International Journal of Scientific & Engineering Research, Volume 12, Issue 6, June-2021
ISSN 2229-5518

A Complete Perspective on Speech Recognition

Nischala Prasad[1], Kushagra Galundia[2], Daksh Gupta[3], Preetha S[4]
Department of Information Science & Engineering, B.M.S. College of Engineering, Bengaluru, VTU, India

Abstract: The most primitive form of human communication is speech. The procedure of recognizing and converting these speech signals that are acoustic in nature, into digital signals which are later converted into text using machine learning algorithms is known as Speech Recognition. Speech Recognition, also called Automatic Speech Recognition has advanced and developed a lot over the years with the improvement in computer technology in the fields of computing power and storage. Speech recognition is a subset of signal processing. In this paper, we have compared various aspects and methodologies of speech recognition systems and discussed the impacts of different techniques and methodologies on speech recognition system.

Keywords: Deep Learning, Dynamic Time Warping, Feature Extraction, Hidden Markov Model, Language Modelling, Neural Network, Speech Recognition System,

I. INTRODUCTION

Speech recognition (SR), also known as Automatic Speech Recognition (ASR), is the act of turning speech sounds into a sequence of words using an algorithm on a computer device. Speech recognition is a subset of signal processing. It aims to develop and enhance the existing techniques of speech input into a machine. Earlier, the computer machines available were restricted in terms of storage and computing power, which in turn restricted the growth of ASR systems. However as time progressed, the advancement of computer machines in terms of storage capacity and computing power has allowed automatic speech recognition to improve. Speech is the most basic mode of communication and there exist several languages all over the world through which people communicate. At present, speech recognition systems have been developed only for a few major languages, providing future scope for speech recognition systems to be developed in native languages. In this paper, we aim at highlighting the fundamentals of speech recognition systems, the various types of speech recognition available, their methodologies, the progression of such systems over the years, their applications and also the challenges faced in developing these systems.

In a Speech Recognition System, speech input is usually captured through a hardware device which captures it in analog form and is later converted into an electrical or digital signal. This conversion is done by the sound card which also has the capabilities to store and replay the speech input multiple times. There are 5 building blocks in a Speech Recognition System:

1. *Signal Pre-processing*
   A speech input accepted from a hardware input is analog in nature, this has to be digitized according to the Nyquist Theorem which asserts that “a signal must be sampled at more than twice the highest frequency component of the signal”. This conversion is done in signal pre-processing.

2. *Feature Extraction*
   The parameterization of the speech signal is done through feature extraction by finding stable and correlated properties. These parameters will form the observation vectors which will be used for providing accurate classification. Various techniques can be used for feature extraction, but there are 3 techniques which are widely used, these are:
   i. Linear Predictive Coding (LPC) - This is the most optimal technique for finding basic speech parameters in a speech signal. The concept behind this technique is that past speech samples can be used to approximate present samples by following a linear combination. The sum of squared differences between samples are minimized to obtain LPC features.
   ii. Mel Frequency Cepstral Coefficients (MFCC) - This is a widely used technique for feature extraction because of the positioning of its feature bands in a

logarithmic way. It is better at approximating human speech signals than any other technique. When a signal is input into the system, it passes through the hamming window to reduce discontinuities present in the signal. The MFCC method makes use of "Discrete Fourier Transform" and "Inverse Discrete Fourier Transform" to calculate Cepstral coefficients.

iii. Perpetual Linear Prediction - This technique makes use of the three main characteristics present in psycho-acoustic properties of human hearing, making it an improvement over LPC. The three characteristics which are used are- Equal loudness curve, Hearing band spectral resolution and Power law for loudness intensity.

3. *Language Modelling*

In language modelling, "mth" word is predicted using the "m-1" preceding words in order to find the perfect fit word sequence. There are 4 types of language modelling:

i. Uniform Model – Each word has an equal probability of occurring
ii. Stochastic model – Probability of a word occurring is dependent on the occurrence of its previous word
iii. Finite State Languages –Word sequence is found by using a finite state network
iv. Context Free Grammar (CFGs) - Sequence of words are encoded using CFGs which are then used for recognition

4. *Decoder*

In the Decode stage, the "most likely" word sequence which fits the speech input is found. This word sequence is found by using dynamic programming algorithms which will search for a sole path through the network that will give the best match. The most popular algorithm used for decoding is Viterbi algorithm.

5. *Speech Recognition*

The recognition of the speech signal is the final stage in a Speech Recognition System. This is divided into two phases: Training and Testing. In the training phase, identification of objects is carried out and the more times this is done, the better the performance and accuracy become. Testing phase is the comparison between the training phase patterns and the input pattern. The closeness of these two phases determines the performance of the system.

Any Speech Recognition System is classified based on the types of speech or utterances which are to be recognized. These types are:

1. Isolated Words - These systems recognize only one word at a time. They have a listen and a not-listen state which means the speaker has to wait between the utterances of two words. During this pause, the words are processed one by one.
2. Connected Words - Similar to isolated words, but can process more than one word at a time and require lesser pauses.
3. Continuous Words - In these systems, the speaker can speak in a neutral way and the system will process what the speaker is saying. These are hard to design as it is difficult to set the utterance boundaries.
4. Spontaneous Words - In this kind of system, any mispronounced or unheard words like "uh" or "um" are covered with false statements.

II. RELATED WORK

From evolution to new breakthroughs in speech recognition systems, Neha Jain et al. [1] presented the ideas of speech recognition systems. They created a small comparison between the algorithms and the models that were and are currently being used to implement Speech Recognition Systems. Dynamic Type Warping (DTW) was made use of in areas of speaker recognition and Hidden Markov Model (HMM) was made use of in areas of speech conversion. Speech recognition faced many challenges and tools and frameworks were developed to overcome these challenges like Voice Activity Detector, AURORA framework, etc.

In terms of throughput, Word Error Rate (WER), and latency, Vineel Pratap et al. [2] developed an online convolutional voice recognition system that beats two strong baselines. The system was built using "Time-Depth Separable (TDS)" convolutions and "Connectionist

Temporal Classification (CTC)". About 1 million utterances of in-house English videos posted by users publicly, made up the training dataset. The machine had a throughput of 147 audio seconds processed per wall clock second and a latency of 1.09 seconds, with a clean WER of 13.19 and a noisy WER of 21.16.

Praveen Edward James et al. [3] put forward a speech recognition system based on an efficient "Recurrent Neural Network (RNN)" using software along with "long short-term memory (LSTM)". The suggested system's five levels were an input layer, a fully linked layer, a secret LSTM layer, a SoftMax layer, and a sequential output layer. The method is taught using an 80-word vocabulary that can be broken down into 20 sentences. When the concealed layer depth was 42, the system had a maximum accuracy of 89 percent.

Mohit Bansal et al. [4] performed an experiment to find out the better model between Convolutional Neural Networks (CNN) and Basic Neural Network for speech recognition. A free Google dataset that consists of around 65,000 one-second sound files was used to train and test the models. In comparison, CNN resulted in a train accuracy of 99% and a validation accuracy of 79% on testing the model with test data which gave 91% of accuracy compared to the basic Neural Network's train accuracy of 20% and validation accuracy of 19% on testing the model with test data which gave 19%. After analysing the results a conclusion was made that CNN performed better than Basic NN on images and that it also gave a better accuracy.

Leonid Velikovich et al. [5] described a contextual ASR system that uses semantic lattice processing and rescoring in order to improve recognition accuracy. It was shown that the system reduces WER by 12% relative to media playing commands test set as well as brings quality gains on live Google Assistant traffic. An analysis on a use case in which a broad group of semantic entities such as songs or films are used in lattice processing was made.

Hardik Dudhrejia et al. [6] described speech recognition as a 4 step process - speech preprocessing, feature extraction, Speech classification and recognition. This model used various deep learning techniques, recurrent neural networks and LSTM models to extract speech features. The main focus was on recognition of speech in English language.

Loc Hoang Tran et al. [7] applied a hybrid of the Principal Component Analysis (PCA) approach and "Kernel Ridge Regression" on automatic speech recognition. A total of 4,500 voice samples from 50 distinct terms were used to train the model (90 speech samples per word). A set of 500 people was used to assess the sensitivity tests. This model gave a sensitivity measure of 95.4% when compared with 89% of traditional HMM method.

Abhinav Jain et al. [8] described how to improve the performance and reliability of accented speech with various languages. The proposed model used a multi task architecture which handled multi accent acoustic models with various related accent information. Common voice corpus dataset from Mozilla was used in this experiment. This model gave a WER reduction of 15% relative in cases of unseen accent and 10% in seen accents.

R.Thiruvengatanadhan [9] suggested an "Associative Neural Network-based (AANN)" speech recognition system. Individual words were segregated out of continuous speeches using "Voice Activity Detection (VAD)". Each individual utterance were modeled using AANN. MFCC was used for feature extraction. The model was trained using a total dataset of 100 separate speech dialogue recordings, varying in length from 5 to 10 seconds, sampled at 16 kHz and encoded by 16-bit, and using a tuner card, was able to gather television broadcast voice data from Tamil news networks. This experiment resulted in a good performance of 92% as the recognized rate.

A link between native rear-based confidence estimation within the acceptor HMM framework along with the training of ANNs in hybrid HMM/ANN based systems was proposed by S. Pavankumar Dubagunta et al. [10]. Mediaparl and AMI datasets were used to conduct the ASR experiments. Through this link, it was shown that based on linguistic segments error functions can be used to train the ANNs. On performing experiments on different datasets, it was found that better ASR systems were yielded on using ANNs trained at segment level. The projected approach conjointly provided a link to the "Kullback-Leibler divergence" based mostly HMM (KL-HMM) framework.

As an alternative to the attention-based approach, Parnia Bahar et al. [11] suggested a basic 2-dimensional sequence-to-sequence model. Input and output

International Journal of Scientific & Engineering Research, Volume 12, Issue 6, June-2021
ISSN 2229-5518

representations were merged with the use of a 2-Dimensional LSTM layer. The production dataset was the Hub5'00, which included “Switchboard (SWB)” and “Callhome (CH)”, and the evaluation dataset was the Hub5'01. The findings of the experiment were comparable to the baseline on the “300h-Switchboard Hub'00” and demonstrated a 0.4% gain on the “Hub'01”.

The “Lattice-Free Maximum Mutual Knowledge (LF-MMI)” objective function was proposed by Hossein Hadian et al. [12] for end-to-end training of acoustic models. The Switchboard and Wall Street Journal (WSJ) databases were used in the model's experiments. Upon running the experiments, the model was found to give a WER of 12.8 in character-based case and WER of 11.0 in phoneme-based setting.

For children speech recognition, S. Pavankumar Dubagunta et al. [13] contrasted the traditional CNN-based end-to-end acoustic modelling method with a cepstral feature-based ASR methodology. According to research on the “PF-STAR corpus”, CNN-based end-to-end acoustic modelling resulted in stronger systems than those with typical characteristics. Children's speech was tested using the PF-STAR dataset, while adult speech was tested using the WSJCAM0 dataset. After the experiment, it was seen that the WER for children speech model was 13.36% with adult data and 15.62% with children data and the WER for adult speech model was 14.09% with adult data and 16.60% with children data.

Purvi Agrawal et al. [14] proposed a technique that preserved the important modulations of voice signals in the spectro-temporal domain using an unsupervised data-driven modulation filter learning technique. This conservation was achieved utilising a deep generative modelling system that uses a “Convolutional Variational Autoencoder (CVAE)” to learn modulation filters. “Aurora-4 (additive noise with channel artifact)” and “CHiME-3 (additive noise with reverberation)” databases were used to execute the ASR experiments. WER over base features showed a relative average improvement of 9% on “Aurora-4 database” and 23% on “CHiME-3 database”. Semi-supervised training of ASR on Aurora-4 database having 30% labelled data gave a relative average improvement of 29% over base features.

Hengguan Huang et al [15] put forward a novel model that can address hybrid acoustic modeling by incorporating the proposed Recurrent Poisson Process (RPP) into a RNN. It was shown that the model can generate much better alignments while performing the HMM state modeling. The experiments performed on CHiME-2, WSJ0 and WSJ0&1 datasets show that the method achieves much better results than several RNN baselines in ASR and provided a WER of 13.2%, 2.5% and 6.2% respectively.

Table. 1. Various methodologies, datasets and findings

Reference No./Year	Methodology / Algorithm	Dataset	Findings

[2] - 2020	Time-Depth Separable convolutions and Connectionist Temporal Classification	1 million utterances of in-house English videos publicly shared by users	Clean WER = 13.19
			Noisy WER = 21.16
			Throughput = 147
			Latency = 1.09 seconds

[3] - 2020	Recurrent Neural Network using software with long short-term memory	A vocabulary of 80 words which constitute 20 sentences	Maximum accuracy = 89%
[4] - 2020	Convolution Neural Network and basic Neural Network	A free Google dataset that consists of around 65,000 one-second sound files	CNN : Train accuracy = 99% Validation accuracy = 79% Test data accuracy = 91% NN :

			Train accuracy = 20% Validation accuracy = 19% Test data accuracy = 19%
[5] - 2018	Semantic Lattice Processing and Rescoring	Live Google Assistant traffic	WER reduced by 12%
[7] - 2018	Principal Component Analysis method and Kernel Ridge Regression	A set of 4,500 speech samples recorded of 50 different words	Sensitivity Measure = 95.4%
[8] - 2018	Multi Task Architecture	Common voice corpus dataset from Mozilla	Unseen Accents : WER Reduction = 15% Seen Accents: WER Reduction = 10%
[9] - 2019	Associative Neural Network, Voice Activity Detection and Mel Frequency Cepstral Coefficients	100 different speech dialogue clips from Tamil news channels	Performance = 92%
[10] - 2019	Artificial Neural Network and HMM Framework	Mediaparl and AMI	A better ASR system and a link to the Kullback-Leibler divergence based HMM (KL-HMM) framework
[11] - 2019	A 2-Dimensional Long-Short Term Memory	Full Hub5'00 including Switchboard and Callhome	Improvement in Performance = 0.4%

[12] - 2018	Lattice-free maximum mutual information objective function	Switchboard and Wall Street Journal	Character-based case:
			WER = 12.8
			Phoneme-based setting:
			WER = 11.0

[13] - 2019	CNN-based end-to-end acoustic modeling approach	PF-STAR dataset for children speech and the WSJCAM0 dataset for adult speech	Children speech model:
			WER = 13.36% (adult data); 15.62% (children data)
			Adult speech model:
			WER = 14.09% (adult data); 16.60% (children data)

[14] - 2019	Convolutional Variational Autoencoder	Aurora-4 and CHiME-3	Aurora-4 :
			Improvement = 9%
			CHiME-3 :
			Improvement = 23%

[15] - 2019	Recurrent Poisson Process (RPP) into a RNN	CHiME-2, WSJ0 and WSJ0&1	CHiME-2 WER = 13.2% WSJ0 WER = 2.5% WSJ0&1 WER = 6.2%

III. METHODOLOGIES OF SPEECH RECOGNITION

Classification of SR methodologies can be made through three categories, which are – “Acoustic-Phonetic Approach”, “Pattern Recognition Approach” and “Artificial Intelligence Approach”.

1. *Acoustic-Phonetic Approach*
   It is a rule based approach. According to this approach, each spoken instance has a finite set of distinctive phonetic units. These are extracted in the form of features through spectral analysis of speech signals. These extracted features are segmented and labelled giving an end product of a valid word.

2. *Pattern Recognition Approach*
   This approach uses two steps – in the first, the system is trained with all possible patterns and the second step involves simply comparing unknown patterns with the trained patterns to recognize them. This can be either a speech template or a statistical model.

3. *Artificial Intelligence Approach*
   It merges the concepts of Acoustic-Phonetic approach and Pattern Recognition approach to build a hybrid model. It uses “Dynamic Time Warping (DTW)” and “Hidden Markov Models (HMM)”. HMM has lower memory requirements, hence it is preferred over DTW. Artificial Intelligence approach is efficient only when the dataset is small and it can also handle high complexity tasks. The most basic approach of artificial neural networks is Phoneme Recognition.

4. *Generative Learning Approach (HMM-GMM)*
   This approach uses a HMM based on the Gaussian-Mixture model for speech recognition. GMM-HMM is a state transition matrix and a parameterized vector with prior probabilities.

5. *Discriminative Learning (HMM-ANN)*
   Neural networks trained using discriminative learning provide an efficient yet natural way of recognising speech signals. This approach is suitable only for small sets of speech recognition like isolated words.

6. *Deep Learning (HMM-DNN)*
   This is an unsupervised feature learning approach. It has 3 architectures – generative, discriminative and hybrid. DNN has replaced all other older techniques of speech recognition.

IV. PROBLEMS IN SPEECH RECOGNITION

The execution of a speech recognition system relies on its ignorance to its surrounding variables. Several factors can influence its performance like the condition of its surroundings, dependency on hardware and speed at which speech input is received. Since these systems get more advanced, it is important to build a system that can modify itself with respect to its surroundings or various circumstances and develop auto-generative algorithms.

V. AREAS OF APPLICATION

1. Improvement of storage technologies to handle big data has resulted in a wide acknowledgement of speech recognition. It has minimized the manual labor required in many fields.
2. Automatic call processing and voice dialing in telephone networks
3. Entry of data into databases
4. Dictation and voice enabled search
5. Natural language understanding and processing
6. Live translators
7. Domestic appliance control
8. Voice assistants
9. Smart homes

VI. PERFORMANCE

Speech recognition system’s performance is measured by the system’s computation speed and the accuracy of its results. A Real Time Factor (RTF) is used to define speed of the system and the system’s accuracy is measured using Word Error Rate (WER). Performance is the complementary part of WER. Word errors are defined by the number of deletions, substitutions and insertions while analysing a speech input.

$WER = (S + D + I) / N$

The number of word substitutions is given by S, the number of word deletions is given by D, the number of insertions is supplied by I, and the number of utterances is given by N.

Word Recognition Rate (WRR) = 1 - WER RTF = T / D
where, T is given as processing time and D as the duration of process.

VII. CONCLUSION

Speech is the most essential method of communication between people, so a practical interface is needed to associate people with machines. Although this field has acquired wide endorsement to mechanize the administrations and applications, there are a few boundaries which influence the precision and proficiency of speech recognition systems. Vigor of the speech system relies upon some steady boundaries / highlights of the speech signal. To improve the force of speech recognition systems, it is needed to plan speech recognizers in local dialects. Multilingualism is another developing field in the space of speech recognition. There is a great deal of improvement and examination in the field of unknown dialects. However, to upgrade its force and utility for local individuals, it is vital to utilize this innovation in local dialects.

VIII. REFERENCES

[1] Jain, Neha, and Somya Rastogi. "Speech Recognition Systems–A Comprehensive Study Of Concepts And Mechanism." Acta Informatica Malaysia (AIM) 3.1 (2019): 1-3.

[2] Pratap, Vineel, et al. "Scaling up online speech recognition using convnets." arXiv preprint arXiv:2001.09727 (2020).

[3] James, Praveen Edward, et al. "Recurrent neural network-based speech recognition using MATLAB." International Journal of Intelligent Enterprise 7.1-3 (2020): 56-66.

[4] Bansal, Mohit and Dr.T.K. Thivakaran. "Analysis of Speech Recognition using Convolutional Neural Network." (2020).

[5] Velikovich, Leonid, et al. "Semantic Lattice Processing in Contextual Automatic Speech Recognition for Google Assistant." Interspeech. 2018.

[6] Dudhrejiya, Hardik, and Shah, Sanket. "Speech Recognition using Neural Networks." International Journal Of Engineering Research & Technology (IJERT) Volume 07, Issue 10 (October – 2018)

[7] Tran, Loc Hoang, and Linh Hoang Tran. "The combination of sparse principle component analysis and kernel ridge regression methods applied to speech recognition problem." Int J Adv Soft Comput Appl 10.2 (2018).

[8] Jain, Abhinav, Minali Upreti, and Preethi Jyothi. "Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning." Interspeech. 2018.

[9] Thiruvengatanadhan, R."Speech Recognition using AANN". International Journal of Innovations in Engineering and Technology (IJIET). (2019).

[10] Dubagunta, S. Pavankumar, and Mathew Magimai Doss. "Segment-level Training of ANNs Based on Acoustic Confidence Measures for Hybrid HMM/ANN Speech Recognition." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[11] Bahar, Parnia, et al. "On using 2d sequence-to-sequence models for speech recognition." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[12] Hadian, Hossein, et al. "End-to-end Speech Recognition Using Lattice-free MMI." Interspeech. 2018.

[13] Dubagunta, S. Pavankumar, Selen Hande Kabil, and Mathew Magimai Doss. "Improving children speech recognition through feature learning from raw speech signal." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[14] Agrawal, Purvi, and Sriram Ganapathy. "Deep variational filter learning models for speech recognition." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[15] Huang, Hengguan, Hao Wang, and Brian Mak. "Recurrent poisson process unit for speech recognition." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 2019.