Speech Recognition Implementation

Kanika Bansalwal, Kamaldeep Sharma, Anuj Jain

Abstract: ***This paper presents a brief review on Automatic Speech Recognition and provide a technical understanding of ASR system. The objective of this review paper is to elaborate one of the best techniques in the field of speech recognition that is hidden Markov model. Hidden Markov model is very popular technique for speech recognition because speech signal is more like piecewise stationary or short time stationary signal and these models can be trained easily and they are computationally feasible. So, this paper gives a proper implementation of hidden Markov model. After so many years of research, the main challenge in speech recognition field is accuracy. The speech recognition system includes feature extraction, building word template, comparing word and selecting the best with maximum likelihood. Hence, this paper will give a great contribution for understanding the concepts of Automatic Speech Recognition system and hidden Markov model.***

Index Terms: Automatic Speech Recognition, Hidden Markov Model, Pre-Processing, Feature Extraction, Word Template, Robust Speech Recognition, Short-Time Fourier Transform, Peak Detection.

I. INTRODUCTION

Speech recognition is a technique through which words or sentences can be recognized and converted into texts by means of an algorithm. It enables the recognition of speech through converting speech signal into sequence of data.

II. SPEECH SIGNAL PROCESSING

A. Basic Model of Speech Recognition

Speech is an analog signal that must be converted into digital signal before recognition. So, firstly speech signal is converted into digital signal. Then important features are extracted from the signal and this information is stored inthe voice template. When input is provided to the system then again features are extracted from the input and thenthey are compared with the voice template and system gives output that matches best with the previously stored features.
Hence, best result is taken out from the system and given as output. This whole process can be explained by the flow chart given below.

In the flow chart, it can be easily seen that speech signal is firstly pre-processed. In pre-processing stage, external noise from the speech signal is removed and only necessary speech signal is extracted. Then feature extraction technique is applied. Hence, number of features are extracted from the speech signal and a reference model is created for comparison.

When input signal is provided for recognition then first of all, it is pre-processed. Then feature extraction technique is applied to this signal. Then we have features of the input signal. Now we can compare these features with the reference model and find out the best match.

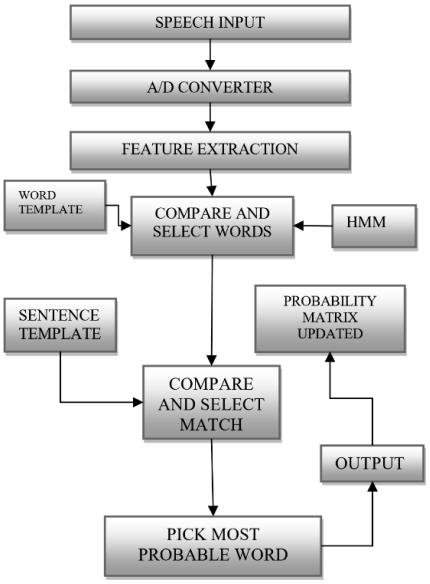

Fig. 1 Basic Model of Speech Recognition

B. Types of Speech Recognition System

According to manner of speech, speech recognition system is divided into three parts-

1. Isolated word recognition that requires pause after every word
2. Connected word recognition that requires properly pronounced every word
3. Continuous speech recognition

According to manner of speech, speech recognition system is divided into three parts-

1. Isolated word recognition that requires pause after every word
2. Connected word recognition that requires properly pronounced every word
3. Continuous speech recognition

According to the vocabulary size, speech recognition

Revised Manuscript Received on July 10, 2019

Kanika Bansalwal, Research Scholar, School of Electronics and Electrical Engineering, Lovely Professional University, Jalandhar, India.
Kamaldeep Sharma, School of Electronics and Electrical Engineering, Lovely Professional University, Jalandhar, India.
Anuj Jain, School of Electronics and Electrical Engineering, Lovely Professional University, Jalandhar, India.

system is also divided into three parts-
1. Small vocabulary that contains dozens of words in its memory
2. Medium vocabulary that contains hundreds or thousands of words
3. Large vocabulary that contains thousands or more words [1]

C. Techniques for Speech Recognition

There are number of techniques for speech recognition. Some of them are listing below-
1. Linear predictive analysis (LPA)
2. Linear predictive cepstral coefficients (LPCC)
3. Perceptual linear predictive coefficients
4. Mel-frequency cepstral coefficients (MFCC)
5. Mel-scale cepstral analysis
6. Power spectral analysis
7. Relative spectra filtering of log domain coefficients
8. First order derivative
9. Artificial neural networks (ANN)
10. Hidden Markov model (HMM)

From all of these techniques best accuracy can be attained by artificial neural networks and hidden Markov model. If we compare between artificial neural networks and hidden Markov models then hidden Markov model is best in accuracy. It gives approx. 98% accuracy in the output. Now we will understand how ASR works.

III. ASR WORKING

When any person or animal produce voice that is called as acoustic signal having some particular frequency is transmitted to the environment. If any other person is listening to that then this frequency will be going through his or her ears and he or she will try to understand the meaning of that acoustic wave. In any machine or system that transmit or record human speech, the acoustic wave is converted into an electrical analogue signal using a microphone. For example, if we speak in telephone then firstly microphone of the telephone converts the acoustic signal into an electrical analogue signal that will be transmitted to the telephone network. The strength of the electrical signal varies in amplitude over time and referred as analogue waveform or analogue signal. Listener's brain and ears receive the analogue signal and then process this signal to figure out the speech. ASR also works on the same principle by taking the speech signal and then performing analysis and synthesis on that signal.

A. Steps that ASR System Involve

1. Speech signal goes through the microphone that converts it into electrical signal.
2. Noise is removed and only words are attained from the speech.
3. The words are divided into particular individual sounds that are the smallest unit of sound and they are known as phonemes. 4. Then sounds are mapped to character groups in the ASR program's database.
5. This program contains a large dictionary of words exists in the languages.
6. Now, phonemes are compared with the sounds already existing in the program and then converted into characters.
7. Then grammar check is done whether the characters are making sense or not.
8. Hence the ASR system works and recognizes the speech.

B. Speech Variations

The process of speech recognition is very complicated. It is because of the production of phonemes and transition between the phonemes that varies person to person and even in the same person. A person cannot speak exact the same as before spoken. There are so many factors that varies the spoken phonemes by the person like age, sex, emotional state, accents, illness, stress or other conditions. These changes in the waveforms can change the performance of the recognition system.

C. Systems to Use for ASR

There are two types of system in the field of speech recognition. They are
1. Speaker independent system
2. Speaker dependent system

Speaker independent systems can provide high recognition accuracies for a wide range of users without taking every user's voice but speaker dependent systems need to be trained using user's voice then only it can achieve high accuracies.

IV. REVIEW OF LITERATURE

Dr. Edward Christopher Wente stuck with the work of developing acoustical devices for speech or music recording. He worked upon the development of high-quality microphones that can able to convert the acoustic signal into electrical signals. The first and foremost work in speech recognition that is conversion of acoustic signal into electrical signal is done by him in Bell Laboratories. He worked on this for approx. twenty years [2].

E. E. David talked about the importance of segmentation of words into utterances or discrete units for better performance of recognition [3].

D.B. Fry told about mechanical speech recognition's theoretical aspects. His technique is based upon the use of language system that includes linguistic units like phonemes, words, sentences etc. He resolves the uncertainties in the speech recognition. He explained that a mechanical speech recognizer containstwo aspects that are inspection of acoustic signal in number of ways and then statistical knowledge applied to the results [4].

H. Olson and H. Belar talked about time compensation for talking speed in machines of speech recognition.

He developed a system that can provide an automatic time compensation for speed variations in talking in the speech recognition systems [5].

Thomas Marill explained in his paper about three fundamental concepts of speech recognition that are intensity frequency time,

linguistic and articulated analysis. In the field of speech recognition, the first hardware system is built by C. P. Smith. After that, the second hardware for automatic speech recognition is built by Davis, Biddulph and Balashek in the Bell Telephone Laboratories. This hardware was able to recognize ten spoken words with 98 percent accuracy. But this system was speaker dependent. It worked accurately for its master voice but it failed for another person's voice [6].

M. Halle and K. Stevens proposed a model in which speech signal is transformed into a sequence of phonemes by an active or feedback process. They explained that for recognition patterns are internally generated in the analyzer according to a sequence of instruction until the best match with the input signal is obtained [7].

Harry F. Olson explained the processing of sound [8].

T. B. Martin and J. J. Talavage described the signal processing techniques for speech phoneme recognition by machine. They told that neural interconnection simulation is done via neural logic. Overall in this research they talk about neural logic networks for speech feature's abstraction [9].

T. Sakai and S. Doshita described a system that overcomes the basic complexities in automatic speech recognition. They told about the method of segmentation of speech to recognize it. This method contains two parts in which first one is about segmentation of the speech sound into small discrete intervals that are also known as phonemes and second part is about pattern recognition of these intervals. After that they talked about two criteria of segmentation that are stability and distance. Then there were three types of classification that are phoneme classification, vowel classification and consonant classification. In this they took time patterns of the speech signals then it is compared with the previously stored patterns and according to that it will give the outputs [10].

W. Bezdel and H.J. Chandler showed the result of analysis or recognition of the vowels like a, e, ou, i, u like that. They showed the results in graphs and tables in detailed form [11]. They published a paper on the results of the analysis and recognition of vowels using zero crossing data by computer [12].

Charles C. Tappert described about the usage of adaptive neural network in the field of automatic speech recognition. He talked about segmentation of speech in means of dyads or possibly syllables or half syllables [13].

M. Lecours and J. Sparkes described about pattern recognition algorithm use to compare spectrum analyzer for speech recognition [14].

R. Putron talked about autocorrelation analysis for speech recognition. This is a technique based upon computer pattern matching. According to him, speech signal is split into two frequency bands and they are quantized into two different amplitude levels. These two signals are fed into autocorrelators that are consisting of digital multipliers, RC integrators or binary shift registers. Then patterns are formed and these pattern at last are compared with the master patterns and then system will give the output according to that [15].

R. De Mori explained about a technique that analyses the timing evolutions of the speech signal based on relational parameters. He showed results in his paper [16].

S. Das described dynamic programming algorithm for speech recognition. In this speech signal is coded in one bit per sample and there is a technique named ADPCM is also used for coding of the speech signal [17].

L.R.Rabiner provided a tutorial on hidden Markov models for speech recognition. In that paper he gave a detailed theory about speech recognition including Markov chains, hidden Markov model implementation [18].

In 2016, paper [19] showed an end-to-end deep learning approach for recognizing English and Mandarin words. This paper used HPC techniques and Batch Dispatch for iterating more quickly to identify algorithms.

This paper [20] presented a speech recognition system model that recognizes speech directly to the words without using any pronunciation model using recurrent neural networks. It describes about two components for recognition of speech that are RNN encoder and RNN decoder.

In 2017, Microsoft updated their 2016 system with neural network-based acoustics to make advancement in the speech recognition task. They added a CNN-BLSTM acoustic model and confusion network in their system. This system resulted with only 5.1%-word error rate [21].

Recently, an improved version of children speech recognizer has built. This is improvised using CNN-based end-to-end acoustic modelling methods. For feature extraction, MFCC is used in this system [22].

In [23] authors demonstrate that electroencephalography can be used for better performance of ASR system in the presence of noise. They used distillation training and showed that this training can improve the accuracy of an ASR system with EEG features in background noise. They have used two datasets vowels and words in their work.

V. MARKOV CHAINS

Markov chains is basically a model that gives information about the probabilities of sequences. It is a stochastic model that models the sequential data that is in an order. It provides a simple way for modelling the dependencies of the information (e.g. weather) with the past information. When we have number of states then based upon the present state, we can make assumption of the next state using Markov chains.

Markov chain assumption is strongly based upon the current state, no matter what the past state is. For example, if we want to predict the tomorrow's weather then we just need the today's weather but we are not allowed to take yesterday's weather. Markov assumption is mathematically expressed as below.

Markov chain can be expressed by following components

1. A set of N states

$$Q = q_1 q_2 \dots q_N \quad (1)$$

2. State transition probabilities, a is the probability of transition from state i to j

$$A = a_{11} a_{12} a_{13} \dots a_{n1} \dots a_{nn} \quad (2)$$

3. The initial probability distribution

$$\pi_i = \pi_1, \pi_2 \dots, \pi_N \quad (3)$$

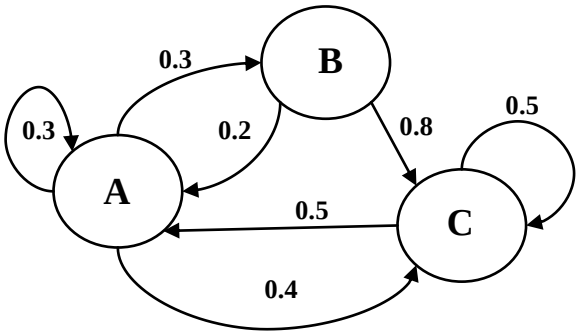

Fig. 2 A Three State Markov Chain

Here, transition probability matrix A is

$$A = \begin{bmatrix} 0.3 & 0.3 & 0.4 \\ 0.2 & 0 & 0.8 \\ 0.5 & 0 & 0.5 \end{bmatrix} \quad (4)$$

VI. HIDDEN MARKOV MODEL

The Markov chains is useful when we are known about the states. But in many cases, the states are hidden. In that case, we use Hidden Markov Models are used. In speech recognition we are known about the words that are normally can be called as events but we are not aware about the part of speech (POS) tags. These are hidden states in the words. For recognition of a word we have words in input that are our observations and we need to compute the hidden states that are POS tags. The HMM can be specified by the following components.

1. The set of N states
$$Q = q_1 q_2 \dots q_N \quad (5)$$
2. The transition probability A
$$A = a_{11} a_{12} a_{13} \dots a_{n1} \dots a_{nn} \quad (6)$$
3. Sequence of T observations
$$O = o_1 o_2 o_3 \dots o_T \quad (7)$$
4. The observation likelihood also called as emission probability B
$$B = b_i(o_t) \quad (8)$$
5. The initial probabilities
$$\pi_i = \pi_1, \pi_2 \dots, \pi_N \quad (9)$$

The three fundamental problems that arises in HMM implementation. They are as follows

1. Given the observation sequence O and the model $\lambda = (A, B)$, how we can compute the probability of the sequence $P = (O|\lambda)$.
2. Given the observation sequence O and the model $\lambda = (A, B)$, how to compute the state sequence which is best and meaningful.
3. Adjusting the model parameters $\lambda = (A, B)$ to maximize the probability $P = (O|\lambda)$.

Problem 1 states that when we have given model and observation sequence then what is the probability that the sequence is produced by the model. It is about the chances of the observation sequence occurring through the model.

Problem 2 states that which of the state sequence from the model is most similar with the given observation sequence.

Problem 3 is about adjusting the values of model parameters. This adjustment is called as training and the observation sequence is known as training data.

To solve the problem 1, we need to calculate the probability of the observation sequence given the model to us. We assume a state sequence

$$Q = q_1 q_2 \dots q_T \quad (10)$$

Here, q_1 is the first state and q_n is the final state. The probability of observation sequence O to be from the assumed state sequence Q is

$$P(O|Q, \lambda) = \prod_{t=1}^{N} P(O_t|q_t, \lambda) \quad (11)$$

Assuming statistical independence

$$P(O|Q, \lambda) = b_{q_1}(O_1).b_{q_2}(O_2) \dots b_{q_T}(O_T) \quad (12)$$

Probability of state sequence Q

$$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \dots a_{q_{T-1} q_T} \quad (13)$$

Now, the probability of O and Q both

$$P(O, Q|\lambda) = P(O|Q, \lambda) P(Q, \lambda) \quad (14)$$

Now, probability of O will be

$$P(O|\lambda) = \sum_{all\ Q} P(O|Q, \lambda) P(Q|\lambda)$$
$$= \sum_{q_1, q_2, \dots, q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \dots$$
$$\dots a_{q_{T-1} q_T} b_{q_T}(O_T) \quad (15)$$

For all of this calculation, we have to do $(2T-1)N^T$ multiplications and (N^T-1) additions. This much big calculation is merely possible for all state sequences.For reducing the calculations, one method can be applied that is forward-backward method.

A. *Forward Algorithm*

In forward algorithm, we define a forward variable $\alpha_T(i)$, for solving this

1. First initialization
$$\alpha_1(i) = \pi_i b_i(O_1), \quad 1 \le i \le N \quad (16)$$
2. Induction
$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \quad 1 \le t \le T-1 \quad (17)$$
3. Termination
$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \quad (18)$$

Firstly, we are initializing the forward variable by the multiplication of initial probabilities and joint probabilities. This calculation of forward variable can be explained by the following fig.

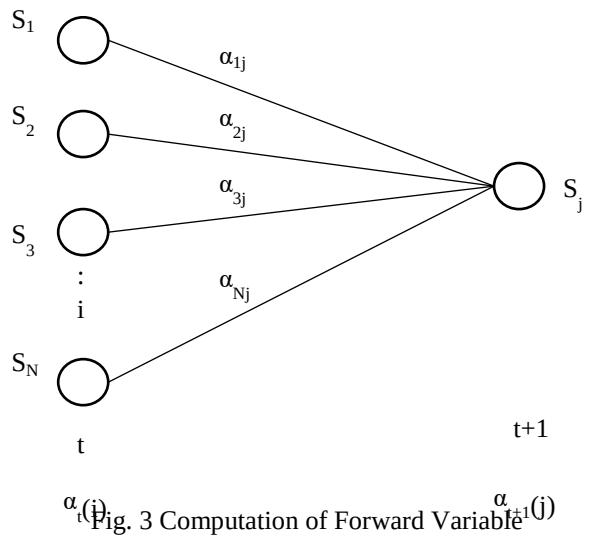

Fig. 3 Computation of Forward Variable

Hence, probability of the observation sequence $P(O|\lambda)$ is the addition of all $\alpha_t(i)$'s. This calculation requires N^2T multiplication that is far more good compare to the first calculation.

B. Backward Algorithm

In this, we will assume the backward variable $\beta_t(i)$, it can be expressed as

$$\beta_t(i) = P(O_{t+1}O_{t+2} \dots O_T | q_t = S_i, \lambda) \quad (19)$$

Similarly, we can solve this inductively

1. Initialization

$$\beta_T(i) = 1, \quad 1 \le i \le N \quad (20)$$

2. Induction

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2 \dots, 1, \quad 1 \le i \le N \quad (21)$$

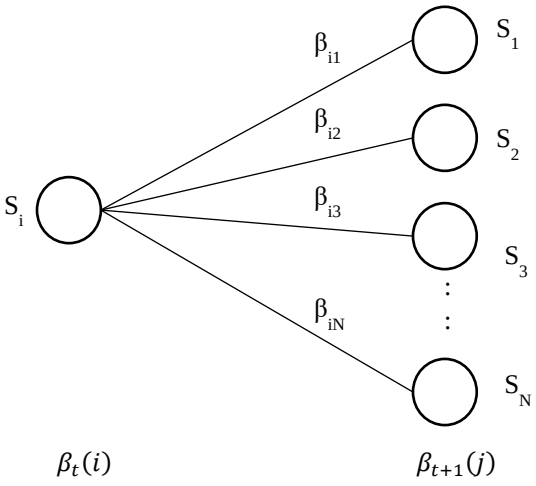

Fig. 4 Computation of Backward Variable

Hence, probability of the observation sequence $P(O|\lambda)$ is the addition of all $\beta_t(i)$. This calculation also requires N^2T multiplication that is far more good compare to the first calculation. This algorithm can understand using the fig. 4.

C. State Likelihood Calculation

Taking one step forward, now we need to find out the optimal state sequence that matches to the observation sequence. Solution for this is to identify the state q_t occur at time individually and then find out which state is occurring maximum time. For solving problem 2, we need to define a variable again $\gamma_t(i)$ as to being in state i at time t, given the observation sequence and the model.

This variable can be expressed as,

$$\gamma_t(i) = P(q_t = S_i | X, \lambda) \quad (22)$$

According to the conditional probability,

$$\gamma_t(i) = \frac{P(X, q_t = S_i, \lambda)}{P(X|\lambda)} = \frac{P(X, q_t = S_i, \lambda)}{\sum_i^N P(X, q_t = S_i, \lambda)} \quad (23)$$

We can rewrite the equation as follows

$$\gamma_t(i) = \frac{\alpha_t(i).\beta_t(i)}{\sum_i^N \alpha_t(i).\beta_t(i)} \quad (24)$$

So, using $\gamma_t(i)$, we can calculate the state that has maximum likelihood. The highest probability of a state $q_t(i)$ being in state S_i at time t can be expressed as

$$q_t^* = \text{argmax}_{1 \le i \le N}[\gamma_t(i)] \ for \ \forall t = 1 \dots N \quad (25)$$

After this, the observation data is trained and tested. Hence, this is the procedure for applying HMM on speech signals.

VII. CONCLUSION

This paper explained about the automatic speech recognition system that how it works and different speech variations that make the speech recognition more difficult. It also gives idea about the different techniques of speech recognition and find out the most efficient technique. While implementing speech recognition through hidden Markov model, basically we need to compute solution for three fundamental problems. This paper gives a mathematical understanding for solving the three fundamental problems that occurs during the application of speech signal. Also, it presented a brief review of literature on the work done in this field till now.

REFERENCES

1. Youhao Yu Research on Speech Recognition Technology and Its Application: 2012 International Conference on Computer Science and Electronics Engineering
2. J. I. Crabtree The Work of Edward Christopher Wente: Journal of The Society of Motion Picture Engineer 1935
3. E. E. DavidArtificial AuditoryRecognitionin Telephony:IBM Journal of Research and Development1958
4. D. B. Fry Theoretical Aspects of Mechanical Speech Recognition: Journal of the British Institution of Radio Engineers 1959
5. H. Olson and H. Belar Time Compensation for Speed of Talking in Speech Recognition Machines: IRE Transactions on Audio(Volume: AU-8 ,Issue: 3, May-June 1960)
6. Thomas Marill, (1961). Automatic Recognition of Speech. IRE Transactions on Human Factors in Electronics
7. M. Halle, K. Stevens 1962 Speech Recognition: A Model and A Program for Research: IRE Transactions on Information theory
8. Harry F. Olson 1962 Processing of Sound: Proceedings of the IRE
9. T. B. Martin, J.J. Talavage Application of Neural Logic to Speech Analysis and Recognition.

10. T. Sakai, S. Doshita The Automatic Speech Recognition System for Conversational Sound: IEEE Transactions on Electronics Computers 1963
11. W. Bezdel, H.J. Chandler Results of an Analysis and Recognition of Vowels by Computer Using Zero-Crossing Data
12. D.A. Bell; W. Bezdel ; H.J. Chandler Results of an analysis andrecognitionof vowels by computer using zero-crossing data:Proceedings of the Institution of Electrical Engineers1966
13. Charles C. TappertApplication of adaptive neural networks tospeech recognition:Sixth Symposium on Adaptive Processes1967
14. M. Lecours; J. Sparkes Adaptive Spectral Analysis for Speech-Sound Recognition: IEEE Transactions on Audio and Electroacoustics
15. R. Purton Speech Recognitionusing Autocorrelation Analysis: IEEE Transactions on Audio and Electroacoustics
16. R. De mori A descriptive technique for automaticspeech recognition: IEEE Transactions on Audio and Electroacoustics
17. S. Das A technique forspeechcoding using dynamic programming:ICASSP '80. IEEE International Conference on Acoustics, Speech, and Signal Processing
18. Rabiner, Lawrence R. "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE 77.2 (1989): 257-286.
19. Amodei, Dario, Sundaram Ananthanarayanan, RishitaAnubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper et al. "Deep speech 2: End-to-end speech recognition in english and mandarin." In International conference on machine learning, pp. 173-182. 2016.
20. Chan, William, et al. "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.
21. Xiong, Wayne, et al. "The Microsoft 2017 conversational speech recognition system." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.
22. Dubagunta, S. Pavankumar, SelenHandeKabil, and Mathew Magimai Doss. Improving Children Speech Recognition through Feature Learning from Raw Speech Signal. No. CONF. 2019.
23. Krishna, Gautam, Co Tran, Jianguo Yu, and Ahmed H. Tewfik. "Speech Recognition with no speech or with noisy speech." *arXiv preprint arXiv:1903.00739* (2019).

AUTHORS PROFILE

Kanika Bansalwalcurrently pursuing her master's degree from Lovely Professional University in the field of Robotics and Automation. She has completed her Bachelor's degree in the field of Electronics and Communication in 2017.

Kamaldeep Sharmais working as an Assistant Professor in Lovely Professional University,Punjab India. She obtained her Bachelor's degree in the field of Instrumentation in 2008 & Master's degree in 2010 with distinction in Instrumentation and Control Engineering. After completing her Masters studies, she is working in teaching field.During her study and teaching Profession she published 5 research papers in different referred journals and Conferences.

Anuj Jainis working as a Professor in Lovely Professional University, Punjab India. He obtained his Bachelor's degree in the field of Instrumentation in 2002 & Master's degree in 2005 with distinction in Electronics and Communication Engineering. After completing his Masters studies, he is working in teaching field. He Completed his Ph.D. in the field of Electronics and Communication at Mewar University, Chittorgarh, Rajasthan in 2016. During his study and teaching Profession he published 21 research papers in different referred journals and Conferences.10 PG(MTech) and 1 PhD completed under his guidance till now.