International Journal of Speech Technology (2022) 25:499–509
https://doi.org/10.1007/s10772-022-09976-7

Soft-computation based speech recognition system for Sylheti language

Gautam Chakraborty[1] · Mridusmita Sharma[2] · Navajit Saikia[1] · Kandarpa Kumar Sarma[2]

Received: 23 December 2020 / Accepted: 2 May 2022 / Published online: 25 May 2022

Abstract

The encouraging trend of usage of human machine interfaces in diverse areas has driven the evolution of Automatic Speech Recognition (ASR) systems during last two decades. Lately, the inclination has been towards the use of machine learning techniques for under-resourced human languages primarily to focus on designing of voice activated digital tool for a sizable portion of computer illiterate speakers. A vast majority of the works in this field have employed shallow models like conventional Artificial Neural Network and Hidden Markov Model in combination with Mel Frequency Cepstral Coefficients and other relevant features for the applications of speech recognition systems. Although these shallow models are found effective, but to minimize human intervention from the approach and also to yield the better system performance, recent research has focused to incorporate deep learning models for ASR applications especially for under-resourced languages. Sylheti language, a member of Indo-Aryan language group, is an under resourced language which has more than 10 million Sylheti speakers living across the world mostly in India and Bangladesh. Focusing on the need of an ASR model for Sylheti, this work aims to design a robust ASR model for an under resourced language Sylheti by employing state-of-the-art deep learning technique Convolutional Neural Network (CNN). To find out the best and suitable ASR model for Sylheti, certain ASR approaches are formulated and trained by Sylheti isolated and connected words. The specially configured ASR model based on CNN is trained with clean, and noisy speech data which are necessary for training and making the system robust. Thereafter, a comparative analysis is presented by configuring the ASR model by some shallow models like Feed-forward neural network, Recurrent neural Network, Hidden Markov model and Time Delay neural Network. Experimental results indicate that the proposed CNN based ASR system works well for Sylheti language and the performance accuracy obtained by the system is found to be satisfactory despite the system demonstrating certain training latency.

Keywords Automatic speech recognition · Mel frequency cepstral coefficient · Under resourced language · Sylheti · Time delay neural network · Convolutional neural network

1 Introduction

Growing popularity of usage of smart devices has led to the attractive application of automatic speech recognition (ASR) framework among people during last two decades. Speech recognition technologies play an important role in increasing the use of speech interface for native and illiterate speakers (in large proportion) who are residing in remote areas in multilingual countries like India. Man–machine interaction system becomes easier and effective through the use of a speech interface software (SIS) which offers easy handling features rather than the conventional input tools like keyboard and mouse, especially for differently-abled person. Automatic speech recognition (ASR) system is considered the most important component of an SIS, which offers

✉ Gautam Chakraborty
gauchak2012@gmail.com

Mridusmita Sharma
mriduzb@gmail.com

Navajit Saikia
navajit.ete@aec.ac.in

Kandarpa Kumar Sarma
kandarpaks@gmail.com

[1] Department of Electronics and Telecommunication Engineering, Assam Engineering College, Guwahati, Assam 781013, India

[2] Department of Electronics and Communication Engineering, Gauhati University, Guwahati, Assam 781014, India

machine learning capability. Despite significant advances, ASR still remains one of the most challenging issues in speech and language research in terms of its robustness and noise tolerability.

ASR is the technology which maps a speech signal from a continuous-time sequence to a sequence of discrete entities called phonemes (or speech sounds), words, and sentences (Rabiner & Juang, 1993). Speech has been the most efficient communication medium through which human beings exchange their thoughts among each other. Being a complex signal, speech carries not only the message but also holds speaker's identity, linguistic entities, dialects, emotions and so on (Sharma & Sarma, 2016). As the application of ASR system has been increasing significantly over the last few decades, development of the ASR system and related research in well-resourced human languages such as English, Russian, German, Japanese, Mandarin, etc. have attained a significant level (Dhanashri & Dhonde, 2017; Kunze et al., 2017; Padmanabhan & Johnson Premkumar, 2015; Waibel et al., 1989; Wang et al., 2019). In most recent years, various researchers have contributed their work in the field of ASR considering various under-resourced (UR) human languages (Alotaibi et al., 2010; Besacier et al., 2014; Nassif et al., 2019; Sharma et al., 2013). Sylheti, an under-resourced language, belongs to Indo-Aryan language group, and according to Ethnologue, more than 10 million people speak Sylheti as their first language across the world. Out of which, major section of Sylheti speakers resides in Bangladesh and a partial section live in Assam and Tripura, India. It has been observed from the literature that Sylheti has been explored with very limited scientific work in linguistic as well as in technological context (Chakraborty & Saikia, 2019; Gope, 2018). Though English is a globally recognized language, out of the world's total population, only 1.5 billion speak English (i.e. 20% of the Earth's population). Moreover, most of those people are not native English speakers. According to 2011 census report in India, it is estimated that only 15% population of India speak English as their native language. As a result, it has prompted the researchers to study about native as well as under resourced (UR) languages like Sylheti, to bridge the digital gap so that the English illiterate section can also take the advantage of the technology. In the field of speech recognition, there are various methods of parametric and non-parametric representation of acoustic waveforms (Sharma & Sarma, 2017). Among the various methods, Mel Frequency Cepstral Coefficients (MFCC) is the most widely used technique because this acoustic feature can provide cepstral description of an audio clip (Shrawankar & Thakare, 2013). A major portion of the previously reported works commonly have preferred to use some shallow models like Recurrent Neural Network (RNN), Time Delay Neural Network (TDNN), Hidden Markov Model (HMM), etc. as common tools for speech-based application(Bhardwaj & Londhe, 2012; Peddinti et al., 2015; Sharma et al., 2013). RNN is considered essential in many speech processing applications as it tracks time variations and mimics the action of human brain with an assembly of feed-forward and feedback connections. TDNN is also another technique which has the ability to capture the relationship between the input and output data with variations in time (Waibel et al., 1989). On the other hand, HMM is a standard statistical tool for signal processing applications. All these tools have established their credentials in a range of applications. These shallow models in combination with MFCCs have been found to be adequate for a range of conditions and demonstrated significant capability to imitate the human brain speech perception mechanism till the concepts of deep learning evolved.

Deep neural network architectures have gained much popularity in the recent attempts to enhance the performance of ASR systems. Further, to a large extent, the deep learning mechanisms have been able to overcome the problems of the shallow features and the conventional classifiers (Young et al., 2018). Recent studies establish that the signal processing in the human brain can be better depicted by the deep learning mechanism because it can find out more abstract features from the input data that represents the spectral and temporal information as interpreted by the brain (Cox & Dean, 2014). It is noteworthy that the language shows some notable variations with respect to geographical distributions and speaker to speaker ethnographic background. These make it important to develop an efficient language specific system based on contemporary processing and classification approaches for reliable speech and speaker recognition system. A remarkable change in the state of the art makes the ASR systems more and more effective and used in many applications like assistance to independent living of people, voice control, language learning and translation, etc.

The organization of the paper is as follows. Section 2 presents some of the related works in speech processing field in recent times. Section 3 gives a brief overview of the basic theoretical considerations related to the proposed work. In Sect. 4, the proposed system model and the speech dataset of the proposed ASR model are discussed. In Sect. 5, details of the experimental findings are presented. Section 6 concludes the work.

2 Related work

Despite significant advances in ASR over the last decades, humans still outperform the machine learning systems, especially on recognition tasks in noisy background conditions. Deep learning algorithms shows the driving force behind state-of-the-art algorithms for machine learning capabilities, and for natural language processing (Goldberg et al., 2018;

Xie et al., 2018) to make available a robust and efficient ASR for the end users. Feeding the inputs into the multiple layered structure in deep learning technology has proven to be more efficient while focusing on non-linear functions with fewer parameters in comparison with the large number of parameters required to represent the same functions in a shallow architecture (Nassif et al., 2019). Two ASR systems in Amazigh under resourced language are presented in (Telmem & Ghanou, 2020) by applying CNN approach and conventional HMM classification model. In this work, CNN based system shows better recognition accuracy than HMM based system. Few works presented by Microsoft in the area of speech in recent times by employing deep learning techniques is deliberated by Deng and his research associates (Deng et al., 2013). This paper shows that speech recognition and related applications such as spoken dialogue and language modeling works can improve the system performance with the application of deep neural networks compared to traditional practice of using GMMs-HMMs. A comprehensive study is carried out and presented in (Nassifet al., 2019) by Nassif and his co-partners providing speech recognition works considering various deep learning methods on multiple languages. Authors in this paper have focused on using DNN-HMM based hybrid models and long short-term memory (LSTM) RNN in designing ASR systems as they yield better results. In another speech related work, the application of DNN in voice command recognition work is presented (Sokolov & Savchenko, 2019). A hybrid architecture mixing of CNN and bidirectional LSTM investigated in (Passricha & Aggarwal, 2020) has proved to be computationally efficient by achieving optimal result performance in speech recognition task.

These scientific developments especially in ASR technologies for various human languages and also the recent trend of application of deep learning in designing a robust ASR system motivate us to explore the Sylheti language in the broad applications of ASR. Sylheti language has its own distinct script Syloti Nagari wherein 32 Sylheti alphabets are inscribed. Distinctive way of pronunciation, de-aspiration and deaffrication, etc. are some unique characteristics of this language (Gope, 2018). Due to the availability of very few open-access resource in this language, Sylheti is still unexplored and under-resourced language. Considering the necessity of a robust ASR model to investigate the unique features of this language, we propose to design two ASR systems for the Sylheti language: one for isolated words and other one for connected words considering speech samples from the Sylheti speech dataset. As deep learning is the state-of-the-art in recent times and CNN becomes the leading architecture of deep learning approach, we are focusing on designing and developing of Sylheti speech recognition system by employing a new CNN based approach. The clean speech samples and derived noisy speech are considered to train with the softmax classifier, and performance is compared with benchmark techniques like RNN, FFNN, TDNN, and HMM.

The next section explains the theoretical definition of relevant modules of a generic ASR system.

3 Basic considerations

In this section we have included basic theoretical aspects that are relevant for designing the ASR system.

3.1 Pre-processing

Pre-processing is a very important stage in ASR system which involves a series of signal analysis on the input of speech data. It is essential in order to achieve high recognition accuracy. After a series of signal analysis steps through analog-to-digital (A/D) conversion, silence and unvoiced part removal, end-point detection, pre-emphasis filtering, and finally windowing, the windowed version of the voiced part of each signal is derived. In this proposed work, each speech signal is recorded with 16 kHz sampling frequency and quantized the sample by 16 bits to derive a digital speech signal. End point detection process is implemented by locating the beginning and end points in the utterance manually to extract the voiced part. Thereafter, a pre-emphasis filter is used to boost the amplitude the high-frequency components of the extracted voiced part. For the proposed ASR systems, a first-order high-pass finite impulse response (FIR) filter is applied on the voiced part of each speech signal according to

$$x_p[n] = x_v[n] - 0.95\, x_v[n-1]; \tag{1}$$

where, $x_v[n]$ is the input signal (voiced part) to the pre-emphasis filter and $x_p[n]$ is the output.

Due to time varying nature of speech signal, voiced part of the signal is segmented into short *frame*s which are assumed to be stationary and speech analysis is carried out on the frames. In this work for Sylheti, frame duration of 32 ms with an overlap of 10 ms are considered. Thereafter, a windowing operation is performed on frames to minimize the signal discontinuities at the borders of the frame. In this study, Hamming windowing w[n]on each frame is implemented using the following equation:

$$w[n] = 0.54 - 0.46\cos\left[\frac{2\prod n}{N-1}\right]; \quad 0 \le n \le N-1 \tag{2}$$

where N is the number of frames in a speech sample and n refers to the current sample.

3.2 MFCC feature extraction

Feature extraction is the process of obtaining parameter representation of the speech signal through differentiation and concatenation processes. There are numerous methods for feature extraction (Shrawankar & Thakare, 2013). MFCC is the most prominent one and established feature extraction technique, the one used in this work, which has vast application in automatic speech and speaker recognition systems using the Mel scale. Mel scale is based on the human ear scale (Shrawankar & Thakare, 2013). MFCCs are the coefficients that collectively make up a Mel frequency cepstrum (MFC). It is based on the nonlinear human perception of the frequency of sounds. These MFCC coefficients represent audio based on perception. Since the MFCC can represent a listener's response system clearly, so people are applying MFCC as an established feature in speech recognition system. The proposed ASR work for Sylheti language derives and employs a set of 13 MFCC coefficients as the representative features for a frame. The basic architecture of MFCC feature extraction process is shown in Fig. 1.

Initially, the discrete Fourier transform (DFT) is applied to the windowed version of each speech frame (in time domain) to derive the frequency components(spectrum). It is computed by employing fast Fourier transform (FFT) for each frame. In next step, the mel-scaled filter bank having linearly spaced filters is applied on the spectrum. The speech signal is not linear and human ear is less sensitive to higher frequencies. Due to wide ranged feature of FFT spectrum, Mel-Scaled Filter Bank is used to convert it to mel-scale. Mel-scale is roughly linear below 1 kHz and logarithmic above 1 kHz. The conversion formula from the frequency scale to the mel-scale is given as-

$$f_{mel} = 2595\log_{10}\left[1 + \frac{f_{linear}}{700}\right]; \tag{3}$$

where f_{mel} is the mel frequency corresponding to the linear frequency f_{linear}. Finally, log is taken from the output f_{mel} and discrete cosine transform (DCT) is applied to it to obtain the magnitudes of the resulting spectrum.

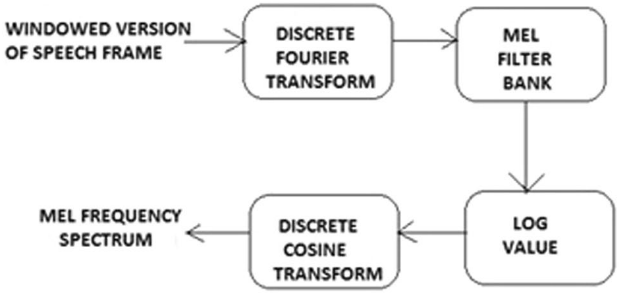
WINDOWED VERSION OF SPEECH FRAME
DISCRETE FOURIER TRANSFORM
MEL FILTER BANK
LOG VALUE
DISCRETE COSINE TRANSFORM
MEL FREQUENCY SPECTRUM

Fig. 1 Computation of MFCC feature

3.3 Classifier

Speech recognition is a kind of pattern recognition problem, and hence the goal of a classifier in ASR system is to partition feature space into class-labelled decision regions. Post-processing operation in ASR system involves the application of classifier. During classification, decisions are made based on the similarity observed from training patterns using information relating to known patterns. Then, they are tested using the unknown patterns. Brief description about CNN along with other four classifiers, which are employed in this proposed work, are stated here.

1. FFNN: FFNN is known as the most popular and simplest type of artificial neural network (ANN) model, which is composed of layered structured of neurons. The concept behind the standard FFNN network is that the flow of information (in the form of neurons) takes place constantly in the forward direction from the single input layer to the output layer (single) through one or multiple hidden layer(s). The general layout of FFNN comprising of one hidden layer is shown in Fig. 2. FFNN algorithm computes a function *f* on fixed size input x such that *y=f(x)* for a training pairs *(x,y)*. During forward propagation, two operations pre activation and activation take place at each node of hidden and output layer in this network (Fausett, 2006).
2. RNN: RNN is another variant of ANNs where information processes in loops from layer to layer so that the

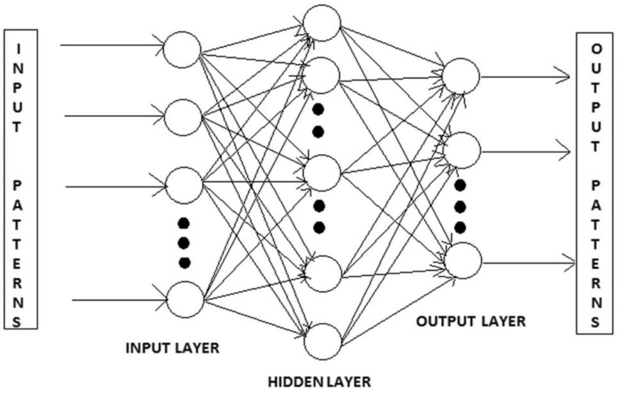

Fig. 2 FFNN architecture

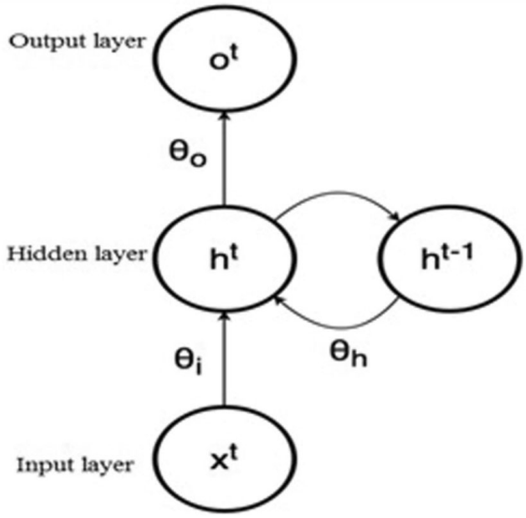

Fig. 3 Working structure of RNN

state of the model is influenced by its previous states. Unlike stateless FFNN, RNN network has an internal memory which allows it to store information about its past computations. Thereby, the RNNs provides dynamic temporal behavior and model sequences of input–output pairs. RNN can compute its output at time *t* based on the information computed at a previous time *(t—1)*. When it is required to predict the next word of a sentence, the state of previous words is required and hence there is a need to remember the previous words. This issue is resolved with the use of the internal memory (Gevaert et al., 2010). A simplest working structure of RNN model is shown in Fig. 3. The notations used in this structure represent as:

x^t denotes the input to the RNN at time t, h^t is the state of the hidden layer(s) at time t, o^t is the output of the RNN at time t, and Θ represents the weight value in the communication link between two layers.

3. TDNN: A TDNN is another class of ANNs which is similar to FFNN except that it is associated with a time delay in its input neuron and hidden neuron. It provides a simple way to represent a mapping between past and present values. The inputs to any node *i* can have the outputs of earlier nodes not only during the current time step, but during some amount of delay *d* of previous time steps (*t*–1, *t*–2,…., *t*–*d*) (Waibel et al., 1989). Figure 4 depicts the general architecture of TDNN.
4. HMM: HMM is a statistical tool that models a system based on Markov chain (Alotaibi et al., 2010; Bhardwaj & Londhe, 2012). A Markov chain tells us something about the probabilities of sequences of random variables, states, each of which can take on values from some sets. These sets may include words, or tags, or symbols representing anything. It attempts to find out the hidden parameters from the observed ones. Speech recognition uses a slightly adapted Markov Model which accepts the smallest speech entities as states in the Markov Model. When a word enters the Hidden Markov Model, it is compared with the best suited model (entity). According to transition probabilities, there exist a transition from one state to another. A state can also have a transition to it's own if the sound repeats itself. An HMM model can be formulated by

   - a set of states Q,
   - a set of transition probabilities A,
   - a set of observation likelihoods B,
   - a defined start state and end state(s), and
   - a set of observation symbols O, which is not drawn from the same alphabet as the state set Q

5. CNN: CNN architecture differs from the traditional multilayer perceptron to ensure some degree of shift and distortion invariance. This type of network is a kind of discriminative deep architecture which consists of one or more convolutional layer and a pooling layer. A common architecture of CNN with basic components is presented in Fig. 5.

   Basically, convolutional networks used in speech processing include trainable multistage architecture with each stage consisting of multiple layers (Nagajyothi & Siddaiah, 2018). The input and output of each stage are sets of 1D arrays called feature maps. The output stage represents features extracted from all locations on the input. Each stage generally consists of a convolution layer, nonlinearity and a pooling layer. A single or multiple fully connected layers operate after several convolution and pooling layers. Similar to shallow ANN, weights are shared in the convolutional layer whereas the pooling layer sub-samples the output generated by the convolutional layer and decreases the data rate of the below layer of the stacked model.

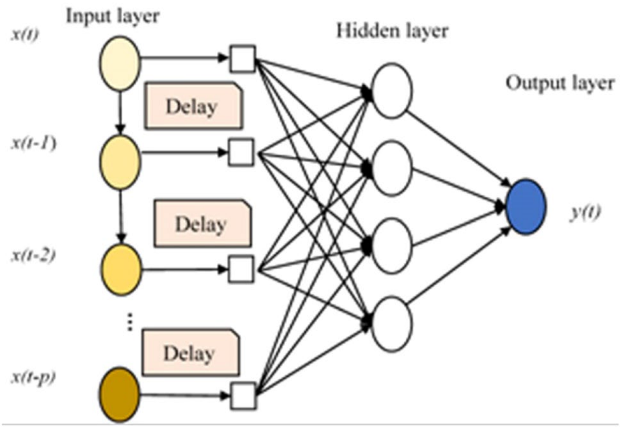

Fig. 4 TDNN architecture

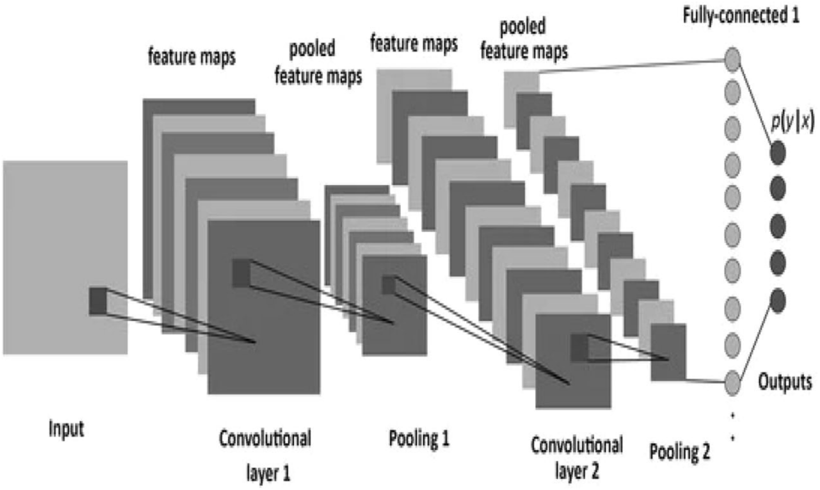

Fig. 5 CNN working model

4 Proposed ASR work

In this section, we discuss the ASR models designed for the recognition of Sylheti words using soft computation models. Proposed ASR model includes pre-processing and post-processing operations. Steps starting from recording of speech signals to feature extraction are termed as pre-processing. On the other hand, post-processing involves modelling of feature vectors and classification of words to be recognized through training and testing phases respectively. The system architecture of the proposed model is depicted in the Fig. 6.

Speech signals are pre-processed first according to the steps mentioned in Sect. 3. Thereafter, pre-processed signals are fetched into convolutional layer followed by a pooling layer in feature learning stage. Finally, after investigating pre-processed signals passing over two convolution layers, feature maps resulting out from the feature learning stage are retrieved by classification stage. A fully connected layer in classification stage accepts the results of the convolution/pooling process and uses them to classify the words into a label. Preparation of speech dataset which is the most important component required for speech recognition is mentioned in the following.

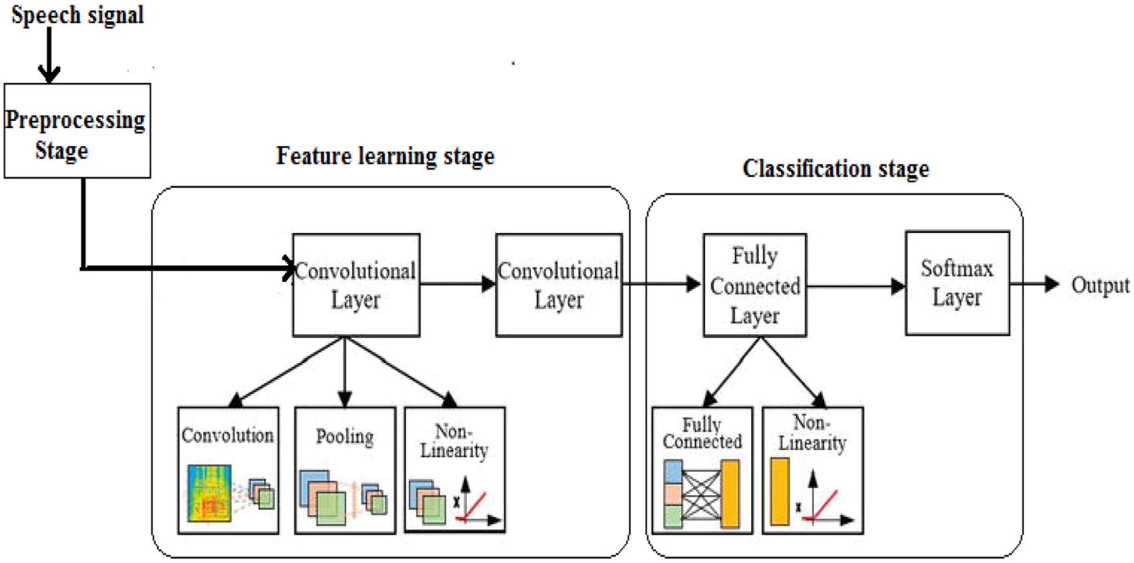

Fig. 6 System architecture of the proposed ASR model

4.1 Dataset

The primary requirement for an ASR system is the data sample of a speech database. Samples from such database are used for training and testing of ASR system to establish the machine learning capability of language. As mentioned in Sect. 1, speech samples are collected for isolated words and connected words to design ASR systems for the Sylheti language in two separate sessions. In first session, 30 isolated Sylheti words including 10 digits (0–9) are selected for creating the database for isolated words. Table 1 presents ten digits from 0 to 9 in Sylheti with their pronunciations. The speech samples are recorded in a closed room environment. The 10 native speakers of Sylheti speaking areas (in the age group of 25 to 70 years) are told to utter the selected 30 isolated words.

Each speaker utters each word 10 times. Thus, a set of 3000 clean speech samples are derived. In order to enhance the training data and also to find the robustness of the proposed ASR systems, noisy speech of different SNR values (1 dB, 0 dB and -1 dB) are obtained by adding Gaussian white noise of these SNR values into the clean speech of 3000samples. Altogether 12,000 speech samples (3000 × 4) are derived, and these samples are used by the proposed ASR models for isolated words in our first set of experiments. Further, to prepare speech data for connected words in Sylheti, the mostly spoken 10 Sylheti sentences are chosen. The same native speakers (10 nos) have contributed their voices by reading out 10 sentences in closed room environment in second phase of recording. Total utterances of 3000 clean samples of connected words are collected in this session when each sentence is uttered by each speaker 30 times.

In order to increase the data set for training, noisy speech of SNR values of 1 dB, 0 dB, and − 1 dB are added to the clean speech of 3000 samples independently. Thus, another set of 9000 samples of noisy speech are derived for connected words. Thus, complete set of 12,000 utterances are considered for designing connected word ASR system for the Sylheti language. As recording essentials, a unidirectional iBall microphone and PRAAT voice recording software are used. In the process of recording in both sessions, the sampling rate of 16 kHz and mono channel mode are chosen. Speakers are asked to sit in front of the microphone at a distance of 10–12 cm before giving their voices. In case of connected words in Sylheti, all the clean speech of 3000 samples along with the noisy speech samples of 3000 each for SNR 1 dB, SNR 0 dB, and SNR − 1dB are selected to exercise independently. From each set of 3000 samples, 1800 samples are considered for training and the remaining 1200 are chosen for testing. These training and testing data are independently used by each classifier of the proposed systems.

5 Experimental details and results of the proposed ASR systems

This section includes discussion on the experimental results obtained from trials carried out to fix the layer sizes of the neural network architectures used for the purpose, performance of the proposed system and comparison with benchmark techniques, covering training time learning curves. The experimental results for the recognition of Sylheti language by using various softmax classifiers are summarized here.

In this work, we propose two sets of ASR systems for Sylheti by employing CNN model: one for isolated words and another for connected words. The first set of experiments are carried out on 3000 clean speech samples of Sylheti isolated words from the constructed Sylheti speech database along with derived noisy speech of 3000 samples each for three SNR values of 0 dB, 1 dB and − 1 dB. Among these 3000 samples for each case, 1500 speech samples (30 words × 5 speakers × 10 times) are used for training and another set of 1500 speech samples (30 words × 5speakers × 10 times) are used for testing independently. Second work concentrates on to carry out experiments on Sylheti connected words. This set of experiments are exercised on 3000 clean speech samples of10 Sylheti sentences and noisy speech of 3000 samples each derived for SNR 0 dB, 1 dB and − 1 dB. Out of 3000 samples for each case of data, training set comprises of 1800 samples (10 sentences × 6 speakers × 30times) and that of testing set includes 1200 samples ((10 sentences × 4 speakers × 30 times).

The following parameters are considered for proposed system in order to carry out the experiments to design machine learning based ASR models for recognizing isolated and connected words in Sylheti:

- CNN Classifier structure: 2 convolutional layers, 1 max pooling layer and 1 fully connected layer.
- Activation function: ReLU.
- Loss function: Softmax cross entropy.
- Performance metric: Recognition accuracy in % (i.e. RA in %).

Recognition accuracy is the ratio of “Number of correct word recognition" to “Total number of word utterances used in testing". In Sylheti isolated word recognition process while exercising 1500 clean speech samples by the proposed CNN

Table 1 Pronunciation of ten English digits from 0 to 9 in SYLHETI

Pronunciation/English digits	0	1	2	3	4	5	6	7	8	9
In Sylheti	Shuinno	Ex	Dui	Tin	Sair	fas	Soe	Shat	At	Noe

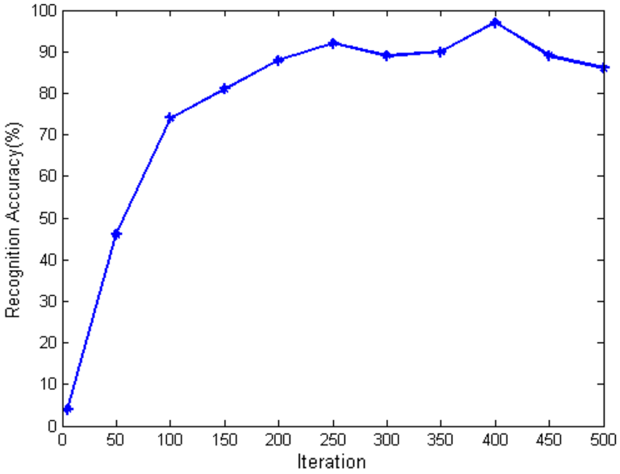

Fig. 7 Training recognition accuracy graph for the ASR model of Sylheti isolated word using clean speech

Table 2 Recognition accuracy and loss with respect to iteration

Batch	Iteration	Accuracy	Loss
1	5	4	3.07
2	25	27	2.46
3	50	46	1.68
4	75	65	1.11
5	100	74	0.76
6	125	82	0.64
7	150	81	0.68
8	175	89	0.37
9	200	88	0.39
10	225	88	0.39
11	250	92	0.29
12	275	85	0.41
13	300	89	0.3
14	325	88	0.33
15	350	90	0.3
16	375	86	0.32
17	400	97	0.11
18	425	92	0.26
19	450	89	0.28
20	475	87	0.35
21	500	86	0.35

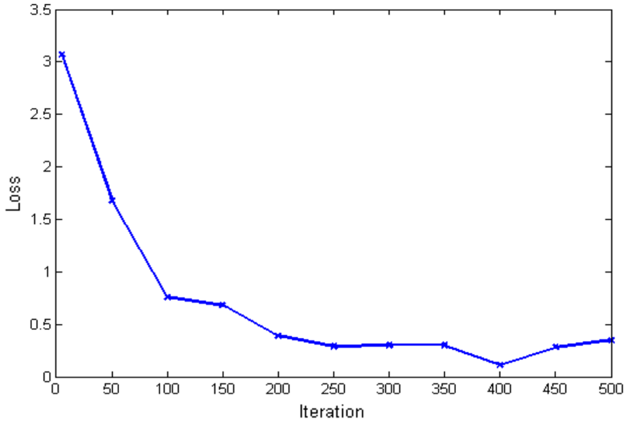

Fig. 8 Loss vs iteration graph for the ASR model of Sylheti isolated word using clean speech

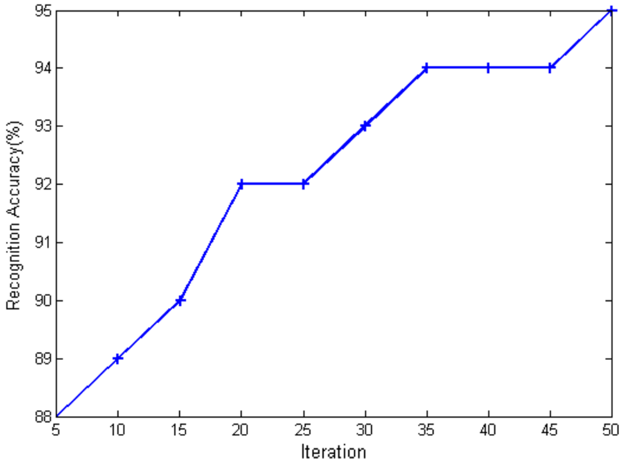

Fig. 9 Testing recognition accuracy graph for the ASR model of Sylheti isolated word using clean speech

based ASR model during training, it is observed that training accuracy tends to increase when training iteration changes where number of iteration is considered 500. After 500 iterations, 88% of average training accuracy is obtained by the model. The training recognition accuracy graph for the ASR model using clean speeches is presented in Fig. 7 based on the recognition accuracy results generated (as listed in Table 2) during training.

Similarly, loss incurred by the model during training with respect to iteration gradually decreases which is shown in Fig. 8. Once training model is built, it is tested with testing dataset of another 1500 clean samples. The average testing accuracy of 93% is achieved in this work. The plot for testing accuracy for clean speech is presented in Fig. 9. Similarly, noisy speeches of SNR values of 0 dB, 1 dB and − 1 dB are also exercised independently by employing CNN algorithm to design ASR models for Sylheti isolated word. The average training and testing recognition accuracies (RA) in % obtained by all the ASR models for Sylheti isolated

Table 3 Training and testing RA obtained by different approaches of the ASR system for SYLHETI isolated word

Speech data	% Training accuracy (in average)	% Testing accuracy (in average)	Computational time during training (in seconds)
Clean speech	88	93	4785.947731
Noisy speech of SNR 0 dB	85	89	4681.845563
Noisy speech of SNR 1 dB	86	90	3789.947732
Noisy speech of SNR -1 dB	82	86	4966.734452

Table 4 Training and testing RA obtained by different approaches of the ASR system for SYLHETI connected word

Speech data	% Training accuracy (in average)	% Testing accuracy (in average)	Computational time during training (in seconds)
Clean speech	98	95	13,594.519325
Noisy speech of SNR 0 dB	93	91	13,455.643312
Noisy speech of SNR 1 dB	90	92	14,677.845561
Noisy speech of SNR—1 dB	88	90	15,122.871281

word by employing CNN model are listed in Table 3. These results are comparable to the similar works reported by the researchers by using CNN model while proposing ASR systems for various other languages (Nagajyothi & Siddaiah, 2018; Wang et al., 2019).

Another experimental model has been designed for Sylheti connected words using the same setup of CNN as described above. Only variation of this setup is to consider 1000 no of iterations. Four sets of speech samples such as clean speech and noisy speech of SNR 0 dB, 1 dB and − 1 dB are considered to carry out the experiments.

Each of the ASR approaches is trained with 1800 speech samples. Once the ASR model is built, it is tested with other 1200 samples. The training and testing recognition accuracies obtained by the models are listed in Table 4. The training and testing recognition accuracy graph of CNN based ASR models for Sylheti connected word using clean speech samples are depicted in Figs. 10 and 11.

These CNN based ASR approaches for Sylheti are compared with conventional acoustic models, and to accomplish this task, experiments are carried out on MFCC features of speech signals by using FFNN, RNN, TDNN, and HMM techniques independently. Table 5 provides a comparative analysis of RA(%) obtained by the proposed machine learning based ASR models for Sylheti isolated and connected words by employing FFNN, RNN, TDNN, HMM, and CNN classifiers.

Further, due to unavailability of any existing ASR system in Sylheti, a comparative performance analysis is carried out among the proposed CNN based ASR systems for Sylheti and those of ASR systems designed for different languages by applying various learning techniques. In Table 6 a comparative analysis of our proposed model with various state-of-art models present in the literature are shown. From

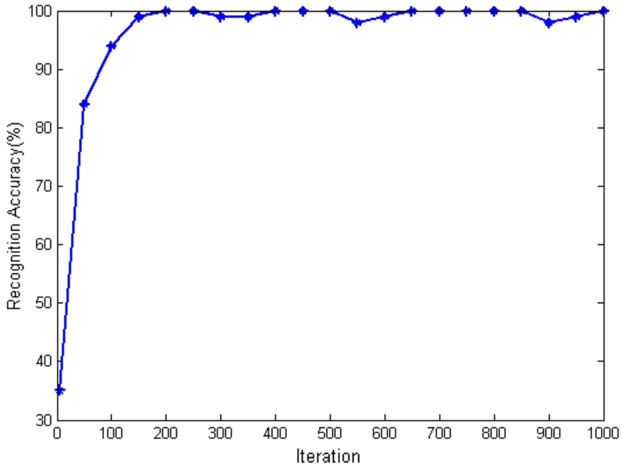

Fig. 10 Training recognition accuracy graph for the ASR model of Sylheti connected word using Clean speech

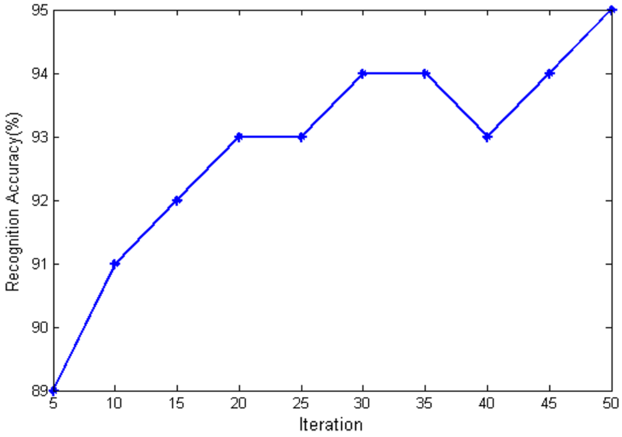

Fig. 11 Testing recognition accuracy graph for the ASR model of Sylheti connected word using clean speech

Table 5 Comparative analysis of RA obtained by the ASR models for the Sylheti language by various classifiers

Speech data	RA (%) obtained by the classifier				
	FFNN	RNN	TDNN	HMM	CNN
ASR models for Sylheti isolated words					
Clean speech	89	91	92	91.7	93
Noisy speech of SNR 0 dB	82	84	86.5	83	89
Noisy speech of SNR 1 dB	81	83.5	85	80	90
Noisy speech of SNR − 1 dB	78.8	82	86	81	86
ASR models for Sylheti connected words					
Clean speech	90.8	91	93.5	91.8	95
Noisy speech of SNR 0 dB	84	86	87	83.8	91
Noisy speech of SNR 1 dB	82.5	84	85	81	92
Noisy speech of SNR − 1 dB	80	83	84.5	78	90

Table 6, it can be stated that the CNN based proposed ASR models for Sylheti language have shown satisfactory performances and provide distinct advantage in terms of recognition accuracy. It is also obvious from the above results that the CNN model is capable of capturing and dealing with the attributes of the Sylheti speech samples. Also, CNN as a deep learning model also has the ability to capture the information present in speech samples in multiple dimensions. The reason behind the fact that CNN acoustic model performs better than TDNN, HMM and RNN in this experiment is due to application of filters and MaxPooling to normalize speaker variance in speech signals. Further, CNN framework has inbuilt denoising capability and can take care of disturbances and small shifts in the feature. This helps to summarise that CNN structure is better than other acoustic models investigated here for speech recognition task.

6 Conclusion

In this paper, a group of learning-based ASR models for an under resourced language Sylheti are presented. CNN model of deep learning is implemented here to obtain the acceptable recognition results. Experimental results obtained in this manner show that the performances of proposed ASR systems for both isolated word and connected word are found to present a satisfactory performance for recognizing Sylheti language. The computational time and the performance accuracy of these systems are also evaluated. The CNN based models presented in this work achieve a recognition accuracy of 93% for isolated Sylheti word recognition and 95% for connected word recognition in Sylheti. Despite of certain limitations in terms of dataset and training latency (which can be addressed using specialized GPUs), the proposed model may be considered a suitable one for recognition of an under resourced language like Sylheti. Further improvements may be achieved by considering unsupervised pre-training and repeated pooling operations in more than one layer. It is also strongly believed that this work on Sylheti facilitates the scope for more research on Sylheti to resolve many challenges like designing of algorithms for speaker and environment adaptation in deep neural network, creation of a large vocabulary Sylheti speech corpus with

Table 6 Performance accuracy of various classifiers in comparison to our proposed model

Reference	Language	Classifier	Recognition rate (%)	Word error rate (%)	Remarks
(Dhanashri & Dhonde, 2017)	English	DNN	86.06	–	The recognition accuracy obtained by our proposed model in recognition of Sylheti words is approximately 10–20% better than the existing models found in literature for the recognition of other human languages
(Kunze et al., 2017)	German	CNN	–	58.36	
(Hori et al., 2016)	Japanese	LSTM RNN	–	22.6	
(Wang et al., 2019)	Mandarin	CNN + BLSTM	–	19.2	
(Sumon et al., 2018)	Bangla	CNN	74.01	–	
(Deka et al., 2018)	Assamese	GMM + HMM	95.7	–	
(Deka et al., 2018)	Assamese	SGMM + HMM	95.9	–	
(Passricha & Aggarwal, 2020)	Hindi	CNN + BLSTM	–	17.8	
(Kimanuka & Buyuk, 2018)	Turkish	DNN + HMM	–	8.47	
(Telmem & Ghanou, 2020)	Amazigh	CNN	92	–	
		HMM	90		
Proposed model	Sylheti isolated words	CNN	93	–	
Proposed model	Sylheti connected words	CNN	95	–	

noise, designing of algorithms for emotional speech recognition for Sylheti, etc.

References

Alotaibi, Y. A., Alghamdi, M., & Alotaiby, F. (2010). Speech recognition system of Arabic alphabet based on a telephony Arabic corpus. In *International conference on image and signal processing*, (pp. 122–129). Springer.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication, 56*, 85–100.

Bhardwaj, I., & Londhe, N. D. (2012). Hidden Markov model based isolated Hindi word recognition. In *2012 2nd International conference on power, control and embedded systems* (pp. 1–6). IEEE.

Chakraborty, G., & Saikia, N. (2019). Speech recognition of isolated words using a new speech database in sylheti. *International Journal of Recent Technology and Engineering, 8*, 6259–6268.

Cox, D. D., & Dean, T. (2014). Neural networks and neuroscience-inspired computer vision. *Current Biology, 24*, R921–R929.

Deka, B., Dey, A., & Nirmala, S. R. (2018). Assamese connected digit recognition system. *International Journal of Research in Signal Processing, Computing and Communication System Design, 4*, 9–12.

Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J. et al. (2013). Recent advances in deep learning for speech research at Microsoft. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8604–8608). IEEE.

Dhanashri, D., & Dhonde, S. (2017). Isolated word speech recognition system using deep neural networks. In *Proceedings of the international conference on data engineering and communication technology* (pp. 9–17). Springer.

Fausett, L. V. (2006). *Fundamentals of neural networks: architectures, algorithms and applications*. Pearson Education India.

Gevaert, W., Tsenov, G., & Mladenov, V. (2010). Neural networks used for speech recognition. *Journal of Automatic Control, 20*, 1–7.

Goldberg, Y., Hirst, G., Liu, Y., & Zhang, M. (2018). Neural Network Methods for Natural Language Processing. *Computational Linguistics, 44*, 193–195.

Gope, A. (2018). The phoneme inventory of Sylheti: Acoustic evidences. *Journal of Advanced Linguistic Studies, 7*, 7–37.

Hori, T., Hori, C., Watanabe, S., & Hershey, J. R. (2016). Minimum word error training of long short-term memory recurrent neural network language models for speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5990–5994). IEEE.

Kimanuka, U. A., & Büyük, O. (2018). Turkish speech recognition based on deep neural networks. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 22*, 319–329.

Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J., & Stober,S. (2017). Transfer learning for speech recognition on a budget. In *Proceedings of the 2nd workshop on representation learning for NLP* (pp. 168–177). Association for Computational Linguistics. https://www.aclweb.org/anthology/W17-2620. https://doi.org/10.18653/v1/W17-2620.

Nagajyothi, D., & Siddaiah, P. (2018). Speech recognition using convolutional neural networks. *International Journal of Engineering and Technology, 7*, 133.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access, 7*, 19143–19165.

Padmanabhan, J., & Johnson Premkumar, M. J. (2015). Machine learning in automatic speech recognition: A survey. *IETE Technical Review, 32*, 240–251.

Passricha, V., & Aggarwal, R. K. (2020). A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition. *Journal of Intelligent Systems, 29*(1), 1261–1274.

Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modelling of long temporal contexts. In *Sixteenth annual conference of the International Speech Communication Association*.

Rabiner, L., & Juang, B. H. (1993). *Fundamentals of speech recognition*. PTR Prentice-Hall. Inc.

Sharma, M., Sarma, M., & Sarma, K. K. (2013). Recurrent neural network based approach to recognize Assamese vowels using experimentally derived acoustic-phonetic features. In *2013 1st international conference on emerging trends and applications in computer science* (pp. 140–143).IEEE.

Sharma, M., & Sarma, K. K. (2016). Learning aided mood and dialect recognition using telephonic speech. In *2016 International conference on accessibility to digital world (ICADW)* (pp. 163–167). IEEE.

Sharma, M., & Sarma, K. K. (2017). Soft computation based spectral and temporal models of linguistically motivated Assamese telephonic conversation recognition. *CSI Transactions on ICT, 5*, 209–216.

Shrawankar, U., & Thakare, V. M. (2013). Techniques for feature extraction in speech recognition system: A comparative study. arXiv:1305.1145.

Sokolov, A., & Savchenko, A. V. (2019). Voice command recognition in intelligent systems using deep neural networks. In *2019 IEEE 17th world symposium on applied machine intelligence and informatics (SAMI)* (pp.113–116). IEEE.

Sumon, S. A., Chowdhury, J., Debnath, S., Mohammed, N., & Momen, S.(2018). Bangla short speech commands recognition using convolutional neural networks. In *2018 International conference on Bangla speech and language processing (ICBSLP)*, (pp. 1–6).

Telmem, M., & Ghanou, Y. (2020). A comparative study of HMMs and CNN acoustic model in Amazigh recognition system. *Advances in Intelligence Systems & Computing A, 1076*, 533–540.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 37*, 328–339.

Wang, D., Wang, X., & Lv, S. (2019). End-to-end mandarin speech recognition combining CNN and BLSTM. *Symmetry, 11*, 644.

Xie, Y., Le, L., Zhou, Y., & Raghavan, V. V. (2018). Deep learning for natural language processing. In *Handbook of statistics* (Vol. 38, pp. 317–328). Elsevier.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine, 13*, 55–75.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.