ASR System for Isolated Words Using ANN with Back Propagation and Fuzzy Based DWT

Sunanda Mendiratta, Neelam Turk, Dipali Bansal

Abstract*: Speech is the primary means through which human beings interact. Speech has become a way for Man Machine Interaction (MMI). The Speech Recognition (SR) systems have been widely used in smart phones to initiate searches or to type certain text messages, and in control devices to perform switch on or off functions etc. This system comprises three blocks: Pre-processing, Feature Extraction and Classification. The input speech signal is pre-processed to remove the noise and to convert it into a digital form for feature extraction. The feature extraction is a significant process during SR systems design because the features extracted form the basis for accurate recognition of the speech. Only a few features of this signal may be selected for classification purposes. For final recognition of the spoken word or the input signal, various optimization algorithms as classifiers are used. This paper presents an extensive literature review on SR Systems. The authors have attempted to do a brief survey to identify the progress in this field. The survey provides the reader with well-known methods used by previous researchers. It also compares the performance metrics for two ASR techniques developed by the authors. The first technique uses Artificial Neural Network with Back Propagation while the second uses Fuzzy based Discrete Wavelet Transform. It was found that the fuzzy based DWT system provided better results in terms of the performance metrics like accuracy, sensitivity, specificity and word error rate. The paper concludes by providing the reader with a direction of future scope in this research area.

Keywords: Speech Recognition (SR), Speech signal preprocessing, Feature extraction, Classification, Speech to text conversion, Man Machine Interaction

I. INTRODUCTION

In a man-machine interaction process, the Speech Recognition step which is also known as an Automatic Speech Recognition (ASR) is defined as the capacity of a computer/machine to recognize the input human voice signal and translate it into a corresponding text. This technology has been used in many sectors such as military, medicine, education and translation. The application of ASR systems is now seen in tasks that require an interface between human and machines. These several tasks include collection of query-based data which provides the latest travel information, inbuilt preprocessing and answering the phone calls through wireless network, predictions of the latest stock prices, dictation of voice, live weather reports and data entry[1]. Moreover, the data transfer to a computer through speech inputs is much faster than that through writing or typing using a keyboard. Just imagine the amount of time that would be saved when the text to be typed into the computer can be dictated. This would significantly speed up the processes involving documentations or writing emails etc. Such machines can be easily operated by people with little knowledge about computers and even the physically disabled persons. Many real-life ASR systems are available today and some of them are Speech Enabled Dialing, Global Positioning System, Voice enabled Google Searches, automatic call routing, Amazon's Alexa, Microsoft's Cortana, Google's Google Assistant and Apple's Siri etc. Speech Processing involves mathematical analysis and application of electrical signals from the acoustic pressure waves that are gathered from human vocalization for data storage and extraction of the result data. The study of Speech Processing includes various processes of analyzing speech, coding of speech, and enhancement of speech, synthesis of speech and finally recognition of speech. Speech analysis involves the production of speech using the properties of speech signal and the mathematical model. The signal properties include bandwidth, frequency, peaks in the spectrum and the envelope of the power spectrum. Speech coding aims to store information regarding a speech specific parameter for subsequent recovery. For retrieval in high quality, the speech signals may be stored or recorded. Speech enhancement improves the quality of the speech by removing the noise in the corrupted signal using several algorithms, whereas, the Speech synthesis involves artificial production of human speech. Speech recognition involves conversion of speech signals to words by using computer algorithms. Therefore, speech synthesis and speech recognition are two opposite processes. The remaining paper has been categorized into following sections: Section 2 presents the literature survey and history of the various ASR techniques, Section 3 discusses the architecture of ASR system, Section 4 compares the performance metrics of two ASR systems developed by the authors and gives a detailed performance analysis, and Section 5 concludes the paper by discussing the future scope for other researchers. It also discusses the methods for recognition accuracy improvement.

II. LITERATURE REVIEW

In the last eight decades, a tremendous growth in the Speech Processing technology has been observed. A lot of researchers have shown interest in the field of ASR technology due to the several reasons such as advanced technological curiosity regarding the process mechanism to the desire of getting things automated and making it simpler through the human-machine interface.

Revised Manuscript Received October 05, 2019
***** Correspondence Author
Sunanda Mendiratta*, Department of Electronics Engineering, J. C. Bose UST, Faridabad, India.
Neelam Turk, Department of Electronics Engineering, J. C. Bose UST, Faridabad, India.
Dipali Bansal, ECE Department, FET, Manav Rachna International Institute of Research and Studies, Faridabad, India.

This technology evolved from the speech analysis and synthesis systems from 1940. In the 1930s, a speech synthesizer has been developed by Homer Dudley and named it as VODER (Voice Operating Demonstrator) in the Bell Telephone Laboratories. The model has been presented at the World Fair in the city of New York (1939) and the developed model is certified as an important milestone for the future speaking machines. The VODER was then modified by addition of certain electrical circuits for the artificial production of speech. This modified version was called the VOCODER [2] or the voice coder. This device was voice-controlled and was fundamentally different from the VODER, which was controlled by keys and pedals. However, these early works were based on speech production or synthesis rather than speech recognition. It was only in the 1950s that the research focus shifted from the Speech Synthesis to the Speech Recognition systems. The first SR system called the 'Audrey' system was implemented by Davis along with his subordinates Biddulph, and Balashek in 1952 century at Bell Laboratories. An isolated recognition of voice in digital format is observed for a single speaker on the basis of the acoustic phonetics. But, it was meant to recognize digits only in a noise free environment. Around the year 1960, various Isolated Speech Recognition systems were developed with small vocabularies of about 10–100 words. In 1962, IBM developed a machine that could understand 16 words of English speech. It was called the "shoebox" and was presented at the technology world fair. Although, this technology was speaker dependent.

The speech recognition systems developed in the following decade saw an increase in the vocabulary size to about 1000 words, until in the 1980s large vocabularies of more than 1000 words could be recognized. This decade saw fast progress in the speech recognition systems. AT & T Bell Labs started the production and manufacturing of independent speech recognition-based speakers by deploying cluster-based techniques and extracting independent voice recognition patterns. In the 1980s, isolated words were connected and new recognition systems came up. These were based on the algorithms that joined isolated words to form connected words being used for recognition. It was during this era that Hidden Markov Models (HMM) [3] were used for Speech recognition. In this paper by Rabiner et al., have presented an introduction to the theory of Markov models and have showed how they used these models for speech recognition problems. The authors used HMM to build an isolated word recognizer. Currently, al-most all the recognizers are based on the statistical framework of Hidden Markov Models. The Neural Networks have been utilized since 1980's for solving classification problems in speech recognition systems. With the substantial developments made in the sub-sequent decades, ASR technology has moved from speaker dependent to speaker independent systems.

In the past 1990s, several unconstrained speech models and constrained work syntax models were used as a setup to build huge vocabulary systems for better understanding and continuous speech recognition. Major advances during this tenure were the techniques for learning the randomness of the speech, statistical analysis and understanding of acoustic and language models. The finite state transducer network was also presented along with the FSM Library. Then during implementation, the techniques for the reduction in size of FSM library for productive execution of huge vocabulary language under-standing frameworks were developed. The FSM (Finite state machine) library is integrated with finite state arrangement approach in a unified transducer system along with the weight search. It has been a noteworthy segment of all modernized language recognition and understanding frameworks.

The next fifteen years have been very fruitful for the SR systems. The size of the vocabulary became infinite. Recognition rates also improved for real time speech recognition problems. The ASR is still considered as a standard classification problem. It can identify sequences of words from speech wave forms. However, it had several issues that prevented it from achieving the desired satisfactory performance. These include multi model recognition, multilingual recognition and noisy environment. Weiner filtering, spectral subtraction or windowing can be used for noise removal and enhancement of the speech. The Gaussian mixture model (GMM) and HMM are widely used for acoustic modeling of the speech inputs [4]. The key technologies developed in the last decade were the ANN, Deep Neural Networks (DNN), etc. These advancements have provided the way for today's speech recognition systems with infinite vocabulary size along with spontaneous speech recognition. A relatively new classification technique of Support Vector Machines for emotion recognition during ASR has been explained by authors [5]. SVM's provide a low cost solution to the classification of high dimensional vectors. The speech recognition systems have grown from template matching to HMMs, from filter banks to cepstral features, from smaller vocabularies to large vocabularies, from the speaker dependent technology to speaker independent technology. Due to the shorter computation time than the other systems, the ANNs can be used to produce an avatar system with real time speech talks. These neural networks deliver superior performance. Although, it takes longer time to train it, particularly when the ANNs have multiple hid-den layers. Moreover, the process through which the ANNs were initialized greatly affect the performance of these networks. Therefore, deep neural networks were preferred for large quantity of unlabeled data [6]. These developments have been listed in Table 1.

In 2006, deep learning algorithms came out as an upcoming research area in machine learning. These algorithms were capable of improving the feature extraction and transformation. Various scientists have found that the DNNs can perform better than the GMMs for speech recognition on acoustic modeling or for performing modeling data correlation[7]. The advances on the computer hardware and algorithms have allowed the researchers to train the neural network in end to end fashion. These neural networks can now work without lesser human intervention and can deliver superior performance. The introduction of neural networks has benefited the ASR to a greater extent.

Table- I: Developments in the Speech Recognition Technology.

Year	Vocabulary size	Type of Speech Recognition	Methods	Key Technology
1960s	Small	Isolated Words	Simple Speech sounds, Phonetic properties	Analysis using Filter Banks
1970s	Moderate Size (100 - 1000 words)	Isolated words, Connected Digits	Template based	Pattern recognition, Linear Predictive Coding, Zero Crossing analysis and Speech segmentation
1980s	Large	Connected Words	Statistical – based	HMM, Modelling for Stochastic speech
1990s	Large	Continuous Speech	Grammar based Sentences.	FSM, Learning by Statistics
2000 -2005	Very Large	Spontaneous speech	Multimodal Dialog in HCI, TTS	Machine learning; ANN
2006-2010	Infinite Vocabulary size	Real time dialogue Speech, Robust speech, Multimodal speech	Variation Bayesian estimation model, Mixed - initiative dialog;	Deep Neural Networks (DNN), Gaussian mixture hidden Markov model (GMHMM)
2011-2015	Infinite Vocabulary size	Automatic language identification	Multimodal, Acoustic and Language modelling	DNN, GMM, DNN-HMM
2016-till date	Infinite Vocabulary size	End-to-end automatic speech recognition, On-device speech recognition	Kernel Approximation, Acoustic and Language modelling	Multi-layer Perceptron (MLP) neural network, Dynamic MLP, DNN, Kernel acoustic model

A SR system which can directly transcribe the audio input into text without using any intermediate phonetic representation has been developed. This system does minimal pre-processing and can be applied to data sets that do not require language models [8]. In 2015, automatic language identification has been done using the DNNs [9]. A multitask speech recognition technique using DNN has been developed recently to enhance the performance of the low-resource ASR. This method does not rely on any additional language resources [10]. The DNNs have been used designing automatic feature extraction system for audio inputs, speaker adoptive training for acoustic modeling [11] and for tracking the dialog state [12]. It has also been used for cross language knowledge transfer [13]. In a comparative review on DNN and HMM methods for SR systems, the authors tried to reduce the mathematical steps required in designing the effective mobile device speech recognition system. The performance analysis for recognition resulted that the HMM methods give better results than the Dynamic MLP. However, the computation time for HMM was quite long as HMMs work on isolated words. DMLP has computation processing speed faster and reduced time that is around 9000 % than HMM for each pulse second of the speech data. Although, as the network size increases for DMLP, the computation time will increase. But this increase is significantly lower than an HMM system [14].

III. ARCHITECTURE OF SPEECH RECOGNITION SYSTEM

Speech Recognition, as the name suggests, is a method in which the human speech is converted into text by a computer/machine in order to have a Human Computer Interaction (HCI). The input to the system is set of spoken words (Isolated/Connected/Real time Speech) and the output is the written text, which the computer/machine understands. So the basic process is conversion of voice to text. The speech recognition system

comprises of three sub processes, i.e., signal preprocessing, feature extraction and classification. The

basic architecture of ASR system is shown in Figure 1.

A. Signal Preprocessing

This is performed for removal of the noise component of the speech signal. The source of the noise could be the environment in which the ASR system is placed, or it

Fig. 1 Basic Blocks of Speech Recognition System.

may be due to the fans, typewriters, computers and other background conversations. These unavoidable and undesirable sources of noise influence the working of ASR system. Thus preprocessing of the input speech signals is the most important step in this system. This step also involves pre-emphasis of the speech signal, Voice Activity Detection (VAD) and the estimation of spectral envelope.

During the preprocessing, initially the input speech signal captured from the data storage system is forwarded to moving average high pass filter and background noise observed is removed. The difference equation representing this filter is given in equation (1):

$$S(n) = X(n) - 0.95X(n-1);\ 1 \le n \le L \quad (1)$$

Where, *S (n)* signifies the output signal, *X (n)* denotes the input speech, *and L* defines the length of the audio frame.

After the removal of the background noise, the speech signals are re-framed and forwarded through the window. This preprocessed signal is used for feature extraction and processing. Since, the length of every speech signal varies, its framing has to be dynamic. Only the speech part of the signal is retained, so that the number of time frames for comparison would decrease leading to faster computation rates for speech recognition. This will also improve the recognition rate. Voice activity detection or End point detection techniques are used for this. VAD is preferred for speech signals in time domain. It is a simple and popular technique. So, the speed of the recognition system will increase.

The audio frames are passed through the window to prevent spectral leakage. For speech recognition problems, Hamming window [15][16][17][18] is the preferred window, which is represented by equation (2). The output is found using equation (3).

$$W(n) = 0.54 - 0.46(2 * \pi * n/N - 1);\ 0 \le n \le N - 1 \quad (2)$$

$$Y(n) = S(n) \times W(n) \quad (3)$$

Where, N denotes the sample set of individual frames, *Y (n)* is the output preprocessed signal, and *S (n)* is the input signal. This output is used for further processing and extraction of efficient features.

B. Feature Extraction

Feature extraction is defined as a process of extracting features from the speech signal. These features represent the speech signal in a compact form. The process of feature extraction involves identification of the components of the input signal that correspond to the linguistic content of the signal. Since, there exists a solid connection amongst every phoneme of VT shape and the spectra of the input, spectral envelope is estimated for ASR. Entire energy of the system is the least difficult element to concentrate, and it is exceptionally valuable for speech/language process detection and low-rate coding applications. Furthermore, its utilization for ASR is less immediate as the energy levels are influenced by transmission conditions. For example, cepstral mean standardization process is used to remove the mean amplitude, however neighborhood adequacy changes ought to be exploited to segregate weak and strong phonemes from the mobile communication system [19].

The output of the hamming window is obtained in the form of spatial time domain. Since, working in the frequency domain is found to be highly convenient rather than representing in the time domain. The time domain output of the hamming window is changed to the frequency domain. Fast Fourier Transform (FFT) technique is performed to obtain the frequency domain signal. Feature Extraction is performed on this frequency domain signal. FFT is defined as high speed processing algorithm using the DFT (Discrete Fourier Transform), which is given by the equation (4) as follows:

$$Y(k) = \sum_{n=0}^{N-1} y(n)\, e^{-j2\pi kn/N} \quad (4)$$

Where, *Y (k)* is the frequency domain signal. Since the speech signal is divided into frames, FFT is calculated for each frame. However, the FFT spectrum range is very wide and a straight scale is not followed by the speech signal, a scale based on pitch comparisons known as Mel scale (Mel from melody) is utilized. It is a linear scale up to 1000 Hz and logarithmic scale above 1 kHz. For each frequency tone, a different frequency based on Mel scale is calculated. The Mel Frequency can be calculated by utilizing equation (5).

$$Mel\ (f) = 2595 * \log_{10}(1 + f/700)$$

Where, *Mel (f)* defines the Mel frequency spectrum obtained from the original input signal frequencies *'f'*. The final step is to extract the Mel Frequency Cepstral Coefficients (MFCC). Short term analysis is used to calculate the MFCC of the speech signal. These MFCC feature vectors are extracted and altered into a set of acoustic vectors.

C. Classification

After feature extraction, the features obtained are classified according to a set of available features. Many classification techniques are available for ASR systems. The Dynamic Time Warping (DTW) based technique and machine learning based Artificial Neural Networks (ANN) are the mainly used classifiers.

For the recognition by DTW [17], the data in the database of the speech signal samples are one by one compared with the tested speech sounds. Subsequently, huge number of tested speech sounds are observed in the storage database, the more will be the computation time for recognition costs. The DTW is regularly utilized for estimating likenesses amongst two transient successions which may change with respect to time or speed. It is likewise used to get adapted with distinct speeds of speaking, speaker recognition, and recognition of online signature. Additionally, it was seen that it can be utilized as a part of partial shape coordinating applications. The DTW is a strategy that can be used to compute an ideal match amongst two given groupings with specific confinements. It can adjust the articulations legitimately and ascertains minimum separation amongst two utterance expressions or the sample examples. It is an effective strategy to take care of the time arrangement issue [20].

Although, DTW has more recognition time, the speech recognition based on ANN is faster. The training of ANN is the only time-consuming part. However, once trained, it requires low computational resources in terms of memory requirements and execution speed, as compared to other methods. Hence, the classification technique of ANN gives faster recognition rates. ANN are also called multi-layer perceptron (MLPs). The neural network works on the basis of number of neurons present and the typical structure comprises of input, hidden and output nodes. Figure 2 shows a neural network with input nodes, '*n*=2', hidden nodes, '*l*=3' and output node, '*k*=1'.

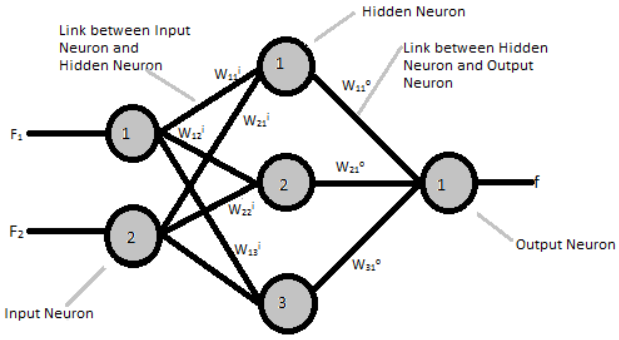

Fig. 2. Neural Network Structure with two input, one output and three hidden neurons.

The features obtained in the Feature Extraction technique is used as an input to the ANN classifier. The captured features are considered and processed as input where the networks get trained with the genuine sample data set and a produces equivalent text as an output. ANNs shows higher accuracy rate for static pattern recognition.

Another class of classifiers are Deep Neural Networks (DNNs). DNN based acoustical modeling is the technology being used by the researchers for SR systems since the last 4–5 years [21]. The latest technique is to utilize DNN in various structural format and as a combined approach for the part of machine learning frameworks being utilized and effectively developed for ASR frameworks. However, optimizing deep learning for different applications is still an open challenge for the researchers. The DNNs are ANN with more layers and with better performance. They attempt to extricate better features of ASR at lower system layers (those close to the input) trailed by classification and grouping at higher layers. The fundamental MFCC or PLP features are modified through the application of multi-layer neural system that yields the output as PDF. Sometimes such networks have relatively few nodes as a hidden layer. The system may have isolated parallel paths which are trained autonomously and consecutively with a few concealed hidden levels[19]. The classifier gives the desired output in the form of text.

IV. PERFORMANCE EVALUATION OF ASR SYSTEM

A. Parameters for checking the performance of an ASR system

The check parameters of SR system depend on its speed and accuracy. The speed in turn depends on the computation times whereas accuracy depends on the percentage of the recognized words. Accuracy is the word recognition rate (WRR), which is given by equation (6) as follows:

$$WRR = H - 1/N$$

Where, H denotes a set of correctly recognized words and N is a set of words.

Word Error Rate measures performance of the system and is given by equation (7),

$$WER = 1 - WRR \quad (7)$$

The following terms explain how the test samples are recognized. They are specificity and sensitivity. The False positive rate is defined as the rate of existence of positive test results, given by

$$FPR = FP/(FP + TN) \quad (8)$$

Specificity is defined as the ability of a method to recognize the negative samples correctly.

$$Specificity = TN/(TN + FP)$$

Sensitivity is defined as the ability of a method to recognize the positive samples properly.

$$Sensitivity = TP/(TP + FN)$$

Where TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

Augmentation of general input features along with ANN provides enhancement of the recognition accuracy of an ASR system. The recognition accuracy is an important performance gradient for a speech recognition system.

B. Performance Analysis and comparison of existing ASR system

The authors in [22] and [23] have proposed an isolated ASR and a fuzzy based ASR using two different methods for the feature extraction process. The first method of isolated word recognition uses Artificial Neural Network (ANN) with Back Propagation (BP), while the other used a fuzzy based Discrete Wavelet Transform (DWT) to extract the features. In the first case recorded audio input is preprocessed to reduce the noise and then the common features (viz. sampling point, word length, pitch) and statistical features (viz. variance, mean, kurtosis, skewness, entropy) are extracted. Later these features are optimized to train the classifier. Finally, the input spoken word is displayed as text output. However, in the second case the input signal undergoes preprocessing where the signal is sampled to produce frames by Hamming windowing process. The signal noise is then removed with the help of harmonic decomposition process. Then the specified features are extracted through DWT and from these, only the required features are chosen by the fuzzy inference system (FIS). The optimally extracted features are then used to train the ANN where neural network optimization is done using cuckoo search (CS) optimization algorithm.

Both the systems are implemented on working platform of MATLAB 2013a and the results have been compared and presented here. The ASR classifier performance has been measured in terms of different performance metrics illustrated in the table 2.

Table- II: Performance Comparison of Isolated Word ASR and Fuzzy based ASR.

Performance Metric	Isolated ASR	Fuzzy based ASR
Sensitivity	50%	95%
Specificity	74%	50%
False Positive rate (FPR)	26%	50%
Recognition accuracy	62%	95%
Word error rate	65%	5%

V. CONCLUSIONS, DISCUSSIONS AND FUTURE SCOPE

The intelligent agents or the machines with the ability to recognize and understand the human speech still need to undergo a lot of improvements to make the conversation between a human and a machine more meaningful and insightful. In spite of great progress in the field of speech recognition, there remains a vast difference in the recognition procedures of a man and a machine.

From the performance analysis of the two techniques discussed, an inference could be drawn that the Fuzzy based ASR with CS-ANN has achieved the recognition accuracy of 95% with an error rate of just 5%. The accuracy is very good when compared to the IWR system using ANN with Back Propagation.

The future researchers will have to face challenges while improving the recognition accuracy to achieve maximum levels. They will have to work upon techniques for optimizing deep learning for different applications. The SR technology has much scope for research in the area of man machine interaction. The methodologies explained in this paper may be used to produce a hybrid model to improve the recognition rates further. This paper may help the researchers with the material and methods to develop future recognition systems. It may pave a path to produce a robust speech recognition system.

REFERENCES

1. M. A. Anusuya and S. K. Katti, “Speech recognition by machine: A review,” Int. J. Comput. Sci. Inf. Secur., vol. 6, no. 3, pp. 181–205, 2009.
2. H. Dudley, “The Vocoder—Electrical Re-Creation of Speech,” J. Soc. Motion Pict. Eng., vol. 34, no. 3, pp. 272–278, Mar. 1940.
3. L. Rabiner and B.-H. Juang, “An introduction to hidden Markov models,” ASSP Mag. IEEE, vol. 3, no. January, p. Appendix 3A, 1986.
4. W. Liu, Z. Wang, X. Liu, Y. Liu, and F. E. Alsaadi, “A survey of deep neural network architectures and their applications,” Neurocomputing, vol. 234, pp. 11–26, 2016.
5. S. Mendiratta, N. Turk, and D. Bansal, “Recognition of Human Emotional States during Automatic Speech Recognition,” Int. J. Innov. Res. Comput. Commun. Eng., vol. 4, no. 4, pp. 301–307, 2016.
6. J. Padmanabhan and M. J. Johnson Premkumar, “Machine Learning in Automatic Speech Recognition: A Survey,” IETE Tech. Rev., vol. 32, no. 4, pp. 240–251, Jul. 2015.
7. L. Deng et al., “Recent advances in deep learning for speech research at Microsoft,” in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 8604–8608.
8. A. Graves and N. Jaitly, “Towards End-To-End Speech Recognition with Recurrent Neural Networks,” in Proceedings of the 31st International Conference on Machine Learning, 2014, vol. 32, no. 2, pp. 1764–1772.

9. J. Gonzalez-Dominguez, I. Lopez-Moreno, P. J. Moreno, and J. Gonzalez-Rodriguez, "Frame-by-frame language identification in short utterances using deep neural networks," Neural Networks, vol. 64, pp. 49–58, Apr. 2015.
10. D. Chen and B. Mak, "Multi-task Learning of Deep Neural Networks for Low-resource Speech Recognition," IEEE/ACM Trans. Audio, Speech, Lang. Process., pp. 1–1, 2015.
11. Y. Miao, H. Zhang, and F. Metze, "Speaker Adaptive Training of Deep Neural Network Acoustic Models Using I-Vectors," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 11, pp. 1938–1949, Nov. 2015.
12. M. Henderson, B. Thomson, and S. Young, "Deep Neural Network Approach for the Dialog State Tracking Challenge," Association for Computational Linguistics, 2013.
13. J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 7304–7308.
14. M. K. Mustafa, T. Allen, and K. Appiah, "A comparative review of dynamic neural networks and hidden Markov model methods for mobile on-device speech recognition," Neural Comput. Appl., vol. 31, no. 2, pp. 891–899, Feb. 2019.
15. A. S. Bhalerao and V. B. Malode, "Implementation of automatic speaker recognition on TMS320C6713 using MFCC," 2013 Int. Conf. Comput. Commun. Informatics, ICCCI 2013, pp. 4–7, 2013.
16. S. D. Daphal and S. K. Jagtap, "Noise Robust Novel Approach to Speech Recognition," 2014 Int. Conf. Electron. Syst. Signal Process. Comput. Technol., pp. 289–294, 2014.
17. A. Pramanik and R. Raha, "Automatic Speeech Recognition using Correlation Analysis," World Congr. Inf. Commun. Technol., pp. 670–674, 2012.
18. L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," J. Comput., vol. 2, no. 3, pp. 138–143, 2010.
19. D. O'Shaughnessy, "Acoustic analysis for automatic speech recognition," Proc. IEEE, vol. 101, no. 5, pp. 1038–1053, 2013.
20. P. P. S. Pratik K. Kurzekar, Ratnadeep R. Deshmukh, Vishal B. Waghmare, "A Comparative Study of Feature Extraction Techniques for Speech Recognition System," Ijirset, vol. 3, no. 12, pp. 18006–18016, 2014.
21. M. Sarma, "Speech recognition using deep neural network – recent trends," Int. J. Intell. Syst. Des. Comput., vol. 1, no. 12, pp. 71–86, 2017.
22. S. Mendiratta, N. Turk, and D. Bansal, "Isolated word recognition system for speech to text conversion using ann," Inst. Integr. Omi. Appl. Biotechnol., vol. 7, no. 11, pp. 78–91, 2016.
23. S. Mendiratta, N. Turk, and D. Bansal, "Fuzzy based selection of DWT features for Automatic speech recognition system for man machine interaction with CS-ANN Classifier," Off. J. Inst. Integr. Omi. Biotechnol., vol. 7, no. 11, pp. 222–240, 2016.

AUTHORS PROFILE

Sunanda Mendiratta received the B.Tech degree in Electronics and Communication Engineering from Kurukshetra University, Kurukshetra, Haryana, India in 1999. And her M.Tech degree in the same discipline in 2011 from Manav Rachna International University, Faridabad, NCR, India. She has an experience of 8 years of teaching in various prestigious engineering colleges in Faridabad, NCR, India. She is currently working towards Ph.D. degree in J. C. Bose University of Science and Technology, Faridabad, NCR, India. Her research interests include digital filtering techniques, artificial neural networks and speech processing.

Neelam Turk received B.E degree from North Maharashtra University Jalgoan, India in 1998. She did her M.Tech. in Electronics and Communication Engg. from National Institute of Technology, Kurukshetra (India) in 2002. She received Ph.D. degree in Electrical Engg. From National Institute of Technology, Kurukshetra (India) in 2011. Currently she is working as Associate Professor in Electronics Engineering Department with the J. C. Bose University of Science and Technology, Faridabad, Haryana, India. Her research interest include MCSA, signal processing, Speech Processing, wireless communication.

Dipali Bansal is a doctorate in Biomedical Instrumentation and Bio signal processing from Jamia Milia University, New Delhi and an upcoming and young scientist. She is a professor in MRIIRS. Her research interests lie primarily in the areas of analysing human physiological signals and developing easy acquisition systems for these bio-signals using PC based systems. Her parallel areas of interest are development of algorithm in MATLAB for digital filtering, deriving HRV and implementing them on Digital Signal Controllers to achieve a compact solution to home health care. She is keen on latest microelectronics technology in developing implanted biomedical devices and other medical products using Smart Sensors and Integrated Microsystems.