Paper

Speech-Based Vehicle Movement Control Solution

Gurpreet Kaur[1], Mohit Srivastava[2], and Amod Kumar[3]

[1] *UIET, Panjab University, Chandigarh, India,* [2] *CEC Landran, India,* [3] *NITTTR, Chandigarh, India*

https://doi.org/10.26636/jtit.2021.149820

Abstract—The article describes a speech-based robotic prototype designed to aid the movement of elderly or handicapped individuals. Mel frequency cepstral coefficients (MFCC) are used for the extraction of speech features and a deep belief network (DBN) is trained for the recognition of commands. The prototype was tested in a real-world environment and achieved an accuracy rate of 87.4%.

Keywords—deep belief networks, mel frequency cepstral coefficients, speech recognition.

1. Introduction

Speech signal processing is an important research field due to the wide variety of applications it may be useful for. It may be used in banking systems, forensics, hospitals, the military, or in day-to-day activities. Regardless of the domain, people desire intelligent and smart systems facilitating their efforts. Many advancements in this field have led to the development of such smart systems as Apple's Siri, Google Assistant, Amazon Alexa, etc. [1], [2]. Voice operated vehicles may also be used to aid the movement of elderly or handicapped individuals. Unfortunately, voice operated wheelchairs are not quite popular yet, because of numerous challenges in speech recognition [3]–[5]. Every single person has their own way of speaking. Their voice may be characterized by a different accent and variations may exist in the use of language, emotions and the environment. All these parameters exert a great impact on speech recognition.

Considerable progress has been made in various domains, such as education or entertainment. However, much less effort has been put into harnessing speed recognition achievements in the process of physical rehabilitation of patients. All existing feature extraction methods have showed a good recognition rate only in a clean environment. Results obtain in noisy environments still require considerable improvement [6]. The recognition rate suffers when the system is used in a real-world scenario. There is a big difference in the operation of the speaker and speech recognition mechanism when it relies on a stored database and on a real-world database. All existing speaker and speech recognition systems perform efficiently when operating based on a stored database. But in real-world applications, their performance is affected adversely because of the wide variety of speaking styles and background noise [7]. In order to deal with all these challenges, artificial intelligence seems to be the best proposal, as it allows to design systems capable of dealing with variations stemming from speaking styles, gender, language and background noise. Trained neural networks are capable of handling all these details. The problem of speech recognition and speaker recognition has been dealt with by many researchers [8]. Technology-related progress observed thus far aims towards implementing numerous speech recognition systems. Speech recognition depends directly upon feature extraction and classification methods [9]. Some authors have already compared the feature extraction methods known [10]. According to the results of their analyses, MFCC is a solution that is best suited for the task at hand. Classification-related approaches may be template-based [11], stochastic-based [12] or artificial neural network-based systems [13]–[15]. In template-based techniques, voice samples were stored in a database and the comparison was made between the database and the uttered word. There were many popular techniques, such as distance calculation with the use of the dynamic time warping algorithm. However, systems of this type were not capable of providing correct results in real-world applications and, hence, their performance degraded in noisy environments.

Later, stochastic-based models were popular in speech recognition. These were based on probability models and could handle incomplete and uncertain input data. Hidden Markov models were best for dealing with variable input data. In the 1970s, the US Department of Defense sponsored numerous projects focusing on stochastic models, for instance Dragon. The technique was widely accepted all over the world before the emergence of artificial intelligence. Artificial intelligence is a knowledge-based system [16] that started to gain in importance after implementation of numerous neural network algorithms. Previously, neural networks were based on a few hidden layers. They were not capable of dealing with the variability of input data. Nowadays, however, deep learning networks [17] are effectively used for this particular purpose. There are many learning-based algorithms used by researchers, for instance convolutional neural network (CNN) [18], multilayer perceptron (MLP), probabilistic neural network (PNN) [19]–[24]. However, the performance of all those systems deteriorates in noisy environments. Learning with the use of noisy data helps improve the system. Many techniques have been implemented by researchers to make the systems easy to use for humans.

2. System Concept

The proposed system is divided into a software and a hardware module. Speech samples are captured with the use of a microphone. Then, all samples are preprocessed in preparation for applying the feature extraction technique. The MFCC feature extraction method is deployed and the extracted features are then used for training a deep neural network (DNN). DNN is emerging field in speech recognition field and its performance exceed that of other methods. In the DNN approach, a deep belief network (DBN) is used. Many learning techniques are available. In this type of work, unsupervised learning contrastive divergence (CD) is used. Matlab is used to implement the application. In the hardware module, a prototype is made that consists of an RF receiver, a microcontroller, a motor driver and a DC gear motor. The command from Matlab software is received using an RF data modem, via the serial communication protocol. The system works in the 2.4 GHz band its rate is adjustable between 9600 and 115200 bps for direct interfacing with the MCU.

2.1. MFCC Features

Speech signals are converted into the frequency domain using FFT. The FFT output contains a lot of data that is not required. In order to calculate the energy level at each frequency, a mel scale analysis is performed using mel filters. Then energy is calculated and after that logarithmic of filter bank energies is taken. This operation is conducted in order to match the features closer to human hearing. Finally, DCT of the log filter bank energies is taken to decorrelate the overlapped values. Only the first 13 coefficients are selected, known as MFCC features. This is because higher feature numbers degrade the recognition-related accuracy of the system [25]. Those features do not carry any speaker and speech-related information. Mathematically, this may be explained using the vocal tract system. The vocal tract articulation equivalent filter is shown in Eq. (1):

$$s(\Omega) = g(\Omega)\, h(\Omega) \ . \tag{1}$$

The equivalent logarithm of $s(\Omega)$ is:

$$\log|s(\Omega)| = \log|g(\Omega)| + \log|h(\Omega)| \ . \tag{2}$$

In this method, the logarithm of the spectrum is taken and then its inverse Fourier transform is found. The cepstrum $C(\partial)$, or cepstral coefficients, is the inverse Fourier transform of $\log|s(\Omega)|$:

$$C(\partial) = F^{-1}\log|s(\Omega)| = F^{-1}\log|g(\Omega)| + F^{-1}\log|h(\Omega)| \ . \tag{3}$$

It represents the vocal tract's parameters corresponding to the phonemes of sound.

2.2. Deep Neural Network

A deep neutral network (DNN) has many hidden layers. The inspiration is taken from the visual cortex (part of brain). Information is processed in the sequence of regions. So, a neural network may be modeled as a multilayer network consisting of low to high level features. Training is the major issue in DNNs, as optimization is a more difficult step. Performance of the system degrades because of under fitting and over fitting. Under fitting is caused by a vanishing gradient problem and over fitting stems form high variance and low bias situations. The unsupervised pre-training approach may be the solution to be used while training DNNs. Training is performed one layer at a time, from the first to the second to the last. Features are fed to the first hidden layer, then the second layer takes the combinations of features from the first layer. This process is repeated until the last layer is reached. Once all layers have been pre-trained, supervised training is performed for the entire network. This stage is known as fine tuning. This means that deep training can be performed in two steps: pre-training the network and fine tuning, using supervised training. Restricted Boltzmann machines (RBMs) are the building blocks of deep networks. It is an unsupervised, undirected graphical model. It is called restricted, because there is no link between the units of each layer, as shown in Fig. 1.

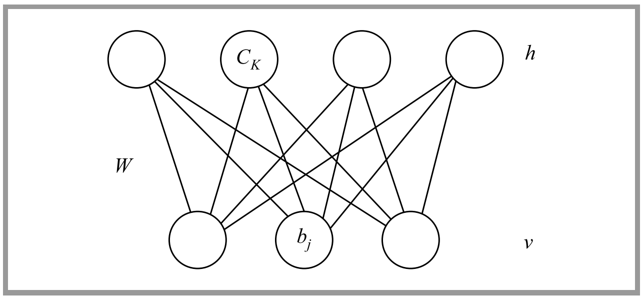


Fig. 1. Undirected graphical RBM model.

The model is of the energy-based variety and its distribution is given by:

$$P(v,h) = \frac{1}{Z}\mathrm{e}^{(-E(v,h))} \ . \tag{4}$$

$E(v,h)$ is the energy function and is defined as:

$$E(v,h) = \ -b^T v - c^T h - v^T W h \ , \tag{5}$$

where b, c are vectors and W is the matrix – these are the parameters of the model. The partition function is defined as:

$$Z = \sum_v \sum_h \mathrm{e}^{(-E(v,h))} \ , \tag{6}$$

$$\begin{aligned} E(v,h) &= \ -b^T v - c^T h - v^T W h \\ &= -\sum_k b_k v_k - \sum_j c_j h_j - \sum_j \sum_k c_j h_j W_{jk} \ . \end{aligned} \tag{7}$$

RBM is trained on binary data. Conditional distribution from the joint distribution is:

$$P(\frac{h}{v}) = \frac{P(h,v)}{P(v)} = \frac{1}{P(v)}\frac{1}{Z}\mathrm{e}^{b^T v+c^T h+v^T Wh}$$
$$= \frac{1}{Z\prime}\mathrm{e}^{c^T h+v^T Wh} \ . \tag{8}$$

This expression can be written in a scalar form as:

$$P(\frac{h}{v}) = \frac{1}{Z\prime}\mathrm{e}^{\sum_{j=1}^{n} c_j h_j+\sum_{j=1}^{n} v^T W_{:j} h_j}$$
$$= (1/z')\prod_{j=1}^{n}\mathrm{e}^{c_j h_j+v^T W_j h_j} \ . \tag{9}$$

The distribution over the individual binary h_j is given as:

$$P(h_j = \frac{1}{v}) = \frac{P(hj=1,v)}{P(hj=0,v+P(hj=1,v)}$$
$$= \frac{\mathrm{e}^{c_j+v^T W_{:j}}}{\mathrm{e}^{\{0\}}+\mathrm{e}^{c_j+v^T W_{:j}}} = \mathrm{sigmoid}\left(c_j+v^T W_{:j}\right) \ , \tag{10}$$

$$P\left(\frac{h}{v}\right) = \prod_{j=1}^{n}\mathrm{sigmoid}\left(c_j+v^T W_{:j}\right) \ , \tag{11}$$

$$P\left(\frac{v}{h}\right) = \prod_{i=1}^{d}\mathrm{sigmoid}\left(b_i+W_{i:}h\right) \ . \tag{12}$$

Parameters of the models are learnt by taking the log likelihood given as:

$$l(W,b,c) = \sum_{t=1}^{n}\log p\left(v^{(t)}\right) = \sum_{t=1}^{n}\log\sum_{h}p\left(v^{(t)},h\right)$$
$$= \sum_{t=1}^{n}\log\sum_{h}\mathrm{e}^{-E\left(v^{(t)},h\right)} - n\log Z$$
$$= \sum_{t=1}^{n}\log\sum_{h}\mathrm{e}^{-E\left(v^{(t)},h\right)} - n\log\sum_{v,h}\mathrm{e}^{-E(v,h)} \ . \tag{13}$$

For maximizing the likelihood, derivative of the log likelihood is taken as:

$$\theta = b,c,W \ , \tag{14}$$

$$l(\theta) = \sum_{t=1}^{n}\log\sum_{h}\mathrm{e}^{-E(v^t,h)} - n\log\sum_{v,h}\mathrm{e}^{-E(v,h)} \ , \tag{15}$$

$$\nabla_\theta(\theta) = \nabla_\theta\sum_{t=1}^{n}\log\sum_{h}\mathrm{e}^{-E(v^t,h)} - n\nabla_\theta\log\sum_{v,h}\mathrm{e}^{-E(v,h)}$$
$$= \sum_{t=1}^{n}E_{p(h/v^t)}[\nabla_\theta(-E\left(v^{(t)},h\right))] - nE_{p(h/v)}|[\nabla_\theta(-E(v,h))] \ . \tag{16}$$

The second half of Eq. (16) i.e. the expectation of h given v, is difficult to learn. Therefore, the contrastive divergence algorithm is used, where the expectation is calculated by a point estimate using Gibb's sampling.

2.3. Deep Belief Network

In a deep belief network (DBN), RBMs are stacked together to form a multilayer network. A graphical model of DBN is shown in Fig. 2. In this model, there are three hidden layers with one input layer. Training is performed with the first layer using the data and then freezing the first layer's parameters. Then, the second layer is trained using the output of the first layer, constituting unsupervised input for the second layer. This process is repeated until the last layer is reached. The following steps are taken for calculating the distribution:

$$P(x) = \sum_{h^{(1)}}p(x,h^{(1)}) \ , \tag{17}$$

$$P\left(x,h^{(1)}\right) = p(\frac{x}{h^{(1)}}\sum_{h^{(2)}}p(x,h^{(1)}) \ , \tag{18}$$

$$P\left(h^{(1)},h^{(2)}\right) = p(\frac{h^{(1)}}{h^{(2)}}\sum_{h^{(3)}}p(h^{(2)},h^{(3)}) \ . \tag{19}$$

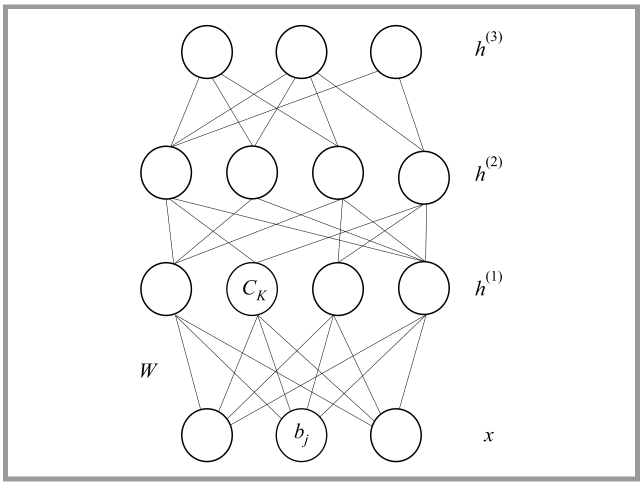


Fig. 2. Graphical model of DBN.

Next, layer-based training is performed for the DBN. Once all the layers have been pre-trained, supervised training is performed for the entire network. This is called fine tuning and the back propagation algorithm can be used to train the network. The output of DBN is, what is recognized who is speaking and what he or she is speaking, and a control signal is generated to validate the proposed technique.

2.4. Hardware Module

The hardware module was based on a 2.4 GHz RF transmitter receiver, a microcontroller (ATMega 8), a motor driver (L293D) and DC gear motors (Fig. 3). The recognized command word is sent from the computer to the hardware unit through RF data modem at 9600 bps to MCU.

The command from Matlab software is received using an RF data modem, using a serial communication protocol. The MCU interprets the commands received. The driver circuit is used to control the DC motors.

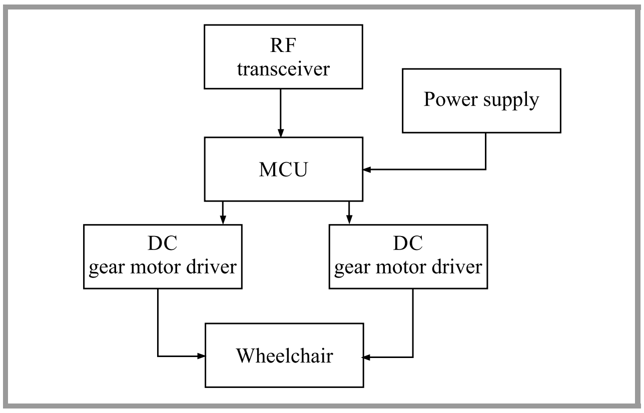
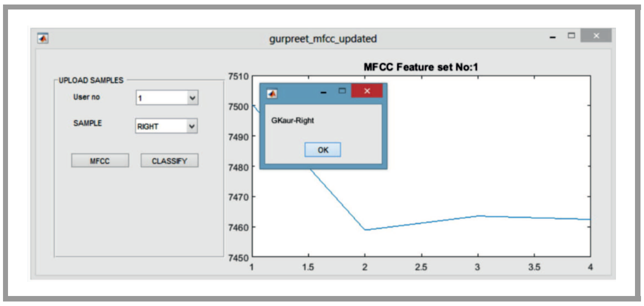


Fig. 3. Hardware of the prototype solution.

3. Results and Discussion

MFCC features are extracted for different words, such as backward, forward, left, right, and stop. Fifty users (25 males and 25 females) took part in the experiment and each of them provided 100 samples of each word. Figure 4 shows one of the examples of the training stage of the speaker-dependent speech recognition with extracted features. Training is performed by the speaker for the word right and the MFCC feature extraction method is used.



User no 1

SAMPLE RIGHT

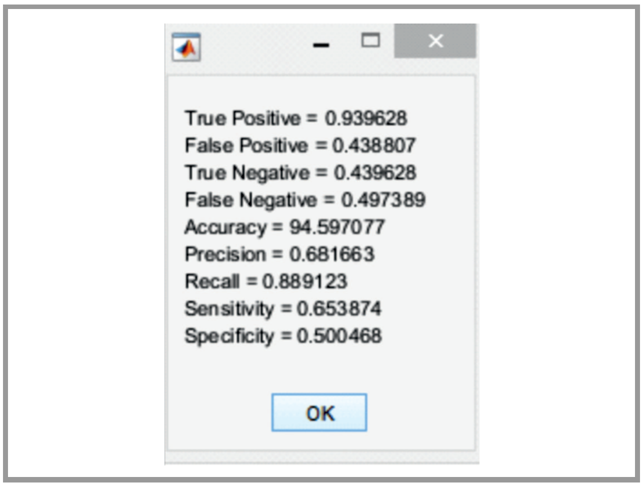
GKaur-Right

Fig. 4. MFCC extracted features.

The extracted MFCC features are fed to DBN network. In DBN, 13 nodes are taken for input and 13 nodes are taken for output. Two hidden layers $h1$ and $h2$ are used. Hidden layer 1 consists of 200 nodes and hidden layer 2 consists of 100 nodes. First pre-training is done with the RBM stacks and then fine tuning is performed with supervised learning method. The matrices used to calculate the accuracy are shown in Fig. 5.

In a real-world implementation, the recognized word is sent from the computer to the hardware unit through an RF data modem at 9600 bps to MCU. After this, serial port is read by the program and according to the command received and motors move accordingly. The prototype has been tested in different environments, such as a playground, an office, a cafeteria and also in a hospital as shown in Table 1.

Recognition accuracy was tested at peak hours (10–11 AM), when maximum levels of background noise are present. The average recognition accuracy achieved is 87.4%.



True Positive = 0.939628
False Positive = 0.438807
True Negative = 0.439628
False Negative = 0.497389
Accuracy = 94.597077
Precision = 0.681663
Recall = 0.889123
Sensitivity = 0.653874
Specificity = 0.500468

Fig. 5. Measuring matrices values.

Table 1
Accuracy (percentage-wise) in different environments

Command	Environment			
	Playground	Office	Cafeteria	Hospital
Backward	88.93	87.01	86.21	85.21
Forward	88.08	87.09	85.09	85.09
Left	89.17	88.07	86.87	85.17
Right	89.00	88.00	86.80	85.10
Stop	89.03	89.07	86.07	85.07

4. Conclusion

A speech-based system has been developed for controlling the movement of a wheelchair. Speech signals have been processed with the use of the MFCC-based feature extraction method and their classification was performed with the help of a DBN-based neural network technique. Five commands have been used to control the vehicle. The prototype has been tested in different environments, such as a playground, an office, a cafeteria and a hospital. The average word recognition accuracy achieved is 87.4%.

References

[1] J. H. L. Hansen and T. Hasan, “Speaker recognition by machines and humans: a tutorial review”, in *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, 2015 (DOI: 10.1109/MSP.2015.2462851).

[2] D. R. Reddy, “Speech recognition by machine: A review”, in *Proc. of the IEEE*, vol. 64, no. 4, 1976, pp. 501–531 (DOI: 10.1109/PROC.1976.10158).

[3] M. Nishimori, T. Saitoh, and R. Konishi, “Voice controlled intelligent wheelchair”, in *Proc. of the SICE Annual Conf.*, Takamatsu, Japan, 2007, pp. 336–340 (DOI: 10.1109/SICE.2007.4421003).

[4] N. Peixoto, H. G. Nik, and H. Charkhkar, “Voice controlled wheelchairs: Fine control by humming”, *Computer Methods and Programs in Biomedicine*, vol. 112, pp. 156–165, 2013 (DOI: 10.1016/j.cmpb.2013.06.009).

[5] V. Partha Saradi and P. Kailasapathi, “Voice-based motion control of a robotic vehicle through visible light communication”, *Computers and Electrical Engineer.*, vol. 76, pp. 154–167, 2019 (DOI: 10.1016/j.compeleceng.2019.03.011).

[6] G. Kaur, M. Srivastava, and A. Kumar, "Integrated speaker and speech recognition for wheel chair movement using artificial intelligence", *Informatica*, vol. 42, pp. 587–594, 2018 (DOI:10.31449/inf.v42i4.2003).

[7] S. Squartini, E. Principi, R. Rotili, and F. Piazza, "Environmental robust speech and speaker recognition through multi-channel histogram equalization", *Neurocomputing*, vol. 78, no. 1, pp. 111–120, 2012 (DOI:10.1016/j.neucom.2011.05.035).

[8] S. Furui, "50 Years of progress in speech and speaker recognition", *ECTI Trans. on Computer and Information Technol.*, pp. 64–74, 2012 (DOI:10.37936/ecti-cit.200512.51834).

[9] G. Kaur, M. Srivastava, and A. Kumar, "Implementation of text dependent speaker verification on Matlab", in *Proc. of 2nd Conf. on Recent Adv. in Engineer. and Comput. Sci. RAECS*, Chandigarh, India, 2015 (DOI: 10.1109/RAECS.2015.7453344).

[10] G. Kaur, M. Srivastava, and A. Kumar, "Analysis of feature extraction methods for speaker dependent speech recognition", *Int. J. of Engineer. and Technol. Innovation*, vol. 7, pp. 78–88, 2017 [Online]. Available: https://ojs.imeti.org/index.php/IJETI/article/view/382/395

[11] S. Narang and D. Gupta, "Speech feature extraction techniques: a review", *Int. J. of Computer Sci. and Mobile Comput.*, vol. 4, no. 3, pp. 107–114, 2015 [Online]. Available: https://www.ijcsmc.com/docs/papers/March2015/V4I3201545.pdf

[12] D. Y. Huang, Z. Zhang, and S. S. Ge, "Speaker state classification based on fusion of asymmetric simple partial least squares (SIMPLS) and support vector machines", *Computer Speech Language*, vol. 28, no. 2, pp. 392–419, 2014 (DOI: 10.1016/j.csl.2013.06.002).

[13] S. M. Siniscalchi, T. Svendsen, and C. H. Lee, "An artificial neural network approach to automatic speech processing", *Neurocomputing*, vol. 140, pp. 326–338, 2014 (DOI: 10.1016/j.neucom.2014.03.005).

[14] N. S. Dey, R. Mohanty, and K. L. Chugh, "Speech and speaker recognition system using artificial neural networks and hidden Markov model", in *Proc. IEEE Int. Conf. on Communication System and Network Technology CSNT*, Rajkot, Gujarat, India, 2012, pp. 311–315 (DOI: 10.1109/CSNT.2012.221).

[15] R. Makhijani and R. Gupta, "Isolated word speech recognition system using Dynamic Time Warping", *Int. J. of Engineering Sciences & Emerging Technologies*, vol. 6, no. 3, pp. 352–367, 2013 [Online]. Available: https://www.ijeset.com/media/0002/9N13-IJESET0603130-v6-iss3-352-363.pdf

[16] G. Dede and M. H. Sazli, "Speech recognition with artificial neural networks", *Digital Signal Process.: A Review Journal*, vol. 20, no. 3, pp. 763–768, 2010 (DOI: 10.1016/j.dsp.2009.10.004).

[17] T. Nikoskinen, "From neural network to deep neural network", *Alto University School of Science*, pp. 1–27, 2015 [Online]. Available: https://sal.aalto.fi/publications/pdf-files/enik15_public.pdf

[18] L. Moreno *et al.*, "On the use of deep feed forward neural networks for automatic language identification", *Computer Speech Language*, vol. 40, pp. 46–59, 2016 (DOI: 10.1016/j.csl.2016.03.001).

[19] T. Alsmadi, H. A. Alissa, E. Trad, and K. Alsmadi, "Artificial intelligence for speech recognition based on neural networks", *J. of Signal and Information Process.*, vol. 6, no. 2, pp. 66–72, 2015 (DOI: 10.4236/jsip.2015.62006).

[20] V. Mitra *et al.*, "Hybrid convolutional neural networks for articulatory and acoustic information-based speech recognition", *Speech Commun.*, vol. 89, pp. 103–112, 2017 (DOI: 10.1016/j.specom.2017.03.003).

[21] M. Farahat and R. Halavati, "Noise robust speech recognition using Deep Belief Networks", *Int. J. of Comput. Intelligence and Applicat.*, vol. 15, no. 1, pp. 1–17, 2016 (DOI: 10.1142/S146902681650005X).

[22] R. Sarikaya, G. E. Hinton, and A. Deoras, "Application of Deep Belief Networks for natural language understanding", *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, pp. 778–784, 2014 (DOI: 10.1109/TASLP.2014.2303296).

[23] A. R. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling Using Deep Belief Networks", *IEEE Trans. on Audio, Speech and Language Process.*, vol. 20, no. 1, pp. 14–22, 2011 (DOI: 10.1109/TASL.2011.2109382).

[24] X. Chen, X. Liu, Y. Wang, M. J. F. Gales, and P. C. Woodland, "Efficient training and evaluation of recurrent neural network language models for automatic speech recognition", *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 24, no. 11, pp. 2146-2157, 2016 (DOI: 10.1109/TASLP.2016.2598304).

[25] R. Ajgou, S. Sbaa, S. Ghendir, and A. Chemsa, "An efficient approach for MFCC feature extraction for text independent speaker identification system", *Int. J. of Commun.*, vol. 9, pp. 114–122, 2015 [Online]. Available: http://www.naun.org/main/NAUN/communications/2015/a382006-081.pdf

Gurpreet Kaur is an Assistant Professor at the Department of Electronics and Communication Engineering at the University Institute of Engineering and Technology, Panjab University, Chandigarh, India. She received her B.Tech. (with Hons) in Electronics and Communication Engineering from Kurukshetra University, Haryana in 2004, M.E. (with distinction) in Electronics and Communication from the University Institute of Engineering and Technology, Panjab University, Chandigarh in 2007, and Ph.D. in Electronics Engineering from IKG Punjab Technical University, Jalandhar in 2018. Her current research interests focus on speech processing and neural networks.

https://orcid.org/0000-0003-3735-0568
E-mail: regs4gurpreet@yahoo.co.in
UIET
Panjab University
Chandigarh, India

Mohit Srivastava is a Professor at the Department of Electronics and Communication Engineering and R&D Dean at Chandigarh Engineering College, Landran, Mohali, Punjab, India. He received his B.Tech. in Electronics and Communication Engineering from Magadh University, Bodh Gaya, M.Tech. in Digital Electronics and Systems from K.N.I.T. Sultanpur and Ph.D. in Image Processing & Remote Sensing from Indian Institute of Technology Roorkee in 2000, 2008 and 2013 respectively. His current research interests focus on digital image and speech processing, remote sensing and their applications in land cover mapping, and communication systems.

https://orcid.org/0000-0002-4566-4279
E-mail: mohitsrivastava.78@gmail.com
CEC Landran
India

Amod Kumar received his B.E. (Hons.) in Electrical and Electronics Engineering from Birla Institute of Technology and Science, Pilani (Raj.), M.E. in Electronics from Punjab University, Chandigarh and Ph.D. in Biomedical Signal Processing from IIT Delhi. He has approximately 38 years of experience in research and development of different instruments used in process control, environmental monitoring, biomedical engineering and prosthetics. He worked as the Chief Scientist at the Central Scientific Instruments Organization (CSIO), Chandigarh, which is a constituent laboratory of CSIR. Currently he is working at NITTTR, Chandigarh, as Professor at the Electronics Department. His areas of interest focus on digital signal processing, image processing and soft computing.

iD https://orcid.org/0000-0003-1177-3191
E-mail: csioamod@yahoo.com
NITTTR
Chandigarh, India