into base phones. Given a base phone, the acoustic model predicts a certain observation.

Language Model

A language model describes the allowable word order in a language. These restrictions may be expressed by generative grammar rules or word pair statistics computed on a training corpus. Although certain words sound like phones, humans can usually distinguish them. This is because they know the context and can guess what words or phrases could be in it. Language models provide voice recognition systems context. The language model defines valid words and their order.

Language models are trained by monitoring word sequences in corpora of text with millions of word tokens and lowering perplexity on training data to estimate n-gram probabilities. However, reducing confusion does not always improve speech recognition. Algorithms that enhance language models based on voice recognition are especially interesting for a language model that describes the probability distribution of words the speaker may utter next given a history of words. Common language models are bigram and trigram. These models calculate probability of two or three word clusters in a sequence. CMU SLM Toolkit and Stanford Research Institute Language Modeling Toolkit are language modeling tools.

Decoder

Decoding involves finding the most probable word sequence W given the observation sequence O and the acoustic-phonetic-language model. Decoding may be addressed via dynamic programming methods. Instead of analyzing likelihoods of all model routes producing O, discover the network path that best matches O.

The Viterbi method is often used to predict the optimum state sequence for an observation series. Larger vocabulary recognition challenges make it difficult to examine all potential words during the Viterbi algorithm's recursive stage.

A beam search may be employed for Viterbi iteration, extending routes to the next time step only for words with path probabilities over a threshold. This method speeds up searches but reduces decoding accuracy. Viterbi method implies best pathways at time t are extensions of best paths ending at time $t - 1$, which is not always true. The sequence's optimum route may be the least likely at first (e.g., the most probable phoneme sequence does not need to match the most probable word sequence). Extended Viterbi and forward-backward algorithms solve this.

Speech Recognition Methodologies

ASR methodologies are broadly classified into three approaches, namely, acousticphonetic approach, pattern-recognition approach and artificial intelligence approach.

Acoustic-Phonetic Approach: It is based on acoustic phonetics that postulates that there exist finite, distinctive phonetic units in spoken language. The phonetic units are characterized by a set of acoustic properties that are manifested in the speech signal, or its spectrum, over time. The first step in the acoustic phonetic approach is a spectral analysis of the speech combined with a feature detection that converts the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. The next step is a segmentation and labeling phase in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech. The last step in this

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, January 2026 Impact Factor: 7.67

approach attempts to determine a valid word (or string of words) from the phonetic label sequences produced by the segmentation to labeling .

Pattern Recognition Approach: The pattern-matching approach involves two essential steps namely, pattern training and pattern comparison. In the pattern-comparison stage of the approach, a direct comparison is made between the unknown speeches with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of match of the patterns. The essential feature of this approach is that it uses a well-formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm. A speech pattern representation can be in the form of a speech template or a statistical model (e.g., Hidden Markov Model) and can be applied to a sound smaller than a word, a word, or a phrase. The pattern-matching approach has become the predominant method for speech recognition in the last six decades.

Artificial Intelligence Approach: The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. In this, it exploits the ideas and concepts of acoustic phonetic and pattern recognition methods. The main methodologies that made significant change in the speech recognition area are elaborated below. Two main approaches to pattern matching have been widely used in

ASR – deterministic pattern matching based on dynamic time warping (DTW) [9], and stochastic pattern matching employing hidden Markov models (HMMs) . In DTW, each class to be recognized is represented by one or several templates. Using more than one reference template per class may be preferable in order to improve the pronunciation/speaker variability modeling. During recognition, a distance between an observed speech sequence and class patterns is calculated. To eliminate the impact of the duration mismatch between test and reference patterns, stretched and warped versions of the reference patterns are also employed in the distance calculation. The recognized word corresponds to the path through the model that minimizes the accumulated distance. Increasing the number of class pattern variants and loosening warping constrains may improve DTW-based recognition performance at the expense of storage space and computational demands. In state of the art systems, HMM-based pattern matching is preferred instead of DTW due to better generalization properties and lower memory requirements.

Generative Learning Approach -HMM-GMM

The most prevalent generative learning method in ASR is Gaussian-MixtureModel-based Hidden Markov models. Speech signals are represented sequentially by Gaussian mixture model (GMM)-based hidden Markov models (HMMs) in conventional speech recognition systems. Speech signals may be piecewise stationary or short-time stationary, hence HMMs are utilized in speech recognition. Speech seems stationary at short timescales. Many stochastic applications use speech as a Markov model. A Gaussian mixture is used in each HMM stage to simulate the sound wave's spectral representation. The parameters for a GMM-HMM are: ( ) - state prior probabilities vector; A=(aij) - state transition probability matrix; B={(b1,…bn}) - Gaussian mixture model of state j. Speech usually associates the state with a phone part.

Modern systems perform well with concealed markov models. HMMs are popular because they can manage variable-length data sequences from word sequence, speaker rate, and accent. Even though HMM-GMM is the norm in ASR, it has pros and cons. Automatically trained HMM-based voice recognition systems are simple and computationally viable. Gaussian mixture models are statistically inefficient for modeling data on or near a non-linear manifold.

Discriminative Learning – HMM-ANN

Using a discriminative model or training a generative model called discriminative learning. In the 1990s, Multilayer Perceptron (MLP) neural networks with softmax nonlinear functions at the last layer were popular. MLP output may be understood as conditional probability [12], therefore feeding it into an HMM creates a good discriminative sequence model, or hybrid MLP-HMM. The challenge of learning MLPs has led to a shift in study, focusing on combining MLPs with classical features for usage in generative HMMs [13]. Back-propagation error derivative-trained neural networks became an interesting acoustic modeling tool for voice recognition in the late 1980s. Neural networks assume no feature statistical qualities, unlike HMMs. When used to estimate voice feature segment probabilities, neural networks provide natural and efficient discriminative training.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, January 2026 Impact Factor: 7.67

Neural networks are good at categorizing short-term units like phones and single words, but they struggle with continuous recognition tasks [14-15] due to their inability to understand temporal connections. Shallow architectures may solve basic or well-constrained issues, but their limited modeling and representational capacity might make it challenging to employ them in more complex real-world speech applications. Using neural networks for feature modification and dimensionality reduction before HMM-based recognition is one option.

Deep Learning -HMM DNN

Deep learning, also known as representation learning or unsupervised feature learning, is new to machine learning. Deep learning has replaced Gaussian mixtures for voice recognition and feature coding on a broad scale. First, generative deep architectures specify high-order correlation features or joint statistical distributions of observable data and their classes. Bayes rule can make this design discriminative. Deep auto-encoders, deep Boltzmann machines, sum-product networks, the original Deep Belief Network (DBN), and its expansion to the factored higher-order Boltzmann machine in its bottom layer are examples.

Second, discriminative deep architectures characterize the posterior distributions of class labels conditioned on visible data to give pattern classification capacity. Deep-structured CRF, tandem-MLP, deep convex or stacking network, tensor version, and detection-based ASR are examples. The third form, hybrid deep architectures, uses generative architectures to discriminate. The hybrid architecture's end purpose is discrimination, hence the generative component is largely used.

Characteristics of Speech Recognition Systems

Automatic speech recognition systems are designed to solve a particular problem. There are numerous parameters that affect the design of the ASR system. There are number of issues that need to be addressed in order to define the operating range of each speech recognizing systems that is built. Some of them are, modeling units like word, syllable, phoneme used for recognition, vocabulary size like small, medium and large, task syntax like simple to complex task using N-gram language models, task perplexity, speaking mode like isolated, connected, continuous, spontaneous, speaker mode like speaker trained, adaptive, speaker independent, dependent, speaking environment as quiet room, noisy places, transducers may be high quality microphone, telephones, cell phones, array microphones, and also transmission channel.

Challenges in Speech Recognition

Robustness of an ASR system is the system's ability to successfully deal with different aspects of variability in the speech signal. There are a number of well-known factors that determine the accuracy of a speech-recognition system. The most noticeable ones are speaker variability, pronunciation variability, region variability, speech rate variability, context variability, channel variability and environment variability. In the design of speech recognition systems, these challenging factors must be considered and effective models to be created to provide good recognition accuracy irrespective of these variabilities [20]. In higher level, speech recognition system design requires the availability of algorithms or processes for automatic generation of word lexicons, automatic generation of language models for new tasks, automatic speech segmentation algorithms, optimal utterance verification-rejection algorithm, achieving or surpassing human performance on ASR tasks.

Applications of Speech Recognition

More recently, with the exponential growth of big data and computing power, ASR technology has advanced to the stage where more challenging applications are becoming a reality. Examples are voice search and interactions with mobile devices (e.g., Siri on iPhone, Bing voice search on winPhone, and Google Now on Andriod), voice control in home entertainment systems (e.g., Kinect on xBox), and various speech-centric information processing applications capitalizing on downstream processing of ASR outputs [21]. Some of these typical applications include dictation systems, voice user interfaces, voice dialling, call routing, domestic appliance control, command and control, voice enabled search, simple data entry, hands and eyes free applications and learning system for disabled people.

Databases and Tools for Speech Recognition

Speech Databases

Many American and European speech datasets are accessible for automated speech recognition research. TIMIT, GlobalPhone, Aurora, Wall Street Journal, AN4, TI Digits, TI46, NTIMIT, RM1, RM2, Switch Board, etc. are popular

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, January 2026
Impact Factor: 7.67

databases. The TIMIT corpus of read speech provides speech data for acoustic-phonetic investigations and automated speech recognition system development and assessment. The DARPA TIMIT acoustic-phonetic continuous speech corpus (TIMIT—Texas Instruments and Massachusetts Institute of Technology) comprises phonetically balanced prompted English speech. A Sennheiser close-talking microphone captured it at 16 kHz with 16-bit sample resolution [22-23]. TIMIT comprises 6300 sentences, 10 from each of 630 speakers from 8 key US dialect areas. All phone-level phrases were manually segmented. GlobalPhone is a 20-language library of high-quality read speech with transcriptions and pronunciation dictionaries. GlobalPhone [24] standardised data, voice quality, collecting, transcription, and phone set protocols across languages. GlobalPhone provides over 400 hours of transcribed audio data from over 2000 native speakers for research in multilingual speech recognition, rapid deployment of speech processing systems to unsupported languages, language identification tasks, speaker recognition in multiple languages, multilingual speech synthesis, and monolingual speech recognition in many languages.
Aurora 2, the first Aurora database, recognizes digit strings in noise and channel distortion. Artificially distorted assessment data. SpeechDatCar's loud automobile voice data is used in Aurora 3. Despite being a digit identification challenge, Aurora 3 collects utterances in loud surroundings [21]. The Aurora 4 challenge is a common big vocabulary continuous voice recognition job that intentionally corrupts pristine Wall Street Journal (WSJ) data. Aurora 5 was designed to study how hands-free voice input affects digit identification in loud rooms and via cellular networks. Artificially simulated assessment data.

Speech Recognition

Tools Researchers on automatic speech recognition have several potential choices of opensource toolkits for building a recognition system. Notable among these are: HTK, Julius (both written in C), Sphinx-4 (written in Java) of the Carnegie Mellon University and Kaldi, a free, open-source toolkit for speech recognition research. Kaldi provides a speech recognition system based on finite-state transducers (using the freely available OpenFst), together with detailed documentation and scripts for building complete recognition systems [25]. Some of the other less popular open-source systems and kits are RWTH Aachen Automatic Speech Recognition System (RASR), Segmental Conditional Random Field Toolkit for Speech Recognition (SCARF), Improved ATROS (iATROS), SRI International's Decipher, idiap's Juicer and SHoUT speech recognition toolkit [26].

Measures of Performance

The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of performance accuracy which is usually rated with word error rate (WER), whereas speed is measured with the real time factor. Other measures of accuracy include Single Word Error Rate (SWER) and Command Success Rate (CSR). The performance of the speech recognizer is measured in terms of Word Error Rate (WER) and Word Recognition Rate (WRR) [3]. Word errors are categorized into number of insertions, substitutions and deletions. Finally, the word error rate and word recognition rate are computed by the following equations.

$$\text{Word Error Rate}(\%) = \frac{\text{Insertion(I)} + \text{Substitution (S)} + \text{Deletion (D)}}{\text{No. of Reference Words (N)}} * 100 \quad (7)$$

$$\text{Word Recognition Rate (WRR)=1 - WER} = \frac{N - S - D - I}{N} \quad (8)$$

II. CONCLUSION

This review included voice recognition architecture, parameterization, methods, characteristics, problems, databases, tools, and applications. Building automated systems that comprehend and recognize speech like humans is difficult. Automatic speech recognition research addresses voice recognition challenges. Research is advancing in robust, multimodal, and multilingual voice recognition. ASR for English, French, and Czech is mature, while Chinese and Japanese are developing. There is little ASR research in Indian languages, however this must change and numerous ASR tasks must be undertaken to provide effective interfaces in local languages to facilitate technology utilization.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, January 2026
Impact Factor: 7.67

REFERENCES

[1]. X. Huang and L. Deng, ―An Overview of Modern Speech Recognition , in Handbook of Natural Language Processing‖, Second Edition, Chapter 15, Chapman & Hall/CRC, (2010), pp. 339-366.

[2]. X. Huang, A. Acero and H.-W. Hon, ―Spoken Language Processing: a guide to theory, algorithm, and system development‖, Prentice Hall, (2001).

[3]. D. Jurafsky and J. H. Martin, ―Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition‖, Prentice Hall, (2009).

[4]. M. A. Anusuya and S. Katti, ―Front end analysis of speech recognition: a review‖, Int. J. Speech Technology, vol. 14, no. 2, (2011), pp. 99–145.

[5]. J. Li, L. Deng, Y. Gong and R.H.-Umbach, ―An Overview of Noise-Robust Automatic Speech Recognition‖, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4,(2014), pp. 745 – 777.

[6]. S. B. Davis, and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, (1980), pp. 357–366.

[7]. R. Lawrence and B.-H. Juang ―Fundamentals of Speech Recognition‖, Prentice-Hall, Inc., (Engelwood, NJ), (1993).

[8]. M. A. Anusuya and S. K. Katti, ―Speech Recognition by Machine:A Review‖, International Journal of Computer Science and Information Security, vol. 6, no. 3, (2009), pp. 181 -205.

[9]. H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 26, no. 1, (1978) pp.43–49.

[10]. J. K. Baker, "The Dragon System-An Overview',' IEEE Trans. on Acoustics Speech Signal Processing, Vol. ASSP-23, no. 1, (1975), pp. 24-9.

[11]. J. Bilmes, ― hat HMMs can do,‖ IEICE Trans. Inf. Syst., vol. E89-D,no. 3,(2006), pp. 869–891.

[12]. S. Renals, N. Morgan, H. Boulard, M. Cohen and H. Franco, ―Connectionist probability estimators in HMM speech recognition‖, IEEE Trans. Speech Audio Processing , vol. 2, no. 1, (1994), pp. 161–174.

[13]. N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki,M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cretin, H. Bourlard and M. Athineos, ―Pushing the envelope—Aside [speech recognition]‖, IEEE Signal Process. Mag., vol.22, no. 5, (2005), pp. 81–88.

[14]. H. A. Bourlard and N. Morgan, ―Connectionist Speech Recognition- A Hybrid Approach", kulwer Academic Publishers, (1994).

[15]. N. Smith and M. J. F. Gales, "Using SVM's and discriminative models for speech recognition", Proc. ICASSP, vol. 1, (2002), pp.77 -80.

[16]. D.Yu and L. Deng, ―Automatic Speech Recognition - A Deep Learning Approach‖, Springer-Verlag London, (2015).

[17]. L. Deng and X. Li, ―Machine Learning Paradigms for Speech Recognition: An Overview‖, IEEE Transactions on Audio, Speech, and Language Processing, vol. 21 no. 5, (2013), pp.1060-1089.

[18]. G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups", IEEE Signal Process. Magazine, vol. 29, no. 6, (2012), pp. 82-97.

[19]. B. Jacob, M.M Sondhi and H.Yiteng, ―Springer Handbook of Speech Processing‖, Springer, (2008).

[20]. M. Forsberg, ― hy Is Speech Recognition Difficult?‖, Chalmers University of Technology, Citeseer, (2003)

DOI: 10.48175/568