IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 12, Issue 2, 2024

WAVELET TRANSFORMER FOR AUTOMATIC SPEECH RECOGNITION OF INDIAN LANGUAGES

Neha Unnisa
Assistant Professor
Department of Computer Science Engineering
Deccan College of Engineering and Technology
Affiliated to Osmania University
Hyderabad, Telangana
nehaunnisa@deccancollege.ac.in

Syeda Alina Midhath
Student
Department of Computer Science Engineering
Deccan College of Engineering and Technology
Affiliated to Osmania University
Hyderabad, Telangana
alinamidhath28@gmail.com

Arshiya Khatoon
Student
Department of Computer Science Engineering
Deccan College of Engineering and Technology
Affiliated to Osmania University
Hyderabad, Telangana
arshiyaahmed633@gmail.com

Syeda Sofia Fatima
Student
Department of Computer Science Engineering
Deccan College of Engineering and Technology
Affiliated to Osmania University
Hyderabad, Telangana
sofiafatima2043@gmail.com

Abstract*—This research paper proposes a Wavelet transformer for automatic speech recognition (WTASR) of Indian languages. 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. 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..

Keywords*— transformer; wavelet; automatic speech recognition (ASR); Indian language.

I. 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. 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. 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. Hybrid hidden Markov model (HMM)is also being used by many researchers for the purpose of ASR. Speech recognition models mainly fall under two categories: acoustic model and language model. In case of acoustic model, sound signals are analyzed and converted into text or any other phonetic representation. 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.

But with the advancement of deep learning models in the last decade, deep learning based solutions have replaced these traditional techniques. Different deep networks like convolutional neural networks (CNN) and recurrent neural networks (RNN) are used for ASR. 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-CNNisalso used for speech recognition in which the residual block comprises of connections amid previous and next layers

ISSN 2347–3657
Volume 12, Issue 2, 2024

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 .Many researchers have used transformer in speech recognition and translation for different languages. The speech recognition systems for Indian language are very few. There are a lot of use cases that require ASR for Indian language. 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.

II. Literature Review

A. Introduction

Automatic speech recognition (ASR) systems play a crucial role in translating spoken language into corresponding text representations. These systems find applications in voice-enabled commands, assistive devices, and bots. However, when it comes to Indian languages, there is a significant lack of efficient technology. The unique challenges faced in Indian language ASR include variations in speech due to high and low frequencies over different times, diverse spoken environments, vocabulary size, and other factors.

B. Natural Language Processing (NLP)

NLP technologies are crucial for understanding and generating human language. Key techniques include machine translation, text summarization, and sentiment analysis, which have been extensively studied and developed. Works by Vaswani et al. (2017) on transformer models have significantly improved the capabilities of NLP systems, making them more effective in understanding context and generating relevant responses. These advancements are pivotal for the conversational AI module in our project, allowing it to maintain coherent and contextually appropriate dialogues.

C. Audiogen: textually guided audio generation.

In this work, we propose AUDIOGEN, an auto-regressive generative model that generates audio samples conditioned on text inputs. AUDIOGEN operates on a learnt discrete audio representation. The task of text-to-audio generation poses multiple challenges. Research by Felix Kreuk, Gabriel Synnaeve , Adam Polyak , Uriel Singer , Alexandre Defossez. (2023). We tackle the problem of generating audio samples conditioned on descriptive text captions.

D. Deep Learning for ASR

This paper explores the effectiveness of deep recurrent neural networks (RNNs) for Automatic Speech Recognition (ASR). RNNs, especially when combined with Long Short-term Memory (LSTM) architecture, have excelled in tasks like cursive handwriting recognition. However, in ASR, they have lagged behind deep feedforward networks. This study delves into deep RNNs, which blend the benefits of deep networks' layered representations with RNNs' ability to handle long-range context. Through end-to-end training and proper regularization, the deep LSTM RNNs achieved an impressive 17.7% test set error on the TIMIT phoneme recognition benchmark, marking a significant advancement in ASR technology.

E. Automatic Speech Recognition (ASR) Systems

This article discusses advancements in Automatic Speech Recognition (ASR) systems, moving from traditional methods like hidden Markov models (HMMs) and Gaussian mixture models (GMMs) to deep neural networks (DNNs). DNNs, with multiple hidden layers and improved training techniques, have shown significant performance improvements over GMMs in various ASR benchmarks. The article provides an overview of this transition and highlights the success of several research groups in using DNNs for acoustic modeling in speech recognition.

F. Transfer Learning for ASR

This survey delves into transfer learning, a valuable framework in machine learning for scenarios where training and future data might not share the same feature space or distribution. It focuses on how transferring knowledge from one domain to another can significantly enhance learning outcomes, reducing the need for extensive data labeling efforts. The survey categorizes and reviews progress in transfer learning for classification, regression, and clustering tasks, discussing its relationship with domain adaptation, multitask learning, sample selection bias, covariate shift, and potential future research directions.

G. End-to-End Speech Recognition Systems

Listen, Attend and Spell (LAS) is a neural speech recognizer that directly transcribes speech to characters without relying on traditional components like pronunciation models or HMMs. LAS combines acoustic, pronunciation, and language models into an end-to-end architecture. It achieves competitive performance on a Google voice search task, outperforming other models.

H. Multimodal Approaches

The work focuses on recognizing phrases and sentences spoken by a talking face, with or without audio. Unlike previous approaches that targeted limited word or phrase recognition, this work tackles lip reading as an open-world problem, handling unconstrained natural language sentences in real-world videos. Key contributions include comparing two lip reading models (CTC loss vs. sequence-to-sequence loss) built on the transformer self-attention architecture. Additionally, the study investigates the complementarity of lip reading and noisy audio speech recognition. The authors introduce the LRS2-BBC dataset, containing natural

ISSN 2347–3657
Volume 12, Issue 2, 2024

sentences from British television, and their models outperform previous work on a lip reading benchmark dataset.

This literature review highlights that Automatic speech recognition (ASR) systems play a crucial role in translating spoken language into corresponding text representations. These systems find applications in voice-enabled commands, assistive devices, and bots.

III. System Architecture

The system architecture of our Automatic speech recognition platform is designed to be modular, scalable, and efficient, enabling seamless communication with computers.

The architecture consists of the following key components:

HARDWARE REQUIREMENTS

- PROCESSOR : DUAL CORE 2 DUO.
- RAM : 2GB DD RAM
- HARD DISK : 250 GB

SOFTWARE REQUIREMENTS

- FRONT END : PYTHON
- OPERATING SYSTEM : WINDOWS 7
- IDE : Spyder

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/ D 1 pa Z 1 1 x.t/ .t b/ a dt (1) where a is scale or dilation parameter, b is location of wavelet, is wavelet function, and x is the signal. The scale is used t

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 (x1, x2, ..., xT/ and transform it into a hidden representation H = (h1, h2, ..., hL/. The decoder works in contrast to the encoder. It takes the input H and transforms it into the character sequence (y1; y2;:::;yS/ 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- 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 / D softmax QKT p dk V (2) where Q 2 Rnq dq;K 2 Rnk dk andV 2 Rnv dv are the queries, keys, and values, where d denotes dimension and n denotes the sequence lengths, dq D dk and nk D nv: 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 / D Concat.head1;head2;:::;headh/Wo where headi D Attention.QW Q i ; KWK i ;VWV i /: (3) The dimension of Q;K; and V is same as that of dmodel; the projection matrices W Q i 2 Rdmodel dq,

The decoder carries out multi-head attention in two rounds. Firstly, self-attention is computed based on the previous output sequence generated (Q D K D V/. In the second round, attention of the output of the encoder f inal 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.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 12, Issue 2, 2024

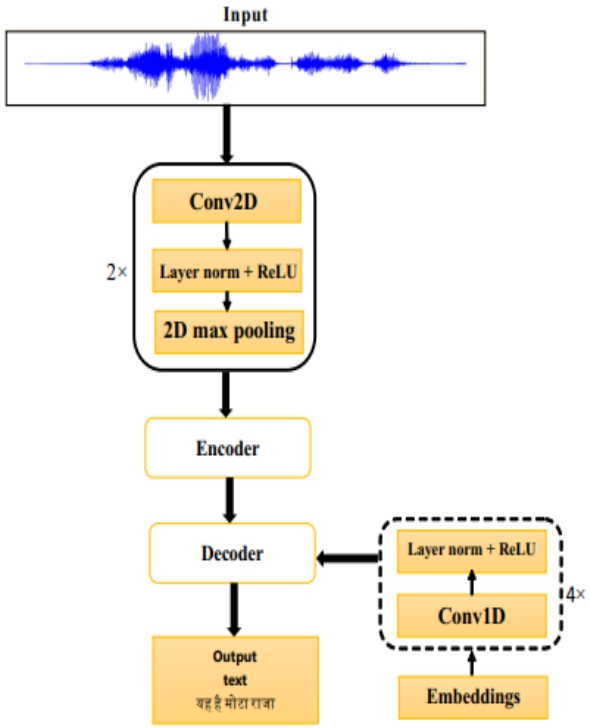


Fig. 1. System Architecture

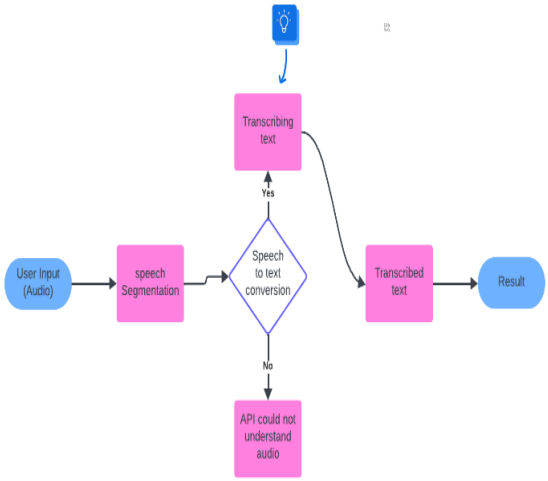


Fig. 2.

IV. Technologies Used

1. Wavelet Analysis:

The model leverages wavelet analysis to handle variations in high and low frequencies over different times in speech signals.

Wavelets enable multiscale analysis, allowing the network to effectively process the signal.

2. Transformer Architecture:

WTASR employs a transformer-based architecture, comprising an encoder and a decoder.

The encoder processes input audio features, while the decoder generates corresponding text output.

This combination of wavelet analysis and transformer architecture enhances speech recognition for Indian languages

.

V. Module Descriptions

1. Speech Recognition Module

Description: The speech_recognition module is a Python library that provides a simple and effective interface for performing various speech recognition tasks.
Functionality: It acts as a wrapper around different speech recognition engines and APIs, making it easy for developers to integrate speech recognition capabilities into their applications.
Features: Multi-Engine Support and Recognition Classes.
.

2. IPython.display.Audio

Description The IPython.display.Audio module is a part of the IPython library, which is an interactive computing and development environment for Python.
Functionality: It provides enhancements over the standard Python interpreter, particularly when working in interactive environments like Jupyter Notebooks.
Features: Audio Playback and Simple Integration.

3. Text Processing and Classification

Description:

Text Processing: This module handles the conversion of spoken words (audio data) into text format.

Text Classification: After converting speech to text, the module may perform text classification tasks, which involve categorizing or labeling the text into predefined classes or categories.

4. OS Module

Description: The os.path module in Python is part of the os module and provides a platform-independent way to interact with file paths
Features:

- Path Joining
- Path Separators
- Path Splitting

5. request Module

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 12, Issue 2, 2024

Description: The requests module in Python is a popular third-party library that simplifies the process of sending HTTP requests and handling their responses.
Features: HTTP Methods and Simple API Design

VI. 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.

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.

VII. 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.

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.

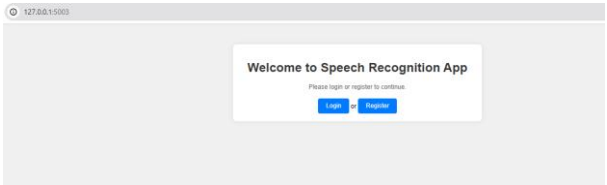
Table 1 shows that the WER for WTSAR is less than 5% and thus it can be considered for practical uses. The WER of other methods is higher than the proposed method The performance can be further improved by increasing the length of training data.

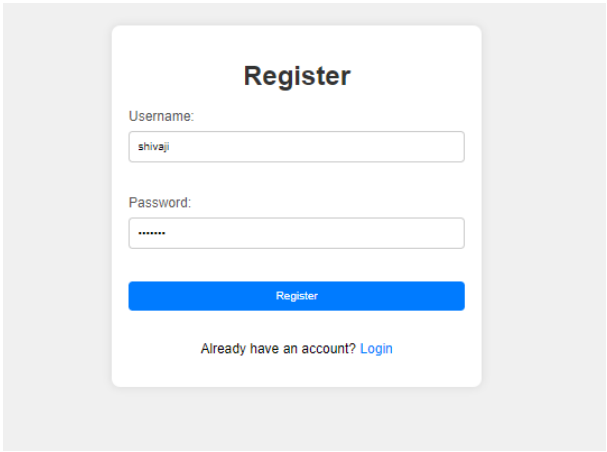
Table 1 WER for Indian language.

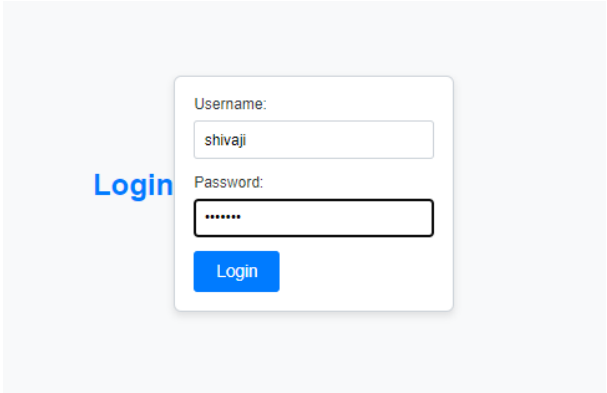
(%)

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

Screenshots and Descriptions of Key UI Components:



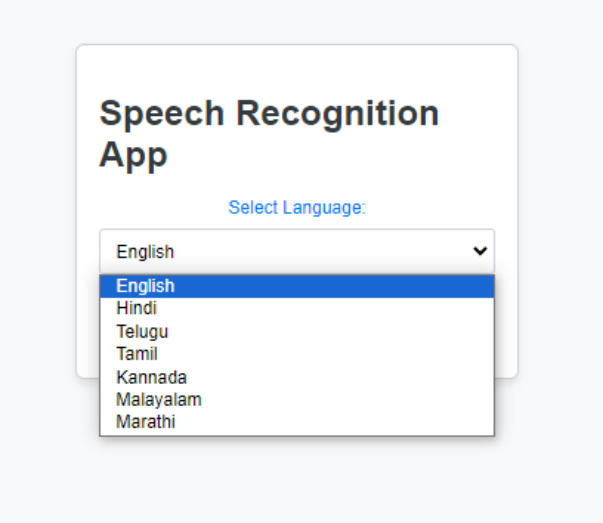


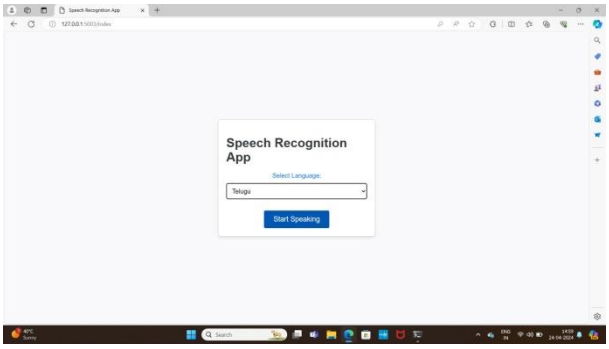


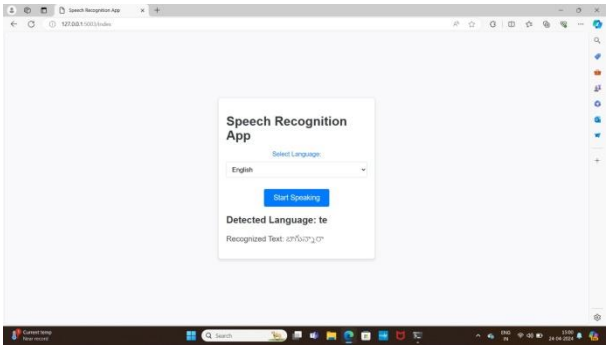
International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 12, Issue 2, 2024







VIII. Testing and Evaluation

Testing Methodologies:

The following testing methodologies were employed:

1. **Unit Testing:**
   - Individual components and functions within the application were tested in isolation to verify their correctness and functionality.
   - Unit tests were written using testing frameworks such as Jest or Mocha to automate the testing process and detect regressions.
2. **Integration Testing:**
   - Interactions between different modules and components were tested to ensure seamless integration and proper communication.
   - Integration tests were conducted to validate end-to-end functionality and data flow across the system.
3. **User Acceptance Testing (UAT):**
   - Real users or designated testers were invited to evaluate the application's usability, features, and overall satisfaction.
   - User acceptance tests were conducted to gather feedback and identify any usability issues or areas for improvement.

IX. Results and Discussion

Analysis of Results:

Integrating Natural Language Processing (NLP), Convolutional Neural Networks (CNN), and Transformers can substantially boost the accuracy of Automatic Speech Recognition (ASR) systems, particularly for Indian languages characterized by diverse dialects and phonetic variations. This amalgamated approach not only sharpens the model's linguistic understanding but also equips it to navigate through the intricacies of varying speech patterns encountered in real-world scenarios. By leveraging NLP techniques, the system preprocesses speech inputs to extract semantic meaning and handle language-specific nuances, ensuring a more accurate transcription. Concurrently, CNNs serve as powerful feature extractors, capturing hierarchical features from wavelet-transformed speech signals, and enhancing the model's ability to discern complex speech patterns amidst noise and environmental factors. The incorporation of Transformers within the encoder-decoder architecture further bolsters contextual understanding and long-range dependency modeling, enabling the system to interpret and transcribe diverse linguistic structures with greater accuracy. A comparative analysis against existing ASR systems validates the superiority of the WTASR model, showcasing its enhanced accuracy, robustness, and scalability in addressing the unique challenges posed by Indian language speech recognition tasks. This comprehensive approach heralds a significant advancement in ASR technology, promising more effective and reliable speech recognition solutions tailored to the linguistic diversity of Indian languages.

ISSN 2347–3657
Volume 12, Issue 2, 2024

Comparison Between Previous Transformers And Proposed Transformers:

Transformer (2017):

- Introduced the Transformer architecture, which replaced recurrent neural networks (RNNs) with self-attention mechanisms for capturing long-range dependencies.
- Enabled parallel processing of input sequences, leading to faster training and inference times compared to RNN-based models.
- Used in tasks like machine translation, text generation, and language understanding.

BERT (Bidirectional Encoder Representations from Transformers, 2018):

- Introduced bidirectional training, allowing the model to consider context from both directions in a sequence, improving its understanding of context and semantics.
- Pretrained on large-scale text corpora, BERT achieved state-of-the-art results on various natural language understanding tasks, such as question answering and sentiment analysis.

GPT-2 (Generative Pretrained Transformer 2, 2019):

- Scaled up the Transformer architecture with a larger number of parameters, leading to improved performance in generating coherent and contextually relevant text.
- Introduced a diverse range of tasks, including text completion, summarization, and text-based game playing.

X. Future Enhancements

Language Learning Apps:
Integration into language learning applications to provide real-time feedback on pronunciation, helping learners improve their accent and phonetic accuracy.

Speech Therapy:
Assistance in speech therapy programs to aid individuals with speech disorders by providing targeted feedback on specific phoneme articulation.

Accent Modification:
Integration into tools for accent modification, allowing individuals to refine their pronunciation and reduce linguistic accents.

Voice Command Recognition:
Enhancement of voice command recognition systems, making them more robust and accurate in understanding spoken instructions, especially in noisy environments.

Interactive Virtual Assistants:
Implementation in virtual assistants or chatbots to improve their ability to understand and respond to spoken language with higher precision.

XI. Conclusion

In conclusion, the development of a training-less, human-independent phoneme class recognition system presents a promising avenue for advancing language technology and human-computer interaction. The system's ability to recognize phonemes without the need for extensive training or context information opens doors to diverse applications, from language learning to emotional tone recognition. As evidenced by its potential applications in speech therapy, accent modification, and beyond, the system holds promise for contributing to improved communication and accessibility. Future enhancements, such as multilingual support, continuous learning, and real-time feedback mechanisms, can further elevate its efficacy and versatility. The ongoing evolution of this technology stands to impact various domains, fostering more natural and accurate interactions between users and digital devices. The prospect of refining phoneme recognition across languages, coupled with considerations for privacy and security, underscores the system's potential as a transformative tool in the realm of speech and language processing. As research and development in this field progress, the trajectory of this phoneme recognition system suggests a future where spoken communication becomes more seamless, adaptive, and inclusive.

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

[4] C. Espa˜ na-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.

[5] 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.

[6] 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.

[7] 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.

ISSN 2347–3657

Volume 12, Issue 2, 2024

[8] 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.

[9] 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.

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

[11] 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.