Multimedia Tools and Applications (2023) 82:16173–16193
https://doi.org/10.1007/s11042-022-14019-z

HindiSpeech-Net: a deep learning based robust automatic speech recognition system for Hindi language

Usha Sharma[1] · Hari Om[1] · A. N. Mishra[2]

Received: 16 August 2021 / Revised: 26 April 2022 / Accepted: 23 September 2022 /
Published online: 24 October 2022

Abstract
Automatic Speech Recognition (ASR) has become one of the major research areas over the past decade and gained a lot of interest. Their system implementation, adaptation to different languages and robustness in the performance are still some of the major challenges. Hindi is one of the most widely spoken languages in the world but it is a complex and resource-constraint language. Thus, speech recognition and classification systems need to be developed for Hindi language to spread the technology and to explore more communication means. But due to its language complexity than other languages and lack of standard databases, it is quite challenging to develop such systems. Deep learning is extensively used in different research fields and has proven its prominence to a broader extent. In this paper, a seven-layer 1D-convolutional neural network HindiSpeech-Net has been proposed to recognise different speech samples of the Hindi language in the respective category. A large dataset of 2400 speech samples in the Hindi language is collected in ten different classes in real-world conditions which is further accompanied by signal filtering and augmentation to enhance the dataset for making a robust model and avoid overfitting. The collected dataset is divided into training, validation and test set which were evaluated in different performance parameters. The trained HindiSpeech-Net model achieved an accuracy of 92.92% on the test set. The proposed framework is computationally less expensive, works in real-time and is suitable for implementation in embedded systems.

Keywords 1D-CNN · Convolutional neural network · Hindi language · Deep learning · Speech recognition

✉ Usha Sharma
ushasharma1529@gmail.com

[1] Department of Computer Science & Engineering, Indian Institute of Technology (ISM) Dhanbad, Dhanbad 826004, India

[2] Krishna Engineering College, Ghaziabad 201001, India

1 Introduction

Speech recognition is one of the leading edge and progressive research fields. Apart from this, computers are becoming very much crucial for people nowadays. But it is also essential to make widely available this technology, especially in the rural area of the nation. Speech recognition systems can help to give input to the computer system and proper interpretation in place of a keyboard and mouse. These systems will be more helpful to the rural population, visually impaired and physically challenged persons. Speech recognition systems can be more advantageous for children, old-aged persons and people who are illiterate, differently-abled or suffering from dyslexia. With the aid of speech recognition, the system is able to recognise the different sounds from the audio input and able to interpret them in a certain known language. Expressed words can be changed to text with a mechanism of speech recognition. The reason behind the widespread use of speech recognition in multiple devices because of its ease of communication in a natural manner. It is used in different sectors such as tourism, social media, health and education. Automatic speech recognition systems have a number of applications in telephony and communication. Systems have different complexity levels ranging from small to medium in terms of vocabulary or hands-free dialling. The study of these signals can give a significant contribution which can be helpful to deal with the noise interference during recognition.

Speech recognition is composed of multiple steps such as voice recording, word boundary detection, extraction of features and recognition through trained models. Word boundary detection is the procedure to identify the starting and end of the spoken word in a given speech signal [4]. Silence and pause detection between the words from an input signal is done for word boundary detection. Extraction of features indicates different parameters such as amplitude, energy etc. convenient for speech recognition to get useful and relevant information. Recognition operation maps the given speech input signal with known labels which is done with lots of training and implementation of different acoustic and language models. During training, multiple-input speech signals with known labels are trained to generate models.

The speech interface has two major components; speech synthesiser and speech recogniser. Written text is converted into speech with the ease of a speech synthesiser whereas the speech recogniser tries to get the spoken words and convert that into the text. Nowadays, the speech recognition systems which are available these days work mostly in the English language. The Dictionary and language model used by these kinds of systems are in the English language. Thus, requires proficiency in the English language to use computer-based systems and interpreters for speech recognition. It arises the need for a system that can be developed for the people who are not proficient in English and uses another language as a mode of communication. If common people can interact with the machine in countries like India, communication and information technologies will have more usability and greater acceptance across the mass. In India, more than 70% of the population lives in villages and rural areas and the language interpreter or recogniser should be built in their native language for wide use. Most of the previous research works were concentrated on English or other European languages but there is not much emphasis on Hindi and other regional languages. As India is in the direction of revolution in computer technology, it can be more beneficial to develop speech recognition system in Indian languages. Technology should not be bound in languages and has to be communicated well in regional languages.

Speech signals have complex features and feature vectors are formed after their integration. These feature vectors can be used to make a classification of speech signals. Various classifiers

are used for this purpose such as support vector machine [12], k-nearest neighbour [1], and hidden Markov models [25]. Machine learning classifiers and different neural networks can also be equipped to solve this problem.

Convolutional neural networks have proved their significance in recent years and demonstrated their high performance in image, video and audio signals using data locality bias. Automated analysis of the signals can be done efficiently with the use of deep learning and artificial intelligence. Mel Frequency Cepstral Coefficient (MFCC), centroid, and entropy are a few of the common handcrafted features for supervised classification. But these are not much effective and sometimes cause miss-classification. It is because there is a limited number of extracted features that are not robust and insufficient to characterise the complexity of an input speech signal. Manual intervention and domain knowledge are also required to make use of these handcrafted features. A convolutional neural network eliminates manual feature engineering and can be used to develop automatic speech recognition. Features are automatically extracted from the input signal with the assistance of deep learning.

1.1 Motivation of the work

In developing countries like India where there is a low literacy rate, speech recognition can help to implement and spread the awareness of technology at the ground level. The development of speech recognition systems for Indian languages is difficult because of the complexity of languages in the country. The Hindi language is one of the most important Indian languages and a large population around the globe speak Hindi and understand it. As per International Phonetics Association, there are 16 vowels and 36 consonant sounds whereas there are 20 vowels and 24 consonant sounds in the English language. Although there is much better precision in utterances as compared to English it is a tedious task to recognise the Hindi language. There is a scarcity of annotated Hindi speech data, even though it is one of the world's most widely spoken languages. Thus, only a few automatic speech recognition systems have evolved in Hindi.

The main contributions of the proposed work are:

1) HindiSpeech-Net, a one-dimensional convolutional neural network has been developed for speech recognition of input speech signals in the Hindi language and their multi-class predictive categorisation.
2) As the Hindi language is a complex language and speech datasets are not available, a custom diverse dataset has been prepared for the task of speech recognition from multiple speakers.
3) The data has been prepared in normal uncontrolled conditions to prepare the computer-aided system to recognise in real-world situations. Signal filtering has also been done on collected speech signals which were further augmented to develop a more robust system. Augmentation of signals has prepared the developed model to deal with a wide variety of real-world signals.
4) The proposed 1D-CNN model of multiple layers and parameters has been constructed which was trained for multiple iterations. The performance of the proposed HindiSpeech-Net is assessed before and after data augmentation on various statistical parameters.
5) The proposed speech recognition framework is automatic and works in real-time to classify the speech signals in known labels with high accuracy, which signifies the better performance of the trained model and its applicability in a real-world scenario.

The rest of the paper is organized as follows: Section 2 illustrates the related work. Section 3 describes the speech dataset preparation and proposed methodology for speech recognition.

Section 4 explains the experiments and results regarding the speech recognition task. Section 5 illustrated the discussion part of the results and the wide applicability of the proposed system. Section 6 describes the conclusion and future scope.

2 Related work

Artificial intelligence has revolutionized the research areas in various domains and speech analysis is one of those prominent research areas. Speech recognition, information retrieval from the music, environmental sound detection, audio tracking, speech synthesis, and audio enhancement is a few of the applications on which major research is undergoing. The advancement of deep learning algorithms gives additional aid to developing more robust systems to demonstrate a majority of applications with enhanced capabilities. The word "deep" in the term "deep learning" signifies those neural network architectures having a huge number of parameters and trained with large datasets using current and advanced machine learning techniques i.e., cloud computing and graphic processing units. Advancement in the area of deep learning has evolved multiple utilities in the field of signal processing, which has outperformed traditional machine learning techniques. Thus, traditionally and most commonly used speech signal processing methods i.e., hidden Markov models (HMM) and Gaussian mixture models (GMM) show efficient performance like deep learning models in those scenarios where data is sufficiently available. It is a little different to process the audio signals as compared to image processing based deep learning methods. It is because raw audio samples are one-dimensional time-series signals whereas images are two-dimensional arrays. Thus, audio signals need to be converted into a 2-D representation of time-frequency signals for processing with deep convolutional neural networks. But the frequency and time axes are not homogeneous as vertical and horizontal axis in the respective 2D array. Images are captured instances of the particular object and analysis of patches is done with other parameters but in the case of time series audio signals, the whole sequence is analysed in a chronological manner [32].

Traditional pattern recognition task is reemphasized due to the wide availability of the data and the development of convolutional neural networks and deep learning. Initially, preprocessing of the data, extraction of features, and their classification were done manually which is quite tedious [24]. Various researches were done to improvise the methodology and parts of speech processing as well as the feature extraction part [39]. These engineered features are also known as handmade or handcrafted features. With the advent of deep learning, these steps get automated which results in more processing of data in less time. The field of audio tagging is also advanced to predict whether the target sound class is present in the given audio clip or not [16]. Different types of features can be extracted from the audio files and all these features can be fused to perform the audio classification in the form of an ensemble [28]. The collected features can give good classification results after their proper comparison and evaluation. Handcrafted features can be combined later to get more accurate classification results.

Most speech recognition systems work on the principle to get a sequence of phonemes from the input audio signal and search for the word to convert into the text. But it is critical and challenging and prone to speech classification errors. Hierarchical phoneme clustering techniques can be equipped for the selection of the best model among the available ones and such implementation has achieved 71.7% accuracy [30]. A clinical depression diagnosis is equipped

using phoneme-level based convolutional neural networks [27]. Depression can be analysed using the evaluation of the speech on different parameters. Vowels and consonants are used as acoustic characteristics to well differentiate normal speech from depressed speech.

Researchers are also utilizing the spectrogram of the speech signals to equip the recognition task in the form of a 2-dimensional image classification task. The speech signals are converted into the respective spectrogram images and the collection of such images of different categories is given to the 2-D convolutional neural network to find the discrimination between different classes of the signals. These signals are processed through multiple numbers of layers to make a differentiation between the signals [26]. Transfer learning is mostly used in this methodology where performance efficient deep neural networks are utilized which has proven their accuracy in other image datasets. These networks are also used in a wide area of applications such as speech-health classification tasks. Recurrent neural networks with different feature extractors such as compact bilinear pooling are also used for this purpose of speech-health classification where suggestions regarding intoxication of the speaker and suggestion of the food can be found [34]. Unmanned aerial vehicles (UAVs) or drones are becoming widely popular and nowadays used in many day-to-day applications like surveillance, journalism, movie-making etc. Speech recognition and classification frameworks are also making technical interventions with UAVs to associate visual data with audio information [31]. These approaches are also used to get visual feedback according to the audio commands.

In the last decade, environmental sound classification has become prominent due to the large construction of cities. But, accurate recognition of such sound is tough due to the vigorous interference due to ambient noise and their non-stationary nature. To deal with such situations, a two-stream convolutional neural network (CNN) with a random-padding method is developed and short length data sequences are also analysed to classify environmental sounds [8]. In the biomedical signal processing domain to develop diagnostic systems, Fen Li et al. proposed a framework using the one-dimensional convolutional neural network to classify the input heart sound signals into normal and abnormal in place of traditional ECG methods. A denoising autoencoder algorithm is used to extract deep features from input cardiac sound [23]. For the applications in the field of end-to-end speech recognition, deep learning architecture named ContextNet is proposed where global context information is incorporated into convolution layers due to the fully convolutional encoder by adding squeeze and excitation modules [14]. ContextNet has shown high computational and accurate performance. Jongpil Lee et al. proposed a CNN architecture for music classification using sample level filters in place of traditional frame-level input representations [22].

An audio-visual speech recognition system is also beneficial in those scenarios where the audio gets corrupted by noise. The selection of important and discriminatory features is an important step to get high performance to recognise the input speech signals. A noise-robust hidden Markov model-based framework is proposed for audio-visual speech recognition. Audio features are extracted with the deep de-noising autoencoder and CNN-model are used to extract visual features from the area of the raw mouth in a captured image. Then, audio and visual features from previous networks are integrated using a multi-stream HMM [29]. In another approach by Sharmila et al., the features with a discrete wavelet of the lip portion that integrated with the Bark frequency Cepstral coefficients (BFCC) while Pseudo Hue and Intensity of Colour were used for localization lip using Hidden Markov Model (HMM) [36]. In another work, hidden-Markov models and MFCC coefficients were used for the Hindi vowel recognition and resulted in a recognition score of 83.19% [5].

Wavelet transformation is also used in many research works for speech recognition in the Hindi language for transient features extraction [11, 15, 35]. The performance of the traditional classifiers can be used when acoustic features are also used with them. Acoustic features and recurrent neural networks are integrated for automatic speech recognition in the Hindi language [17]. Different discriminative methods were used for training the acoustic model to improve the performance and recurrent neural networks for language modelling from the data in the form of text.

Some experiments have been also done in the direction of recognition of vowel, consonant, and phoneme in the Hindi language for extraction of various features and their modelling. A hybrid feature extractor is developed by combining Mel frequency cepstral coefficients and perceptual linear predictive for improving the performance of speech recognition [19]. Voice activity and detection-based frame dropping formula have been utilised for removing distorted elements of speech and pauses to improve the modelling of phonemes. Shobha Bhatt et al. has developed a continuous speech recognition system in the Hindi language using syllables which are the larger acoustic units [10]. Hidden Markov models with perceptual linear predictive coefficients are equipped for this purpose. Research findings indicate that by selecting an appropriate acoustic unit for Hindi, the performance of the speech recognition system may be improved. The Hindi language is an Indian language that is a complex and limited resource language, research has been done in the field of speech classification. Time-delay Neural Network (TDNN) based approach is proposed using acoustic modelling with i-vector adaptation to capture the extended temporal context of acoustic events [33]. The accuracy of 89.9% is used after implementing data augmentation techniques. Anik Dey et al. proposed an approach for acoustic modelling of Hindi speech by borrowing from English data, where a baseline Gaussian model-sharing is used with DNN training [6]. In this method, DNN is trained in English and then the last layer is fine-tuned after training by using the target Hindi data. The trained model is accessed on one hour of transcribed Hindi data and achieved an accuracy of 59.9%.

Most state-of-the-art automatic speech recognition systems use Perceptual Linear Prediction (PLP) and MFCC for feature extraction to train and develop speech recognition systems. But, the sensitivity of PLP and MFCC is not much in the case of background noise and it gives rise to more noise-robust features like Basilar-membrane Frequency-band Cepstral Coefficient (BFCC) and Gammatone Frequency Cepstral Coefficient (GFCC). However, many issues associated with these feature extraction methods such as a standard number of filters and accepted bandwidth are still needed to be explored more for more efficient performance. In another work, a differential evolution algorithm-based approach is presented for optimization of the filter number and their spacing in, BFCC, GFCC and MFCC methods [18]. Experiments conclude that BFCC based speech recognition systems perform superior as compared to GFCC and MFCC under various conditions.

Over the last two decades, there has been a lot of interest in multimedia content analysis. The impact of background noise from Indian musical instruments on automatic speech recognition has been studied [21]. The capacity to recognise distinct words is affected differently by various instruments. In noisy circumstances, it has been observed that traditional automatic speech recognition algorithms that use the MFCC or perceptual linear prediction as features perform poorly. Numerous feature extraction strategies in noisy situations are evaluated and the result indicated that gammatone frequency cepstral coefficients perform well [13]. Other Indian languages, such as Kannad, are also being investigated for automatic speech recognition purposes [20]. In big data and other artificial intelligence domains, feature

selection is considered a difficult optimization problem. Different optimization techniques are designed to use with different classifiers, such as the monarch butterfly optimization algorithm [2]. To make artificial intelligence systems more resilient, many optimization strategies are being developed [3, 37]. Various efforts have also been made to develop encryption techniques in order to improve privacy and security [38].

But a lack of standard datasets and pronunciation dictionaries in the Hindi language which is spoken in India and across the globe, and there is very less work reported in this language [7]. Retroflexion, nasalization, gemination, and aspiration are salient features of Hindi language speech creating more complexity to develop speech recognition systems [9].

It creates a wide scope to explore more these low resource and most spoken languages like Hindi to improve the existing systems as well as to develop new speech recognition systems. There is a scarcity of word datasets in the Hindi language and deep learning-based approaches can be used to develop more intelligent systems. To achieve this purpose, a huge dataset is collected in the proposed work which is further augmented to increase the dataset in the development of a robust trained model. A novel CNN model represented as HindiSpeech-Net, is proposed and trained with the speech dataset.

3 Materials and methods

The artificial intelligence-based speech signal classification method is presented in the proposed work using a one-dimensional convolutional neural network, HindiSpeech-Net. The proposed speech classification framework is shown in Fig. 1.

Different speech signals which were collected from different speakers are processed. Proposed 1D-CNN architecture of HindiSpeech-Net has been used to extract prominent and discriminatory features from the input speech signals to distinguish different classes of Hindi digits. After the training of the 1D-CNN model with a dataset of input speech signals, the saved deep learning model tries to classify the input test speech signal into one of the 10 different classes of digits in the Hindi language. This multi-class classification problem is made up of multiple steps i.e., acquisition of input speech signals, pre-processing of signals, augmentation of captured speech signals, designing of 1D-CNN architecture, and evaluation of the proposed model.

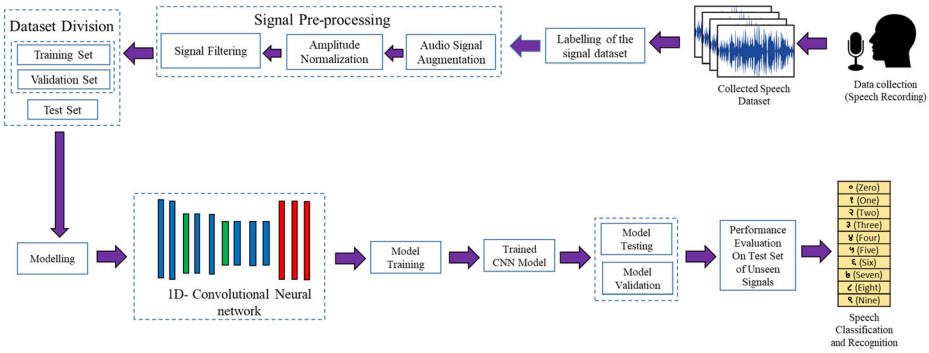

Fig. 1 Block diagram of the proposed methodology

3.1 Data collection

As there is a scarcity of Hindi language public speech dataset, a diverse audio speech dataset of Hindi isolated digits is formed by the recordings of the speech of different speakers. A total of 24 speakers are involved in preparing this dataset and they belong to the age group of 21 to 30 years. A total of eighteen female and six male speakers records the Hindi language numeric digits. Ten isolated Hindi digits have the pronunciation as 'Shoonya (0)', 'Ek (1)', 'Do (2)', 'Teen (3)', 'Chaar (4)', 'Paanch (5)', 'Chhey (6)', 'Saat (7)', 'Aath (8)' and 'Nau (9)'. Different isolated Hindi numeric digits are elaborated in Table 1 with their corresponding English pronunciation and numerals.

Each digit is recorded ten times by every speaker. A total of 2400 audio speech samples in the Hindi language were recorded by all speakers. The audio files were combined at16 kHz sampling frequency and saved in Microsoft Wave (.WAV) file format using uncompressed Pulse Code Modulation. Samples of audio signals for each digit and respective spectrograms are shown in Fig. 2.

3.2 Signal augmentation

Feature extraction and their classification is challenging in the case of short clips of speech. Background noise and the short duration of audio are some of the issues which make it more challenging for a classification model to perform its task. Thus, it is required to train with a wide variety of data which can contribute to preparing a more efficient model. As a vast number of parameters are required to be tuned while training with the dataset, it will require a large amount of data. Data augmentation techniques can help to prepare more robust models.

Data augmentation methods can not only introduce the variation in the training dataset but also increases the number of samples in the training dataset. In case of insufficiency of data, the convolutional neural network can suffer overfitting. Overfitting signifies a huge variation in the training and testing accuracy. Deep neural networks mostly use data augmentation for images but it can be equally useful for audio signals.

In the collected dataset, the speech signals for each Hindi digit have 240 samples which is relatively limited for the multi-class classification task. A sufficient number of speech data

Table 1 Different Hindi numeric digits and their English pronunciation with numerals

Hindi Digits	Hindi Pronunciation	English Digits	English Pronunciation
०	Shoonya	0	Zero
१	Ek	1	One
२	Do	2	Two
३	Teen	3	Three
४	Chaar	4	Four
५	Paanch	5	Five
६	Chhey	6	Six
७	Saat	7	Seven
८	Aath	8	Eight
९	Nau	9	Nine

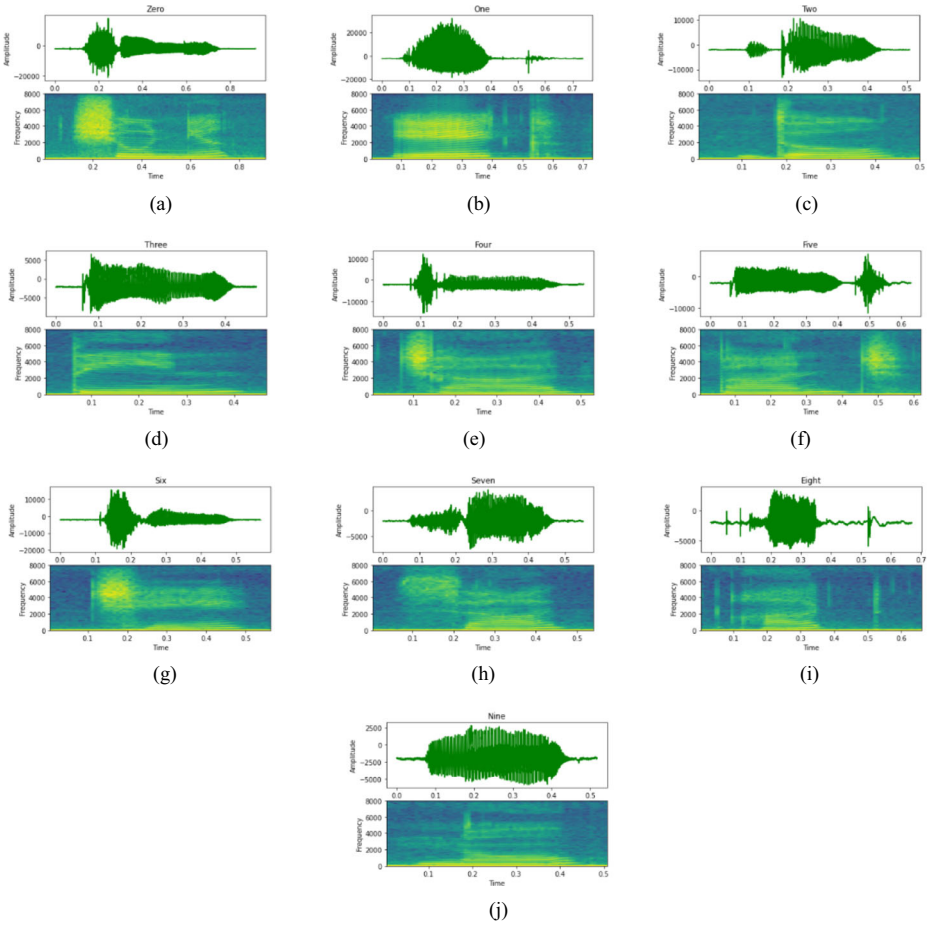

Fig. 2 Speech signals for different recoded speech of different digits in Hindi language and their spectrogram, **a** Zero, **b** One, **c** Two, **d** Three, **e** Four, **f** Five, **g** Six, **h** Seven, **i** Eight, **j** Nine

samples are required for proper training of a deep learning model to avoid overfitting in the trained model. To achieve this objective speech signals are augmented to enhance the number of speech samples in the collected dataset. Three audio augmentation techniques are equipped in this work, i.e., changing the speed, pitch variation and addition of noise. Input signals with respective augmented signals are shown in Fig. 3. The pitch of the signal is varied with the pitch factor of 2 and 4. Similarly, speech signals are also augmented to make the speed of audio fast and slow. A speed factor of 1.5 and 0.75 is taken to make the speed of the audio fast and slow, respectively.

As real-world situations where speech is recorded may have a noisy environment, the system should be made robust to deal with such situations. Thus, audio signals are augmented with the background deformation technique in the proposed methodology. A given speech signal is represented as $S = [s_1, s_2, s_3, s_4, \ldots\ldots\ldots, s_n]$ where n is the length of the signal and s denotes the amplitude of the speech signal. Background deformations denoted by $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \ldots\ldots\ldots, \alpha_n)$ which has the same length n as that of an input speech signal. The

deformation control parameter is selected as μ. The augmented signal ϒ is generated from the equation as –

$$\Upsilon = S + (\alpha \times \mu) \tag{1}$$

The value of α ranges in the interval (0,1) and μ is taken as 1500 Hz for the conversion of deformation value. After the generation of noise, the input speech signal is mixed with the generated noise signal in the background.

3.3 Signal pre-processing

Multi-class classification of Hindi language digit speech signals is quite challenging. Thus, all signal need to be pre-processed before training a deep learning model. This will help to get remedy from the noise and unwanted signals. There may be a large variation in the amplitude of the given speech signal. It may be possible that recorded signals have different time length of the signals. As signals range from 0.7 to 1.3 s. For a safety reason, a time range of a signal is taken as 1.375 s. The frequency range of a final processed signal is taken between 1000 and 4800 Hz where most of the speech sample's frequency is normally found. The recorded signals have a frequency of 16 kHz and the final length of the signal is taken as 22,000 vectors.

3.4 Convolutional neural network architecture of proposed HindiSpeech-Net

Convolutional neural networks have made a significant advancement in the domain of speech signal analysis and recognition. Traditional methods which require hand-crafted features and tedious processing steps can be replaced by the automatic convolutional neural network. These kinds of networks try to find the hidden pattern in the given data and try to relate it with the

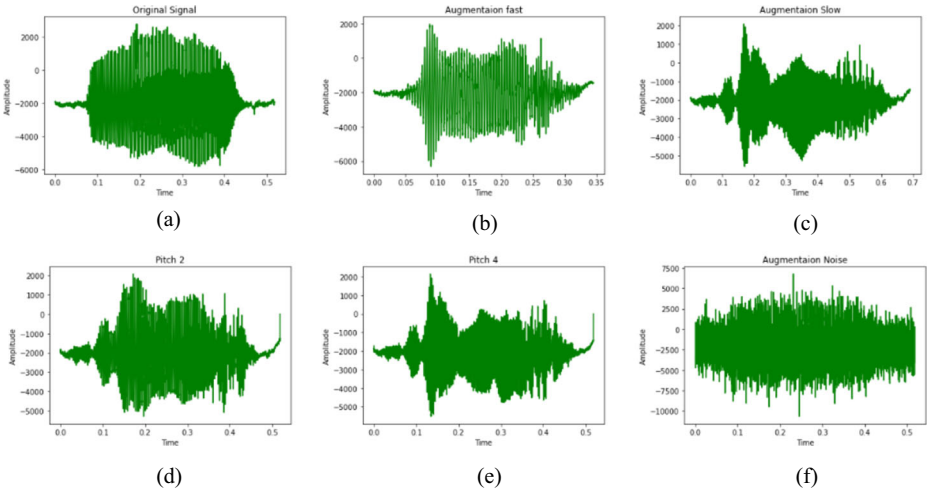

Fig. 3 Augmentation of sample 1D speech signals, **a** Original Signal, **b** Speeding up, **c** Slowing Down, **d** Changing Pitch with pitch factor of 2, **e** Changing Pitch with pitch factor of 4, **f** Noise Injection

output in an automatic manner. Convolutional neural network characterises a function f having parameter σ which tries to predict the output y for given input x and all these are related as:

$$\hat{y} = f_{\sigma}(x) \tag{2}$$

The value of the parameter σ is obtained by the training a neural network based on the relationship between input x_i and output y_i. For a data having N number of input-output pairs, the value of i ranges from 1 to N.one-dimensional convolutional neural network extracts features from data sequence and internal features are mapped with different categories of speech signals.

The advantage of using neural network for classification of raw time series speech data is that the model tries to learn raw signals themselves and there is no specific requirement of domain expertise for manual feature extraction. Such models can learn the internal representation of the data and get comparable outcomes obtained from expert assisted hand-crafted features.

The proposed sound recognition architecture of HindiSpeech-Net is composed of various layers and has multiple network parameters. The architecture of the proposed sequential deep learning model to predict different sounds of digits in the Hindi language is given in Fig. 4 with a respective number of parameters and their output shape. The input layer has the dimension of 1 × 22,000 which is trained on speech dataset Hindi digits. All the speech signals of the collected dataset are pre-processed and reshaped according to the input convolutional layer of deep learning architecture. The proposed deep learning model is trained for a large number of iterations and the best model which obtained lower loss and high accuracy is saved for further evaluation. Weight adjustment is done concerning the validation samples of the Hindi speech provided to the model. The best model is tested for different speech signals in the test dataset for the analysis of the results. CNN models have proved their superiority and effectiveness in the domain of speech classification and recognition. The general architecture of any deep neural network composed of one or more convolution layers that are connected alternatively with pooling layers and activation functions. In the end, a fully connected or flatten layer is attached to perform the classification of an input signal into defined categories.

Convolution is an important operation of deep neural networks; thus, convolutional layer are the important building blocks of such architectures. The convolution of any two input signals resulted in the third signal. In speech processing, two different speech signals as an array of numbers of different sizes but of the same dimensionality produces a third number array having the same dimensions. A kernel which is a small array of numbers, slides over the one-dimensional speech signal through all positions within the duration or length of the speech signal and single convolved value has resulted from each position of the kernel for generation of a feature map.

1D-Convolutional neural network of HindiSpeech-Net takes time-series signals or waveforms as an input represented in the form of a 1D vector. For t seconds long waveforms, the input layer is a 1D vector of dimensions 22,000 × 1. Small-sized kernels are used in convolutional layers to build the proposed CNN architecture of HindiSpeech-Net. Basic modules like activation function and batch normalization are attached after each convolutional layer. Depth of the neural network architecture is an important constituent to obtain more accuracy.

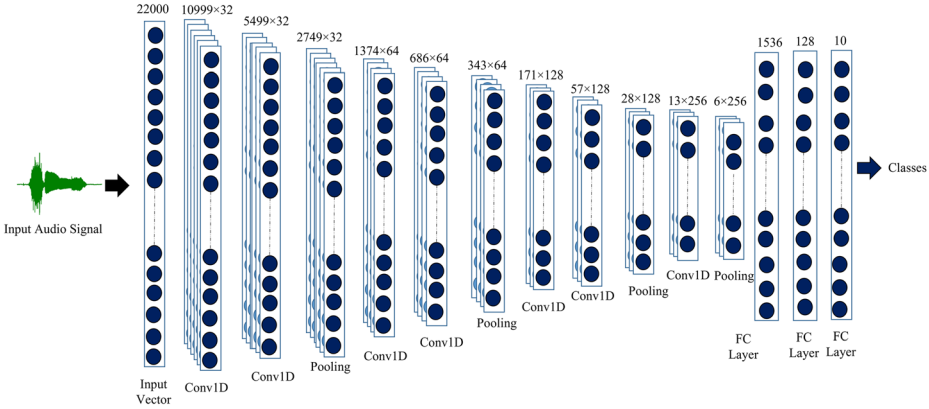

Fig. 4 Proposed 1D-CNN architecture of HindiSpeech-Net for speech classification

Rectified Linear Unit (ReLU) is used in the proposed architecture which is an unsaturated non-linear activation function. It is chosen because it gives computationally efficient performance than other saturated non-linear functions such as Sigmoid and Tanh. The training process gets accelerated and accuracy is significantly improved with the use of ReLU function.

$$\text{Rectified Linear Unit, } ReLU(x) = \begin{cases} x, & if\ x > 0 \\ 0, & if\ x \leq 0 \end{cases} \tag{3}$$

Number of network parameters are exponentially increased with an increasing number of convolutional layers. Hence, pooling operation is applied after convolution to scale down the number of network parameters. Max pooling is chosen for the proposed speech classification model. The pooling window slides over the feature map to get high activation values to select high response generating neurons.

The dropout layer is used in the proposed architecture of HindiSpeech-Net to make the CNN model more generalized and save the model from the overfitting scenarios. On applying the dropout function in the HindiSpeech-Net model, a certain number of neurons gets suppressed according to the used dropout value. The dropout layer is mainly used in neural networks as an effective regularization technique where noise is added to the hidden layers in a stochastic manner. A fully connected or flatten layer is one of the final layers of the CNN model of the proposed Hindi language speech classification and recognition framework. Each node in fully connected layer is connected to all nodes in previous layers and weights for the connections are fixed for each of the nodes. The number of neurons in the flatten layer are treated as hyperparameters which are to be selected empirically. The loss value is taken as a performance parameter for evaluation of the trained model where the difference between the predicted and actual label is computed for the given input speech sample. Feature maps are generated from the input speech signals after applying the convolution operations.

In a proposed CNN model of HindiSpeech-Net, each layer has a different number of neurons and activation functions. The proposed architecture has seven convolutional layers as seven is considered as standard for such neural network implementations due to their optimal performance. The total 393,770 number of weight parameters used in the proposed CNN architecture. The structure of the proposed 1D-CNN is described in detail in Fig. 5.

Layer	Output Shape	Parameters
Input	(22000, 1)	0
Conv1D	(10999, 32)	128
Batch Normalization	(10999, 32)	128
Conv1D	(5499, 32)	3104
Batch Normalization	(5499, 32)	128
Max Pooling 1D	(2749, 32)	0
Conv1D	(1374, 64)	6208
Batch Normalization	(1374, 64)	256
Conv1D	(686, 64)	12352
Batch Normalization	(686, 64)	256
Max Pooling 1D	(343, 64)	0
Conv1D	(171, 128)	24704
Batch Normalization	(171, 128)	512
Conv1D	(57, 128)	49280
Batch Normalization	(57, 128)	512
Max Pooling 1D	(28, 128)	0
Conv1D	(13, 256)	98560
Batch Normalization	(13, 256)	1024
Max Pooling 1D	(6, 256)	0
Dropout (15%)	(6, 256)	0
Flatten	1536	0
Dense	128	196736
Dropout (20%)	128	0
Dense	10	1290
Total parameters: 395,178		
Trainable parameters: 393,770		
Non-trainable parameters: 1,408		

Fig. 5 Parametric description of the Proposed 1D-CNN Architecture of HindiSpeech-Net

The kernel size of 3 is taken for all filters in seven convolution layers. Each convolution layer is followed by the stride of 2 with valid padding. ReLU activation function is used in all the layers except for the final classification layer where the softmax function is taken. Batch Normalization with kernel size 2 is done after each convolution layer. After seven convolution layers, flatten or fully connected layer is equipped with 128 output units having a dropout of 0.3 value. The last layer is the classification layer and has an output shape of 10 to classify the input speech signal into the labelled 10 classes of Hindi digits.

4 Experimental results and discussion

Training of CNN model is one of the crucial steps in the deep learning frameworks. The proposed method has been used on a collected dataset of a wide variety of speech signals in various conditions. Collected speech samples are grouped into training, validation and testing with no repetition. Each input speech signal is transformed into a semantic vector of 22,000 bins. Categorical cross-entropy loss is used to optimize the training process. Signals are processed in a batch size of 32 samples. A batch of 32 random signals was selected from the training data without any repetition to train the model in each step of a training epoch or iteration. The model is trained for the 2000 number of epochs. A validation set is used for

analysing the classification accuracy of the trained model and fine-tuning network parameters of the CNN model.

Various hyper-parameters used for training the proposed deep learning architecture for the multi-class classification of input speech signals such as the filters, stride, epoch with early stopping criteria, kernel size and have been determined empirically by performing several experiments. Early stopping is also used where a criterion of 100 epochs is been set and if performance is not improved training is halted. Valid padding is used and adam is used as an optimization algorithm. Categorical cross-entropy is chosen as a loss hyperparameter. The trained deep neural network predicts into 10 different classes. The CNN is trained on Python using the Keras library and the training system has an Intel i-7 processor with 16 GB RAM.

The collected dataset of speech signals is divided into training, validation and testing set. The test set is composed of 20% audio samples or 48 samples from every 240 samples of each Hindi digit speech from 24 different speakers. Two speech samples are randomly chosen from each speaker to select the 48 speech samples in the test set to evaluate the performance. This will remain the same throughout the experiment to properly assess the trained model accuracy and prediction capability. Augmentation is applied only to the training and validation set and after training the model, already separated test set is used to analyse in different parameters. The number of speech samples before and after augmentation is shown as in Table 2 in training, validation and test set data configuration for respective Hindi digit speech recorded signals.

Accuracy is the measurement of correctly predicted classes out of total classes, which implies the rate of accurate classification. Average accuracy is considered for the final evaluation. Precision is the ratio of correctly predicted cervical cancer (True Positive) out of the total predicted cervical cancer. Recall or sensitivity is another important measurement for performance evaluation. It is the ratio of actually predicted cervical cancer (True Positive) out of the total available cervical cancer patients. F1-score is a single number for evaluation of the model's performance unlike precision and recall always discussed together. F1-score is the harmonic mean of precision and recall. The value of the F1 score varies between 0 and 1, higher value represents a better performance of the model.

$$Accuracy = \frac{\text{Sum of all True Positives}}{\text{Total Number of samples}} \tag{4}$$

Table 2 Dataset configuration for different speech signals before and after augmentation

Hindi Digit	Collected Speech Samples	Before Augmentation			After Augmentation		
		Training	Validation	Test	Training	Validation	Test
० (0)	240	154	38	48	924	228	48
१ (1)	240	154	38	48	924	228	48
२ (2)	240	154	38	48	924	228	48
३ (3)	240	154	38	48	924	228	48
४ (4)	240	154	38	48	924	228	48
५ (5)	240	154	38	48	924	228	48
६ (6)	240	154	38	48	924	228	48
७ (7)	240	154	38	48	924	228	48
८ (8)	240	154	38	48	924	228	48
९ (9)	240	154	38	48	924	228	48
Total	2400	1540	380	480	9240	2280	480

$$\text{Precision} = \frac{(\text{True Positives})}{(\text{True Positives} + \text{False Positives})} \tag{5}$$

$$\text{Recall or Sensitivity} = \frac{(\text{True Positives})}{(\text{True Positives} + \text{False Negatives})} \tag{6}$$

$$F1 \text{ Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{7}$$

Model is trained for 2000 epochs in a batch size of 32. First, the model is trained with 192 speech samples which were divided into training and validation set with 154 and 38 samples, respectively. The trained model with lower validation loss is saved for different epochs or training iterations. The final saved CNN model of HindiSpeech-Net after the defined number of epochs has been used to make predictions in the test set. The testing results are tested on the 480 samples having 48 samples belonging to every single digit in the Hindi Language. TP, TN, FP, and FN are the acronym of True positive, True Negative, False positive, and False Negative speech samples. Calculated values of precision, recall, and F-1 score shown there in Table 3 for the original dataset. The obtained accuracy for classifying the speech signals is 85.21%.

Confusion matrices are generated after evaluation of the performance of the trained CNN models on the testing speech dataset. The confusion matrix depicts the capability of the proposed CNN model to differentiate two different classes of speech signals in the Hindi language. The confusion matrix for the test set using the trained model is shown in Fig. 6 which shows True positive values in the diagonals for all trained classes.

After having the augmentation, augmented data is five times that of the original data after separating the testing set. There are a total of 11,520 audio samples which are divided into the training and validation set in the ratio of 4:1. The original dataset and augmented dataset are combined. So, the proposed CNN architecture of HindiSpeech-Net has been trained with 9240 samples and validated with 2280 samples. After training the model for 2000 epochs, the trained model with lower validation loss is saved. The accuracy and loss parameter for each epoch for both training and validation set is shown in Fig. 7.

Table 3 Performance evaluation of trained model for original dataset

Hindi Digits	TP	TN	FP	FN	Precision (%)	Recall (%)	F1-Score (%)
० (0)	39	423	6	9	86.67	81.25	83.87
१ (1)	41	425	10	7	80.39	85.42	82.83
२ (2)	43	427	7	5	86.00	89.58	87.76
३ (3)	41	425	13	7	75.93	85.42	80.39
४ (4)	38	422	11	10	77.55	79.17	78.35
५ (5)	42	426	9	6	82.35	87.50	84.85
६ (6)	43	427	13	5	76.79	89.58	82.69
७ (7)	42	426	14	6	75.00	87.50	80.77
८ (8)	43	427	15	5	74.14	89.58	81.13
९ (9)	37	421	18	11	67.27	77.08	71.84
Average					78.21	85.21	81.45

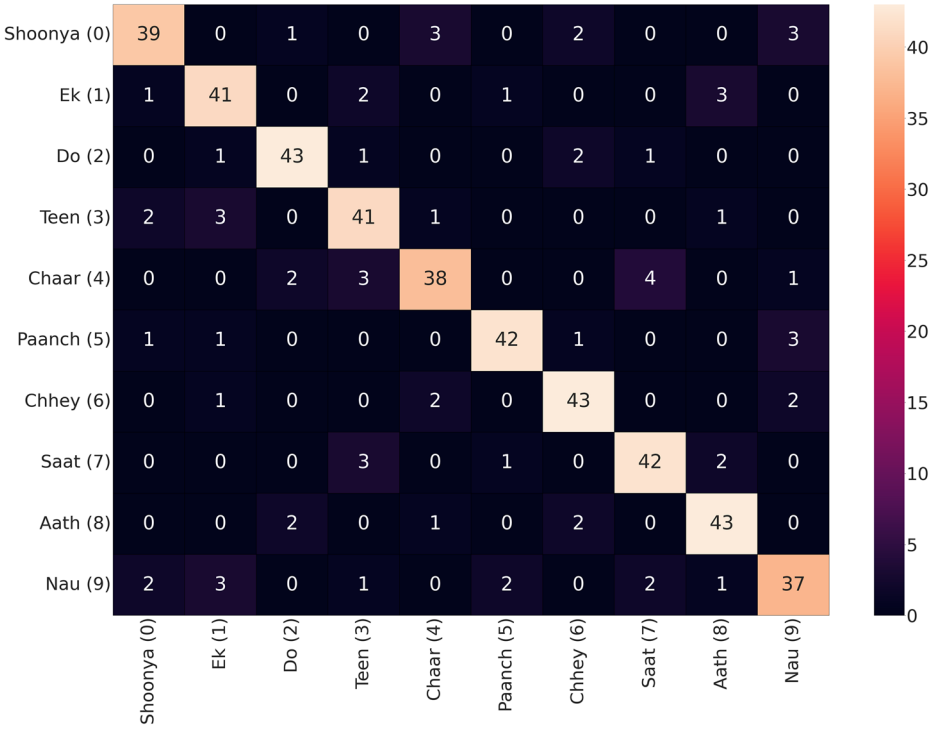

Fig. 6 Confusion Matrix for model trained on original dataset

The trained model performs better in comparison to the model trained only on the original dataset. Precision, recall, F-1 score for all Hindi digits are given as given in Table 4 after performing augmentation. After training the 1D-CNN model of HindiSpeech-Net after the augmentation of training speech signals, the classification accuracy on the test dataset is obtained as 92.92%.

The confusion matrix for the test set using the trained model on the augmented dataset is shown in Fig. 8 which shows True positive values in the diagonals for all trained classes. The trained model achieved high accuracy with less false positives and true negatives.

The number of convolutional layers in the given architectures are also analysed from four number of layers to eight number of layers. The proposed architecture with seven number of convolutional layers performs best among all in comparison with other architectures having

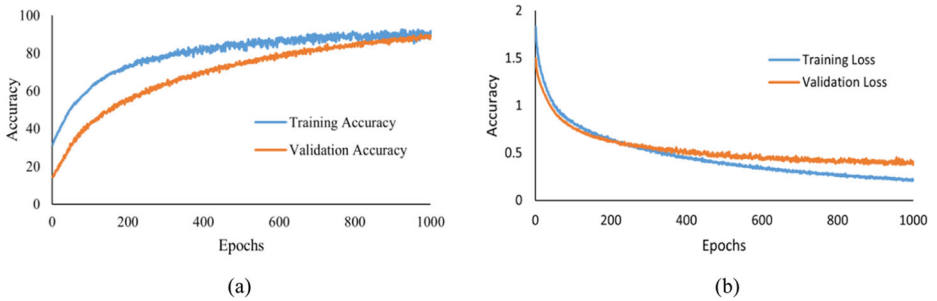

Fig. 7 Training and Validation parameter **a** Accuracy, **b** Loss

Table 4 Performance evaluation of trained model after augmentation

Hindi Digits	TP	TN	FP	FN	Precision (%)	Recall (%)	F1-Score (%)
० (0)	44	428	3	4	93.62	91.67	92.63
१ (1)	46	430	3	2	93.88	95.83	94.85
२ (2)	47	431	4	1	92.16	97.92	94.95
३ (3)	44	428	5	4	89.80	91.67	90.72
४ (4)	41	425	11	7	78.85	85.42	82.00
५ (5)	45	429	8	3	84.91	93.75	89.11
६ (6)	44	428	10	4	81.48	91.67	86.27
७ (7)	45	429	10	3	81.82	93.75	87.38
८ (8)	47	431	12	1	79.66	97.92	87.85
९ (9)	43	427	13	5	76.79	89.58	82.69
Average					85.29	92.92	88.85

different number of parameters. These architectures are evaluated for both datasets i.e., with and without augmentation, as given in Table 5.

The proposed architecture used adam as the optimizer and different optimizers are also explored in augmented and raw dataset, as given in Table 6. The architecture with adam optimizer performs best in comparison to rest of the optimizers.

Different state of the art techniques such as hidden Markov models (HMM) and different machine learning classifiers are assessed on the same datasets. Features extraction techniques such as Mel-frequency cepstral coefficients (MFCC) features and Gammatone Frequency

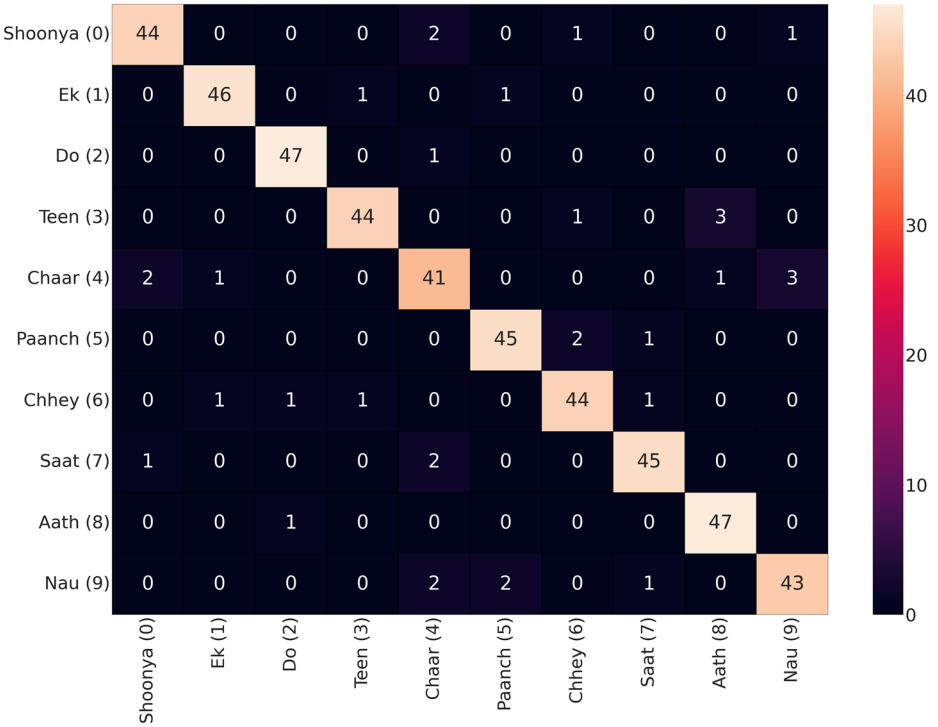

Fig. 8 Confusion Matrix - After Augmentation

Table 5 Performance analysis of different models with different number of convolutional layers

Number of Convolution Layers	Highest Number of Filters	Number of Parameters	Without Augmentation		After Augmentation	
			Validation Accuracy (%)	Testing Accuracy (%)	Validation Accuracy (%)	Testing Accuracy (%)
4	64	2,833,834	80.79	75.83	84.47	86.04
5	128	1,441,834	84.47	77.29	87.11	84.38
6	128	557,738	87.63	80.83	89.21	87.29
7	256	395,178	88.42	85.21	91.58	92.92
8	256	494,762	87.11	83.54	90.26	91.04

Table 6 Performance comparison of different optimizer on performance of the architecture

Optimizer	Without Augmentation		After Augmentation	
	Validation Accuracy (%)	Testing Accuracy (%)	Validation Accuracy (%)	Testing Accuracy (%)
Adam	88.42	85.21	91.58	92.92
RMSprop	87.11	83.54	89.74	90.63
Adagrad	86.58	83.96	88.68	89.38
Adadelta	86.05	82.71	87.11	87.92
SGD	88.42	85.21	91.58	92.92

Cepstral Coefficients (GFCC) features are also considered and detailed comparison with proposed 1D-CNN model has been given in Table 7.

The proposed model takes 87 min to train the model and 1.05 min to assess the validation set. The computation time to analyse an input speech signal takes 0.1 s which can be suitable to process the input speech signals in real-time. These systems can be effectively utilised in a wide variety of applications.

5 Discussion

In the proposed work, a 1D-convolutional neural network architecture of HindiSpeech-Net is presented for speech recognition. The speech-language is chosen as Hindi which is widely

Table 7 State-of-the-art comparison with the proposed work

Method	Test Accuracy (%)	Average Precision (%)	Average Recall (%)	F1-Score (%)
Hidden Markov model (HMM) with Mel-frequency cepstral coefficients (MFCC) features	82.08	78.21	81.37	79.76
Hidden Markov model (HMM) with Gammatone Frequency Cepstral Coefficients (GFCC) features	78.54	74.29	76.12	75.19
Random Forest classifier after extraction of MFCC features	83.54	79.11	82.97	80.99
Support Vector Machine after extraction of MFCC features	84.17	81.22	84.73	82.94
Proposed Method (1D-CNN)	92.92	85.29	92.92	88.85

spoken around the globe but dataset availability is limited. Little work has been done in the field of Hindi speech recognition and classification. Hindi is also a complex language and has different involved complexities like retroflexion, nasalization, gemination, and aspiration. These complexities make it more challenging to recognise and classify sounds. Deep learning which has proven its own advantages and significance in different fields of research, is utilised in this research work. A large dataset is collected for this purpose in real-world conditions from both male and female speakers. As different variety of signals for a single sound can be present while implementing in real-world conditions, data augmentation is also equipped. Speech signals are various signals is varied in terms of their speed and pitch. All the signals are converted into the length of 22,000 to give them as input to the 1-D convolutional neural network. HindiSpeech-Net has been trained for multiple epochs and achieved significant accuracy on the test dataset.

As the developed system works in real-time, automated and trained with a wide variety of signals, it is suitable to implement in real-world conditions with different applications of speech recognition and classification. The same approach can be used to deal with signals in biomedical domains such as EEG, ECG and EMG signals which are quite complex but the proposed approach can play a crucial role make proper classification and precise diagnostics. Thus, the developed approach is robust and has wide applicability in various time-series signals.

6 Conclusion

This paper presents the one-dimensional convolutional neural network i.e. HindiSpeech-Net for the classification of speech for speech recognition in the Hindi language. For this purpose, different Hindi speech signals are recorded for ten different number from zero to nine in the voice of different speakers. Recorded speech signals are pre-processed before training a neural network. Signals are augmented as well to make a robust recognition system that can help to develop the robust speech recognition system. The proposed architecture of HindiSpeech-Net is composed of seven convolutional layers with flatten layer which classifies the input speech signal into respective classes. The proposed system is evaluated in different parameters. The proposed model achieved an accuracy of 85.12% on the test set of speech signal and performance is enhanced after the model trained with augmented signals, and an accuracy of 92.92% is obtained. High precision and recall value are achieved for the trained model of HindiSpeech-Net which shows its applicability on a wide range of audio signals for speech recognition or classification. The proposed framework can be extended to a large database and different regional languages. Speaker identification will be another future scope to determine the speaker according to the captured speech signals and evaluating its parameters.

Data availability The data used in the proposed work are available from the corresponding author upon reasonable request.

Declarations

Competing interest The authors declared no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

References

1. Adiwijaya, Aulia MN, Mubarok MS, Novia U, Nhita F (2017) A comparative study of MFCC-KNN and LPC-KNN for hijaiyyah letters pronounciation classification system. 2017 5th International Conference on Information and Communication Technology, ICoIC7 2017. https://doi.org/10.1109/ICoICT.2017.8074689
2. Alweshah M, Khalaileh S, Al, Gupta BB et al (2020) The monarch butterfly optimization algorithm for solving feature selection problems. Neural Comput Appl. https://doi.org/10.1007/s00521-020-05210-0
3. AlZu'bi S, Shehab M, Al-Ayyoub M et al (2020) Parallel implementation for 3D medical volume fuzzy segmentation. Pattern Recognit Lett. https://doi.org/10.1016/j.patrec.2018.07.026
4. Benzeghiba M, De Mori R, Deroo O, Dupont S, Erbes T, Jouvet D, Fissore L, Laface P, Mertins A, Ris C, Rose R, Tyagi V, Wellekens C (2007) Automatic speech recognition and speech variability: a review. Speech Commun. https://doi.org/10.1016/j.specom.2007.02.006
5. Bhatt S, Dev A, Jain A (2018) Hindi speech vowel recognition using hidden Markov model. The 6th intl. workshop on spoken language technologies for under-resourced languages, pp 196–199. https://doi.org/10.21437/SLTU.2018-41
6. Bhatt S, Jain A, Dev A (2020) Syllable based Hindi speech recognition. J Inform Optim Sci 41(6):1333–1351. https://doi.org/10.1080/02522667.2020.1809091
7. Dey A, Zhang W, Fung P (2014) Acoustic modeling for hindi speech recognition in low-resource settings. 2014 international conference on audio, language and image processing, pp 891–894. https://doi.org/10.1109/ICALIP.2014.7009923
8. Dong X, Yin B, Cong Y, Du Z, Huang X (2020) Environment Sound event classification with a two-stream convolutional neural network. IEEE Access 8:125714–125721. https://doi.org/10.1109/ACCESS.2020.3007906
9. Dua M, Aggarwal RK, Biswas M (2018) Performance evaluation of Hindi speech recognition system using optimized filterbanks. Eng Sci Technol Int J 21(3):389–398. https://doi.org/10.1016/j.jestch.2018.04.005
10. Dua M, Aggarwal RK, Biswas M (2019) Discriminatively trained continuous Hindi speech recognition system using interpolated recurrent neural network language modeling. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3499-9
11. Farooq O, Datta S, Shrotriya MC (2010) Wavelet sub-band based temporal features for robust hindi phoneme recognition. Int J Wavelets Multiresolut Inf Process. https://doi.org/10.1142/S0219691310003845
12. Ganapathiraju A, Hamaker J, Picone J (2004) Applications of support vector machines to speech recognition. IEEE Trans Signal Process 52(8):2348–2355. https://doi.org/10.1109/TSP.2004.831018
13. Gaudani H, Patel NM (2022) Comparative study of robust feature extraction techniques for ASR for Limited Resource Hindi Language, pp 763–775
14. Han W, Zhang Z, Zhang Y, Yu J, Chiu C-C, Qin J, Gulati A, Pang R, Wu Y (2020) ContextNet: improving convolutional neural networks for automatic speech recognition with global context. Interspeech 2020, pp 3610–3614. https://doi.org/10.21437/Interspeech.2020-2059
15. Ishizuka K, Nakatani T (2006) A feature extraction method using subband based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition. Speech Commun. https://doi.org/10.1016/j.specom.2006.06.008
16. Kong Q, Yu C, Xu Y, Iqbal T, Wang W, Plumbley MD (2019) Weakly labelled audioset tagging with attention neural networks. IEEE/ACM Trans Audio Speech Lang Process. https://doi.org/10.1109/TASLP.2019.2930913
17. Kumar A, Aggarwal RK (2020) Discriminatively trained continuous Hindi speech recognition using integrated acoustic features and recurrent neural network language modeling. J Intell Syst 30(1):165–179. https://doi.org/10.1515/jisys-2018-0417
18. Kumar A, Aggarwal RK (2020) Hindi speech recognition using time delay neural network acoustic modeling with i-vector adaptation. Int J Speech Technol. https://doi.org/10.1007/s10772-020-09757-0
19. Kumar A, Mittal V (2021) Hindi speech recognition in noisy environment using hybrid technique. Int J Inform Technol. https://doi.org/10.1007/s41870-020-00586-7
20. Kumar P, Jayanna HS (2022) Development of speaker-independent automatic speech recognition system for Kannada language. Indian J Sci Technol 15:333–342. https://doi.org/10.17485/IJST/v15i8.2322
21. Kumar A, Solanki SS, Chandra M (2022) Effect of background Indian music on performance of speech recognition models for Hindi databases. Int J Speech Technol. https://doi.org/10.1007/s10772-021-09948-3
22. Lee J, Park J, Kim K, Nam J (2018) SampleCNN: end-to-end deep convolutional neural networks using very small filters for music classification. Appl Sci 8(1):150. https://doi.org/10.3390/app8010150
23. Li F, Liu M, Zhao Y, Kong L, Dong L, Liu X, Hui M (2019) Feature extraction and classification of heart sound using 1D convolutional neural networks. EURASIP J Adv Signal Process 2019(1):59. https://doi.org/10.1186/s13634-019-0651-3

24. Liu Z, Wang Y, Chen T (1998) Audio feature extraction and analysis for scene segmentation and classification. Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology. https://doi.org/10.1023/A:1008066223044
25. Mustafa MK, Allen T, Appiah K (2019) A comparative review of dynamic neural networks and hidden Markov model methods for mobile on-device speech recognition. Neural Comput Appl. https://doi.org/10.1007/s00521-017-3028-2
26. Mustaqeem, Kwon S (2020) A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sens (Switzerland). https://doi.org/10.3390/s20010183
27. Muzammel M, Salam H, Hoffmann Y, Chetouani M, Othmani A (2020) AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis. Mach Learn Appl. https://doi.org/10.1016/j.mlwa.2020.100005
28. Nanni L, Costa YMG, Aguiar RL, Mangolin RB, Brahnam S, Silla CN (2020) Ensemble of convolutional neural networks to improve animal audio classification. Eurasip J Audio Speech Music Process. https://doi.org/10.1186/s13636-020-00175-3
29. Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T (2015) Audio-visual speech recognition using deep learning. Appl Intell 42(4):722–737. https://doi.org/10.1007/s10489-014-0629-7
30. Oh D, Park J-S, Kim J-H, Jang G-J (2021) Hierarchical Phoneme Classification for Improved Speech Recognition. Appl Sci 11(1):428. https://doi.org/10.3390/app11010428
31. Oneață D, Cucu H (2019) Kite: automatic speech recognition for unmanned aerial vehicles. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. https://doi.org/10.21437/Interspeech.2019-1390
32. Purwins H, Li B, Virtanen T, Schluter J, Chang S-Y, Sainath T (2019) Deep learning for audio signal processing. IEEE J Selec Topics Signal Process 13(2):206–219. https://doi.org/10.1109/JSTSP.2019.2908700
33. Samudravijaya K, Murthy HA (2012) Indian language speech sound label set (ILSL12), 2012 developed by Indian Language TTS Consortium & ASR Consortium retrieved from https://www.iitm.ac.in/donlab/tts/downloads/cls/cls_v2.1.6.pdf. Accessed 21 Feb 2021
34. Sertolli B, Ren Z, Schuller BW, Cummins N (2021) Representation transfer learning from deep end-to-end speech recognition networks for the classification of health states from speech. Comput Speech Lang 101204. https://doi.org/10.1016/j.csl.2021.101204
35. Sharma A, Shrotriya MC, Farooq O, Abbasi ZA (2008) Hybrid wavelet based LPC features for Hindi speech recognition. Int J Inf Commun Technol 1(3/4):373. https://doi.org/10.1504/IJICT.2008.024008
36. Sharmila, Mishra AN, Awasthy N, Verma V, Malhotra S (2020) Hindi speech audio visual feature recognition. Int J Adv Sci Technol
37. Wang H, Li Z, Li Y et al (2020) Visual saliency guided complex image retrieval. Pattern Recognit Lett. https://doi.org/10.1016/j.patrec.2018.08.010
38. Yu C, Li J, Li X et al (2018) Four-image encryption scheme based on quaternion Fresnel transform, chaos and computer generated hologram. Multimed Tools Appl. https://doi.org/10.1007/s11042-017-4637-6
39. Zahid S, Hussain F, Rashid M, Yousaf MH, Habib HA (2015) Optimized audio classification and segmentation algorithm by using ensemble methods. Math Probl Eng. https://doi.org/10.1155/2015/209814