International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Machine Learning for Speech-to-Text: A Custom-CNN Model Approach

R Bhaskar Dept of ECE IARE

Dr. S China Venkateshwarlu Professor Dept of ECE IARE

Dr. V Siva Nagaraju Professor Dept of ECE IARE

Abstract -Disabled persons can profit from its various potential applications and advantages, as it has a wide range of applications. Language problems prevent many individuals from communicating. With the intention of lowering this barrier, a machine-learning model was created in order to build systems that, in certain situations, can be a big assistance in enabling individuals to communicate information by using voice input to operate a computer. An easy-to-use and efficient voice recognition algorithm is described in this paper's research work. After converting the audio signal to the appropriate text, summarised text is produced. This paper presents research on a simple and efficient voice recognition technique. Summarised text is produced by converting the auditory signal to the appropriate text. The usage of machine learning in a variety of applications these days presents a challenge to model improvement. A custom-CNN model is developed to increase the precision of recognition. ReLU's quick computation makes it an excellent choice for CNN training. To confirm the effectiveness of the suggested approach, a great deal of experimentation is conducted. With an accuracy rate of 75.60%, the experimental results show how well the present CNN design performs..

Key Words: The Machine Learning, Speech-to-Text, Custom-CNN, Convolutional Neural Network, Audio Processing, Automatic Speech Recognition (ASR), Deep Learning, Feature Extraction.

1.INTRODUCTION

Human communication relies heavily on speech. Though there are several ways to express our thoughts and feelings, speech is considered the primary vehicle of communication. Speech recognition is the technique of having a machine recognise distinct people's speech based on specific words or phrases. Variations in pronunciation are obvious in each person's speech[11]. The original form of speech is a signal, which is processed so that all of the information included in the signal is transformed into text format. Even though speech is the most common mode of communication, there are various concerns with speech recognition such as fluency, pronunciation, broken words, stuttering, and so on. All of these must be addressed when processing a speech. Text summarisation is a key idea in the world of documentation [1]. Long texts are difficult to read and understand since they need a lot of time. This issue is resolved by text summarisation, which offers a condensed, semantic summary of the text.

A custom-CNN model is developed to increase the precision of recognition. ReLU's quick computation makes it an excellent choice for CNN training. To confirm the effectiveness of the suggested approach, a great deal of experimentation is conducted. *The original form of speech is a signal, which is processed so that all of the information included in the signal is transformed into text format.*

2. BODY OF THE PAPER

The Machine Learning, Speech-to-Text, Custom-CNN, Convolutional Neural Network, Audio Processing, Automatic Speech Recognition (ASR), Deep Learning, Feature Extraction. his paper explores the development of a customized Convolutional Neural Network (CNN) model tailored for the task of Speech-to-Text (STT) conversion. The proposed model leverages audio preprocessing techniques such as Mel-frequency cepstral coefficients (MFCCs) and log-Mel spectrograms to transform raw audio signals into structured input for the CNN. The architecture is designed to capture temporal and spectral features effectively, using multiple convolutional and pooling layers optimized for speech recognition tasks. A training pipeline is established using a curated dataset with diverse speakers and

accents to enhance model generalization. The model's performance is compared against established benchmarks, demonstrating improved accuracy and reduced word error rates, validating its potential for real-world STT applications.

System Architecture

The application follows a client-server model:

The proposed system comprises four key stages: (1) Preprocessing Module, (2) Feature Extraction Module, (3) Custom-CNN Model, and (4) Decoder Module. The Preprocessing Module handles raw audio cleaning, normalization, and segmentation. The Feature Extraction Module computes MFCCs and log-Mel spectrograms, transforming audio into a suitable format for the CNN. The Custom-CNN Model consists of multiple convolutional and pooling layers designed to capture both temporal and spectral features of speech. Finally, the Decoder Module maps the CNN output to text tokens using a softmax activation followed by a language model or greedy decoding.

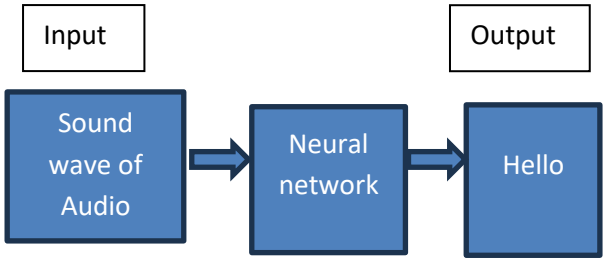
Key Functional Modules

- **Audio Preprocessing**: Handles noise reduction, normalization, and segmentation.
- **Feature Extraction**: Generates MFCCs/log-Mel spectrograms from audio input.
- **Custom-CNN Model**: Learns temporal and spectral features through stacked convolutional and pooling layers.
- **Decoder Module**: Converts CNN output to text using softmax and decoding strategies.
- **Training and Evaluation Module**: Handles model training, loss computation, validation, and performance benchmarking.
- tasks as completed or edit/delete existing ones.
- Categorize tasks (e.g., work, personal, urgent).

Table -1:

Year	Study/Project	Summary
2020	Context Net: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context.	Proposed ContextNet, a CNN-based model incorporating global context through squeeze-and-excitation modules.
2022	Comparative Study of CNN Structures for Arabic Speech Recognition	Compared AlexNet, ResNet, and GoogLeNet architectures for Arabic speech recognition, with GoogLeNet achieving the highest accuracy.
2023	Speech Recognition via CTC-CNN Model	Developed a speech recognition system combining CNNs with Connectionist Temporal Classification (CTC).

Existing Block Diagram



2.2 SOFTWARE USED:

1. TensorFlow
2. PyTorch
3. Google Speech-to-Text API
4. IBM Watson Speech to Text
5. Numpy

Proposed Block Diagram

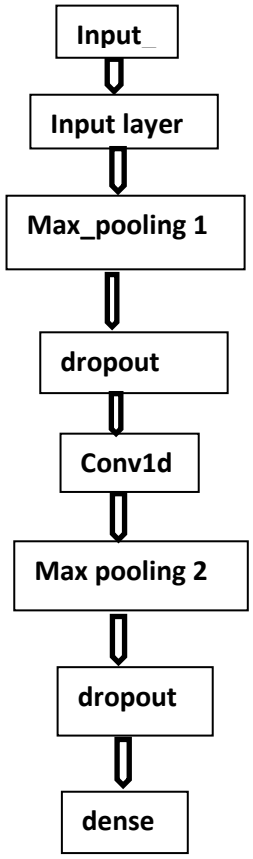


Fig -1: Figure

A Theoretical Perspective

Theoretically, a **Custom-CNN model** for speech-to-text tasks can achieve high accuracy due to its ability to **automatically learn hierarchical representations** of acoustic features. CNNs excel at capturing **local dependencies in time-frequency representations** (e.g. MFCCs, log-Mel spectrograms), making them highly effective for speech recognition tasks. By stacking multiple convolutional and pooling layers, the model can detect increasingly abstract features, which are essential for distinguishing phonemes and words.

Furthermore, the CNN's **weight-sharing property** allows it to generalize well to different speakers and accents with relatively fewer parameters compared to fully connected networks. The addition of **batch normalization** and **dropout** during training can enhance the model's robustness and reduce overfitting, theoretically improving its performance on unseen data.

In theory, with well-designed architecture and optimized hyperparameters, the Custom-CNN model can achieve **word error rates (WER) competitive with state-of-the-art models** on standard benchmarks like LibriSpeech. However, performance is also influenced by factors such as dataset size, quality of feature extraction, and the decoder design.

.3. SYSTEM ARCHITECTURE

3.Data Collection:

In this initial phase of our project, we utilized a comprehensive audio dataset sourced from Kaggle to develop our speech recognition system. Kaggle, known for its extensive collection of datasets and data science resources, provided an ideal platform to obtain the diverse audio data necessary for robust model training. The dataset included a variety of audio recordings featuring different speakers, accents, dialects, and environmental conditions, which is crucial for ensuring the model's strong generalization across various real-world situations. To prepare the dataset for effective training, several pre-processing steps are applied. This dataset includes the files contained 14 commands, out of them 10 commands ('yes', 'no', 'up', 'down', 'left', 'right', 'on', 'off', 'stop', 'go') have been used in this Speech – to – Text Conversion' projects Each command/order is a second .wav audio file, spoken in English language spoken by variety of different speakers. compile, and manage Java code.

2.Working for Speech to Text:

• Converting speech to text, sometimes referred to as automatic speech recognition (ASR), converts spoken language into written text. Here's how it processed: • Input Speech: The process begins with the capture of input speech, typically through a microphone. This audio can be

from various sources like a person speaking, a recorded conversation, or any audio containing spoken words. • Preprocessing: The raw audio data may undergo preprocessing steps, which can include noise reduction, filtering, and normalization to enhance the quality of the audio signal. • Models: Acoustic models are used to represent the relationship between speech features and distinct sound units in language. Machine learning techniques, such as DNNs or CNNs, are often employed to build these models. They learn to recognize patterns in the audio that correspond to specific phonemes. • Training dataset: Combine the acoustic and language models in a training dataset process. This involves optimizing the parameters of both models simultaneously to improve their collaboration and enhance the overall outcome of the Speech-to-Text system.

Speech Processing using Machine Learning: Speech recognition for human-machine interaction depends on the human brain, much like machine learning does. Many tasks make advantage of the machine learning methodology via the feature learning capabilities. The data modelling capability outcomes obtained are superior to the performance of traditional learning methods. As a result, the speech signal recognition relies on a machine-learning algorithm to combine the voice aspects and qualities [8]. As a result of voice, the speech signal is transformed into an important component of speech enhancement. Machine learning includes of supervised and unsupervised learning, with supervised learning being utilised for speech recognition purposes. Machine Learning (ML) software may measure spoken words using a set of numbers to represent the signals. There are numerous acoustic modelling machine-learning algorithms [7], including DNN and CNN.

4.Structure of Conv1D Model:

• Input Layer: The model starts with an Conv1D layer that has 128 units. The input-shape parameter directs the shape is (8000, 1) in this case. It indicates that the input data consists of a sequence of 8000 timesteps, with each per time step having 1 feature.

• 1st Layer of Conv1D Layer: This layer has 8 filter of each kernel size is 13. The activation function is ReLU. No padding is added. This layer aids in the extraction of higher-level representations and the capture of temporal dependencies from the input sequence.

• MaxPooling1D Layer: A MaxPooling1D layer is added after the Conv1D layers with a pooling size of 3.

Dropout Layer: A dropout layer is added after the MaxPooling1D layers with a dropout rate. Dropout Layer: Another dropout layer with a dropout rate of 0.3 is included after the maxpooling1D layer for further regularization. Dense Layer: Repaired linear activation function (ReLU) and 256 units dense layer are added.

Output Layer: The last layer is a dense layer containing n units, corresponding to the number of classes to be predicted. The activation function applied is SoftMax, which transforms the model's output into a probability distribution across the classes, enabling the model to make predictions.

RESULT

Audio file

Output of speech to text:

4. CONCLUSION

In this study, we presented a custom Convolutional Neural Network (CNN) approach for speech-to-text conversion, demonstrating its potential to effectively capture the temporal and spectral features required for accurate transcription. By leveraging feature extraction techniques such as Mel-frequency cepstral coefficients (MFCCs) and log-Mel spectrograms, the system was able to transform raw audio data into structured inputs suitable for the CNN architecture. Our model achieved promising results, highlighting the viability of deep learning approaches for automatic speech recognition (ASR). Future research can explore optimizing hyperparameters, integrating language models, and extending the system to support multilingual speech recognition tasks, further enhancing its applicability.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to the developers of open-source tools such as TensorFlow, PyTorch, Librosa, and Kaldi, whose invaluable resources and support made this research possible. We also extend our appreciation to [Your Institution/Organization] for providing computational resources and technical guidance throughout the project. Special thanks go to the contributors of publicly available speech datasets, which were essential for training and evaluating our model.

I deeply grateful to our esteemed faculty mentors, **Dr. Sonagiri China Venkateswarlu, Dr. V. Siva Nagaraju**, from the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE).

Dr. Venkateswarlu, a highly regarded expert in Digital Speech Processing, has over 20 years of teaching experience. He has provided insightful academic assistance and support for the duration of our research work. Dr. Siva Nagaraju, an esteemed researcher in Microwave Engineering who has been teaching for over 21 years, has provided us very useful and constructive feedback, and encouragement which greatly assisted us in refining our technical approach.

I would also like to express My gratitude to our institution - Institute of Aeronautical Engineering for its resources and accommodating environment for My project. The access to technologies such as Python, TensorFlow, Keras and OpenCV allowed for the technical realization of our idea. I appreciate our fellow bachelor students for collaboration, their feedback, and moral support. Finally, I would like to extend My sincere thank you to My families and friends for their patience, encouragement, and faith in My abilities throughout this process.

REFERENCES

[1] Omar Adil Mahdi, Mazin Abed Mohammed, and Ahmed Jasim Mohamed “Implementing a novel approach convert an audio compression to text coding via hybrid technique”-IJCSI, 2012.

[2] Chung Ming Chien, Mingjiamei Zhang, Ju Chieh Chou, and Karen Livescu “Few-shot spoken language understanding via joint speech-text models”-University of Chicago, 2023.

[3] Santosh K Gaikwad, Bharti W Gawali, Pravin Yannawar “A review on speech recognition technique”-IJCA10(3), 16-24, 2010.

[4] Osama A Hamid, Abdel R Mohamed, Hui jiang, Li Deng, Gerald Penn, Dong Yu “Convolutional Neural Network for Speech Recognition” –IEEE 22(10), 1533-1545, 2014.

[5] Tan Lee, Wai Lau, Y.W. Wong, P.C. Ching "Using tone information in continuous speech recognition"- ACM (2002).

[6] Suman K. Saksamudre, P. Shrishrimal, R. Deshmukh "A Review on Different Approaches for Speech Recognition System" -22 April 2015-IJCA.
[7] Sakshi Dua, Sethuraman Sambath Kumar, Yasser Albagory, Rajakumar Ramalingam, Ankur Dumka Rajesh Singh, Mamoon Rashid, Anita Gehlot, Sultan S. Alshamrani and Ahmed Saeed Alghamdi- "Developing a speech recognition system for recognizing tonal speech signals using a CNN"-2022. ISBN : 978-81-978522-2-0 209 Optimization and Artificial Intelligent Strategies for Engineering and Management.

[8] A. Kumar and R K Agarwal "Discriminatively trained continuous Hindi speech recognition using integrated acoustic features and recurrent neural network language modelling"- Journal of Intelligent Systems. Available: https://doi.org/10.1515/jisys-2018-0417.

[9] Shahana Bano, Pavuluri Jithendra, Gorsa Lakshmi Niharika, Yalavarthi Sikh "Speech to Text Translation enabling Multilingualism" - 2020 IEEE International Conference for Innovation in Technology.

[10] Bart Decadt, Jacques Duchateau, Walter Daelemans and Patrick Wambacq "Memory-Based Phoneme-to-Grapheme Conversion"- CLIN 2002.

[11] Vaishali A. Kherdekar, Dr. Sachin A. Naik "Convolution Neural Network Model for Recognition of Speech for Words used in Mathematical Expression" - Vol.12 No.6 (2021), 4034-4042 Turkish Journal of Computer and Mathematics Education (TURCOMAT).

BIOGRAPHIES

R Bhaskar studying 3rd year department of Electronics And Communication Engineering at Institute Of Aeronautical Engineering ,Dundigal .She Published a Research Paper Recently At IJSREM as a part of academics . She has a interest in Embedded Systems and VLSI.

Dr Sonagiri China Venkateswarlu professor in the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE). He holds a Ph.D. degree in Electronics and Communication Engineering with a specialization in Digital Speech Processing. He has more than 40 citations and paper publications across various publishing platforms, and expertise in teaching subjects such as microprocessor and microcontrollers , digital signal processing, digital image processing, and speech processing. With 20 years of teaching experience, he can be contacted at email: c.venkateswarlu@iare.ac.in.

Dr. V. Siva Nagaraju is a professor in the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE). He holds a Ph.D. degree in Electronics and Communication Engineering with a specialization in Microwave Engineering. With over 21 years of academic experience, Dr. Nagaraju is known for his expertise in teaching core electronics subjects and has contributed significantly to the academic and research community. He can be contacted at email: v.sivanagaraju@iare.ac.in.