BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 08/10 pp85–91
Volume 6, Number 1, March 2023
DOI: 10.26599/BDMA.2022.9020017

WTASR: Wavelet Transformer for Automatic Speech Recognition of Indian Languages

Tripti Choudhary*, Vishal Goyal, and Atul Bansal

Abstract: Automatic speech recognition systems are developed for translating the speech signals into the corresponding text representation. This translation is used in a variety of applications like voice enabled commands, assistive devices and bots, etc. There is a significant lack of efficient technology for Indian languages. In this paper, an wavelet transformer for automatic speech recognition (WTASR) of Indian language is proposed. The speech signals suffer from the problem of high and low frequency over different times due to variation in speech of the speaker. Thus, wavelets enable the network to analyze the signal in multiscale. The wavelet decomposition of the signal is fed in the network for generating the text. The transformer network comprises an encoder decoder system for speech translation. The model is trained on Indian language dataset for translation of speech into corresponding text. The proposed method is compared with other state of the art methods. The results show that the proposed WTASR has a low word error rate and can be used for effective speech recognition for Indian language.

Key words: transformer; wavelet; automatic speech recognition (ASR); Indian language

1 Introduction

Automatic speech recognition (ASR) is a significant area of research under the pattern recognition field. ASR comprises of multiple technologies for transforming the speech signals into its corresponding text. The objective of ASR is to enable machines to translate the speech signal into textual form. Many researchers worldwide are working on this problem to further improve efficiency and accuracy. And even the organizations like Amazon, Apple, Google, IBM, etc. have also developed high end speech recognition systems for English language[1]. The development of ASR for Indian languages is limited. Thus, there is a great need of development of algorithms for Indian languages. The speech signals comprise of a lot of heterogeneity in terms of language, speaker's voice, variations in the channel and so on. This heterogeneity may be owed to various factors like the gender, accent, age, environmental conditions and also the speed of the speaker. The ASR systems must be trained in a manner to overcome all these limitations. In this regard the training data length and the device with which the signal is recorded also play important roles. An ASR system is considered to be efficient if it is able to translate the speech into its corresponding text despite all these challenges. The training data for Indian languages are still scarce and the text corpus for Indian languages is limited[2]. Thus, the Indian languages need efficient ASR systems that can perform the recognition in such limited resources. Many researchers are working in the field of ASR and multiple techniques are being applied to speech-to-text conversion. Artificial neural networks (ANN) have been widely used for providing speech recognition systems[3]. Hybrid hidden Markov model (HMM) is also being used by many researchers for the purpose of ASR[4]. Speech recognition models mainly fall under two categories: acoustic model and language model. In case of acoustic model, sound signals are

• Tripti Choudhary and Vishal Goyal are with the Department of Electronics and Communication, GLA University, Mathura 281406, India. E-mail: triptichoudhary06@gmail.com; vishal.goyal@gla.ac.in.
• Atul Bansal is with Chandigarh University, Mohali 140413, India. E-mail: atul.bansal@cumail.in.
* To whom correspondence should be addressed.
Manuscript received: 2022-05-31; revised: 2022-06-06; accepted: 2022-06-21

analyzed and converted into text or any other phonetic representation[5]. However, the language models work towards discovering the grammar, words, and sentence structure of any language. Multiple machine learning and HMM based techniques are used traditionally for ASR[6–10].

But with the advancement of deep learning models in the last decade, deep learning based solutions have replaced these traditional techniques[11]. Different deep networks like convolutional neural networks (CNN) and recurrent neural networks (RNN) are used for ASR[12]. In Ref. [13], an encoder decoder RNN is presented for ASR. The encoder comprises of multiple long short-term memory (LSTM) layers. These layers are pre-trained to identify phonemes, graphemes, and words. A residual 2D-CNN is also used for speech recognition in which the residual block comprises of connections amid previous and next layers[14].

In speech recognition the number of frames in an audio signal is much higher as compared to other forms. Therefore, the CNN model was modified, and the transformer network evolved. Transformers are widely used in various natural language processing (NLP) applications. One of the most successful areas is speech recognition. Transformers provide the ability of sequence-to-sequence translation[15]. Many researchers have used transformer in speech recognition and translation for different languages[16]. The speech recognition systems for Indian language are very few. There are a lot of use cases that require ASR for Indian language[17]. Thus, in this paper a transformer model for Hindi language speech recognition is proposed. The transformer model is augmented with wavelets for feature extraction. The wavelet can analyze the acoustic signal at multiscale. Thus, the features are extracted using discrete wavelet transform (DWT). These features are fed in the transformer model to generate the corresponding text. The transformer model is trained using Indian dataset for efficient speech-to-text translation.

The paper is organized in five sections. Introduction to the research problem and literature review is proposed in Section 1. Section 2 introduces the proposed methodology. Section 3 details the experiments. The results are discussed in Section 4. Section 5 gives the conclusion of the proposed work.

2 Proposed Methodology

The proposed methodology is shown in Fig. 1. The method hybrids the power of wavelet feature extraction with a transformer deep learning network for speech-to-text conversion. Wavelet transform can change the "scale" parameter to find different frequencies in the signal along with their location. So, now we know which frequencies exist in the time signal and where they exist. Smaller scale means that wavelets are squished. So, they can capture higher frequencies. On the other hand, a larger scale can capture lower frequencies. Therefore, the use of wavelets overcomes the problem of pitch and frequency of the audio signal. As shown in the block diagram, there are two main stages: wavelet feature extraction and the transformer network. These are discussed in the following sections.

2.1 Wavelet feature extraction

Wavelet transforms (WT) are extremely useful for the analysis of signals as they are able to perform multiscale analysis. More explicitly, dissimilar to the short-time Fourier trasform (STFT) that gives uniform time resolution to all frequencies, DWT gives high time resolution and low recurrence resolution for high frequencies, and high recurrence resolution and low time resolution for low frequencies. In that regard, it is like the human ear which displays comparable time-recurrence resolution qualities. DWT is a unique instance of WT that gives a minimal portrayal of a sign on schedule and recurrence that can be figured proficiently[18]. The discrete wavelet transform is used for audio signals.

$$T(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \times \psi \frac{(t-b)}{a} \mathrm{d}t \qquad (1)$$

where a is scale or dilation parameter, b is location of wavelet, ψ is wavelet function, and x is the signal.

The scale is used to represent the spread or squish of the wavelet.

2.2 Transformer network for speech recognition

Transformer networks are widely used for speech recognition tasks. A speech transformer is made up of two main parts, i.e. encoder and decoder. The task of encoder is to take a speech feature sequence $(x_1, x_2, \ldots, x_T)$ and transform it into a hidden representation $H = (h_1, h_2, \ldots, h_L)$. The decoder works in contrast to the encoder. It takes the input H and transforms it into the character sequence $(y_1, y_2, \ldots, y_S)$ of the corresponding text. The decoder considers the previous output when predicting the next character of the sequence. Conventionally, spectrogram inputs and word embeddings were used for speech-to-

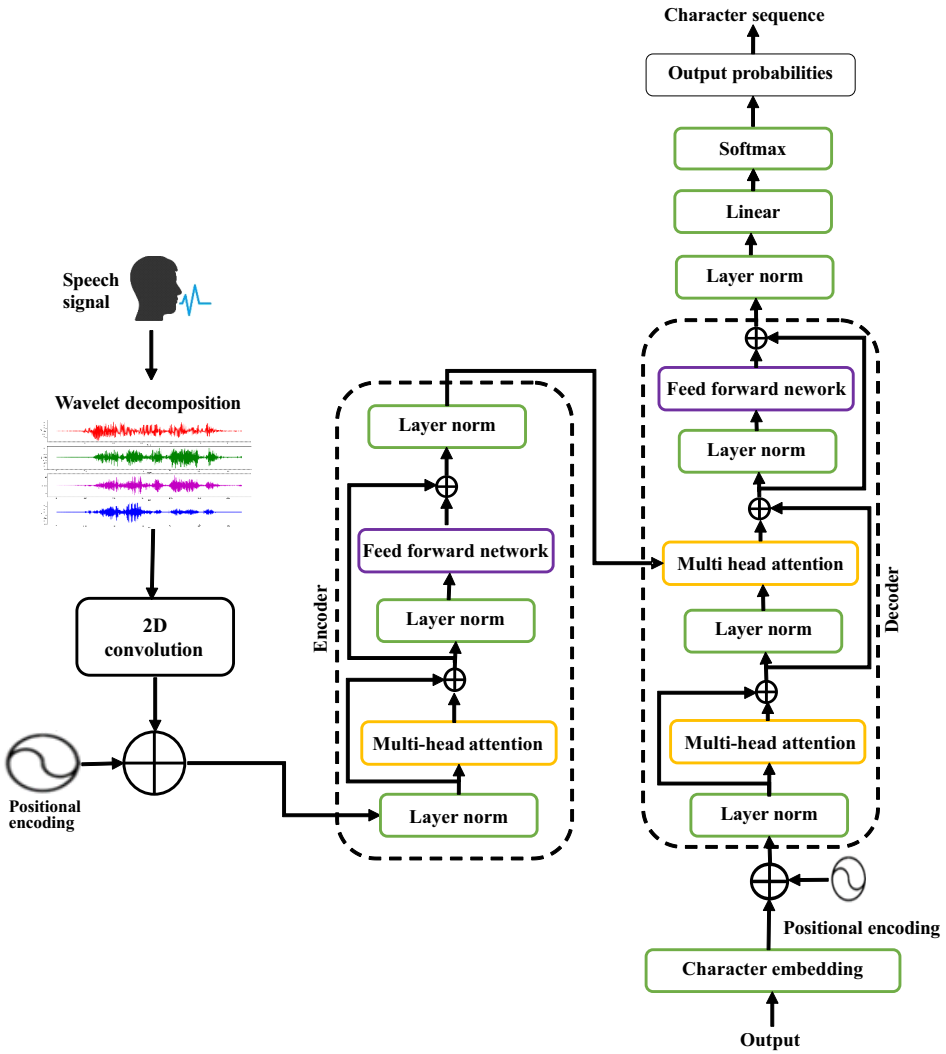

Fig. 1 Proposed methodology.

text conversion. But the transformer network replaces these by using the concept of multi-head attention and position wise feed forward networks.

The encoder and decoder comprise of N transformer layers. The encoder layers work continuously for refining the input sequence representation. These layers combine multi-head self-attention and frame-level affine transformations for the refining process. Self-attention refers to the process which communicates the different positions of input sequences to compute representations for the inputs.

The input for computing self-attention is a combination of three components: keys (K), values (V), and queries (Q). The attention value is computed using scaled dot product as shown in Eq. (1).

$$Attention\,(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \quad (2)$$

where $Q \in \mathbb{R}^{n_q \times d_q}$, $K \in \mathbb{R}^{n_k \times d_k}$ and $V \in \mathbb{R}^{n_v \times d_v}$ are the queries, keys, and values, where d denotes dimension and n denotes the sequence lengths, $d_q = d_k$ and $n_k = n_v$.

The output of a query is calculated by computing the weighted sum of the values. The weight of the query is calculated through the query function along with the related key. The multiple attentions are combined together using multi-head attention. It is calculated by taking the product of head number (h) and scaled dot-product *Attention*. The multi-head attention is computed using Eq. (3).

$$MultiHead\,(Q, K, V) = Concat(head_1, head_2, \ldots, head_h)W^o \quad (3)$$

where

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V).$$

The dimension of Q, K, and V is same as that of d_{model}, the projection matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_q}$,

$W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, $W^o \in \mathbb{R}^{hd_v \times d_{model}}$.

The decoder carries out multi-head attention in two rounds. Firstly, self-attention is computed based on the previous output sequence generated ($Q = K = V$). In the second round, attention of the output of the encoder final layer is computed. The output character sequence at each layer is predicted by making use of the previous layer.

The flowchart of the proposed techniques is shown in Fig. 2. The input feature sequence used here is wavelets. The wavelet features are fed into the transformer network. The transformer network comprises of 2D convolution layers along with normalization layer and ReLU activation function. Further 2D max pooling is done. This goes as an input sequence to the encoder followed by the decoder. The decoder generates the corresponding character sequence.

3 Experiment

The experiments are conducted using the speech dataset for Hindi Language. The length of the training data is 95.05 hours and testing data length is 5.55 hours (http://www.openslr.org/103/). The dataset comprises of unique sentences from Hindi stories. The data has high variability with a total of 78 different speakers. The sampling rate of the audio is 8 kHz with an encoding of 16 bit. The vocabulary size is 6542 including both training and testing datasets. Some sample speech text is as follows:

यह है मोटा राजा

मोटे राजा का है दुबला कुत्ता

मोटा राजा व दुबला कुत्ता घूमने निकले

वह उसके पीछे भागा

The model is developed using the Python Keras module. Thr transformer model is trained using GPU support in Google Colab. The optimization is done using the Adam Optimizer. The learning rate is initialized to 10 and the number of epochs used is 100. Each speech signal was decomposed using wavelet transform. A sample wavelet decomposition of the signal is shown in Fig. 3.

4 Result

The performance of the proposed system is computed using the word error rate (WER). It is a metric used for speech recognition or machine translation system. WER is computed as follows.

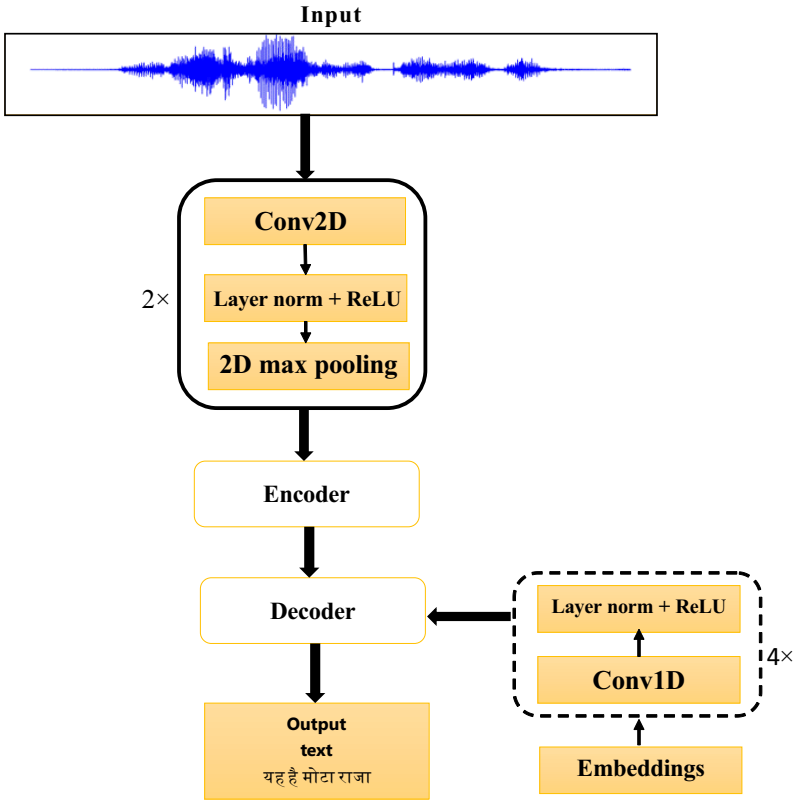

Fig. 2 Proposed model for ASR.

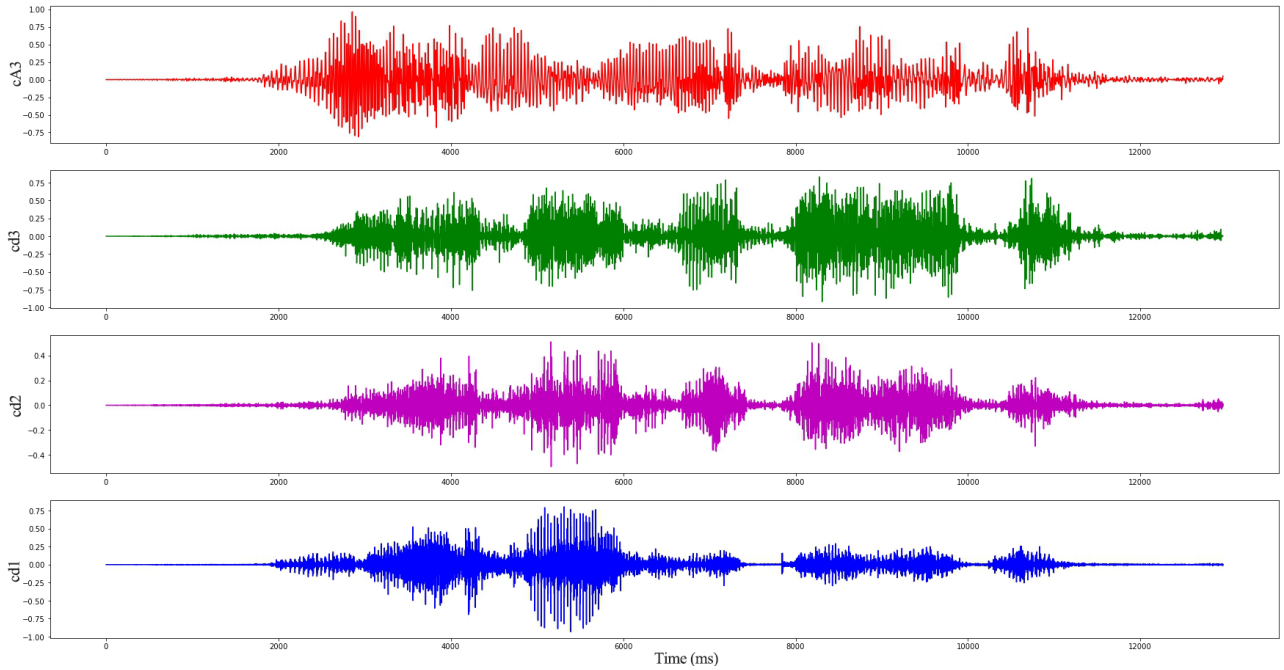

Fig. 3 Wavelet decomposition.

$$WER = (S + I + D)/N$$

where S denotes number of substitutions, D denotes deletions, and I denotes insertions.

• Substitutions. When the system transcribes one word in place of another. Transcribing the fifth word as "this" instead of "the" is an example of a substitution error.

• Deletions. When the system misses a word entirely. In the example, the system deleted the first word "well".

• Insertions. When the system adds a word into the transcript that the speaker did not say, such as "or" inserted at the end of the example.

The proposed method is compared with other state-of-the-art methods for Indian language. The WER of these systems is recorded and shown in the following Table 1.

Table 1 shows that the WER for WTSAR is less than 5% and thus it can be considered for practical uses. The WER of other methods is higher than the proposed method (see Fig. 4). The performance can be further improved by increasing the length of training data.

Table 1 WER for Indian language.

(%)

Method	Hindi	Marathi
CNN	6.3	6.1
RNN	5.9	6.2
Transformer	5.2	5.0
WTASR	4.8	4.9

5 Conclusion

Transformer networks are being widely used for ASR. In this paper a wavelet enabled transformer model for Indian language is proposed. The proposed model overcomes the variations of speech signals like variability in voice, gender, speed of utterance, etc. The wavelets can analyze a signal at multiscale. With the use of wavelets, the features are extracted from speech signals. These features consider the variability of speech signals. These features are used by the Transformer model to predict the corresponding character sequence. The limited data sources available for Indian language make it significant to develop an efficient model. The performance of the proposed model is compared with other state-of-the-art methods. The model gives a significant WER for Indian language and may be used for multiple ASR applications.

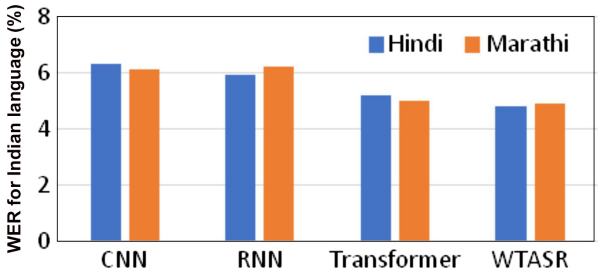

Fig. 4 Graph of WER for Indian languages.

References

[1] L. Deng, G. Hinton, and B. Kingsbury, New types of deep neural network learning for speech recognition and related applications: An overview, in *Proc. 2013 IEEE Int. Conf. on Acoustics*, *Speech and Signal Processing*, Vancouver, Canada, 2013, pp. 8599–8603.

[2] S. R. Shahamiri and S. S. B. Salim, A multi-views multi-learners approach towards dysarthric speech recognition using multi-nets artificial neural networks, *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 5, pp. 1053–1063, 2014.

[3] S. R. Shahamiri and S. S. B. Salim, Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach, *Adv. Eng. Inf.*, vol. 28, no. 1, pp. 102–110, 2014.

[4] H. Bourlard and N. Morgan, *Connectionist Speech Recognition*: *A Hybrid Approach*. Boston, MA, USA: Kluwer Academic Publishers, 1994.

[5] C. España-Bonet and J. A. R. Fonollosa, Automatic speech recognition with deep neural networks for impaired speech, in *Proc. 3rd Int. Conf. on Advances in Speech and Language Technologies for Iberian Languages*, Lisbon, Portugal, 2016, pp. 97–107.

[6] H. Sak, A. W. Senior, K. Rao, and F. Beaufays, Fast and accurate recurrent neural network acoustic models for speech recognition, in *Proc. 16th Annu. Conf. of the Int. Speech Communication Association*, Dresden, Germany, 2015, pp. 1468–1472.

[7] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, in *Proc. 2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Shanghai, China, 2016, pp. 4960–4964.

[8] C. C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, et al., State-of-the-art speech recognition with sequence-to-sequence models, in *Proc. 2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Calgary, Canada, 2018, pp. 4774–4778.

[9] T. Hori, J. Cho, and S. Watanabe, End-to-end speech recognition with word-based Rnn language models, in *Proc. 2018 IEEE Spoken Language Technology Workshop*, Athens, Greece, 2018, pp. 389–396.

[10] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, Convolutional neural networks for speech recognition, *IEEE/ACM Trans. Audio Speech Lang Process.*, vol. 22, no. 10, pp. 1533–1545, 2014.

[11] B. Vachhani, C. Bhat, B. Das, and S. K. Kopparapu, Deep autoencoder based speech features for improved dysarthric speech recognition, in *Proc. 18th Annu. Conf. of the Int. Speech Communication Association*, Stockholm, Sweden, 2017, pp. 1854–1858.

[12] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss, in *Proc. 2020 IEEE Int. Conf. on Acoustics*, *Speech and Signal Processing*, Barcelona, Spain, 2020, pp. 7829–7833.

[13] K. Rao, H. Sak, and R. Prabhavalkar, Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer, in *Proc. 2017 IEEE Automatic Speech Recognition and Understanding Workshop*, Okinawa, Japan, 2017, pp. 193–199.

[14] Y. Wang, X. Deng, S. Pu, and Z. Huang, Residual convolutional CTC networks for automatic speech recognition, arXiv preprint arXiv: 1702.07793, 2017.

[15] L. Dong, S. Xu, and B. Xu, Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition, in *Proc. 2018 IEEE Int. Conf. on Acoustics*, *Speech and Signal Processing*, Calgary, Canada, 2018, 5884–5888.

[16] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li, Developing real-time streaming transformer transducer for speech recognition on large-scale dataset, in *Proc. 2021 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Toronto, Canada, 2021, 5904–5908.

[17] A. Singh, V. Kadyan, M. Kumar, and N. Bassan, ASRoIL: A comprehensive survey for automatic speech recognition of Indian languages, *Artif. Intell. Rev.*, vol. 53, no. 5, pp. 3673–3704, 2020.

[18] S. Jaglan, S. Dhull, and K. K. Singh, Tertiary wavelet model based automatic epilepsy classification system, *Int. J. Intell. Unmanned. Syst.*, doi: 10.1108/ijius-10-2021-0115.

Vishal Goyal is currently a professor at the Department of Electronics & Communication, GLA University, Mathura, India. He obtained the PhD in 2016 from GLA University, Mathura, India. He has participated in several high-profile conferences and published many high-quality journal papers. In addition to his academic career, he held several managerial positions in GLA University.

Tripti Choudhary is a research scholar at the Department of Electronics & Communication, GLA University, Mathura, India. She obtained the MS degree in 2012. Currently, she is working on low-resource Indian languages and published several good-quality papers on the Automatic Speech recognition domain.

Atul Bansal is currently a professor at Chandigarh University, Punjab, India. Prior to his recent appointment at Chandigarh University, he was a professor in the Department of Electronics & Communication, GLA University, Mathura, India. He received the BS degree from Punjab Technical University, India in 2001. He received the MS degree from Indian Institute of Technology, Delhi, India, and the PhD degree from Thapar Institute of Engineering & Technology, Punjab, India in 2015. He published several papers in preferred journals and chapters in books and participated in a range of forums on the electronics domain. He also presented various academic as well as research-based papers at several national and international conferences.