# Exploring Innovations in Speech Recognition: An in-depth Analysis and Real-world Applications

Dishank Kumar Yadav
*Information Technology*
*KIET Group of Institutions*
Ghaziabad, India
dishankkumaryadav@gmail.com

Chirag Sharma
*Information Technology*
*KIET Group of Institutions*
Ghaziabad, India
sharma.cs121s@gmail.com

Archay Singh
*Information Technology*
*KIET Group of Institutions*
Ghaziabad, India
archaysingh14@gmail.com

Dev Sriwal
*Information Technology*
*KIET Group of Institutions*
Ghaziabad, India
devsriwal93@gmail.com

Sanjeev Kumar
*Information Technology*
*KIET Group of Institutions*
Ghaziabad, India
sanjeev.kumar@kiet.com

***Abstract*— Speech recognition, a field with a rich history, enables computers to accurately transcribe spoken words captured via a microphone, facilitating human-computer interaction. In the last ten years, speech recognition has become much better by using deep neural networks. These networks replaced old methods and improved how accurately computers understand speech. We have used DeepSpeech2 model for speech recognition and added some changes which helps computers understand different kinds of speech, like accents and noisy backgrounds. Deep learning has played a big role in making this happen. It helps computers learn from sound directly and makes them better at understanding speech. Overall, these improvements mean that computers can now understand speech much better, no matter how it sounds. Recurrent Neural Networks, Convolutional Neural Networks had also played crucial roles in this advancement, learning patterns from sequential data, and enhancing system robustness. These innovations have led to more accurate and adaptable speech-to-text systems, fostering broader applications, and meeting diverse user needs and convert into more simpler words which helped us to achieve error rate of 17 percent.**

**Keywords— CNN (Convolutional neural network), Speech Recognition, RNN (Recurrent neural network), ASR (Automatic speech recognition), DNN (Deep neural network).**

## I. INTRODUCTION

Speech recognition is a subject of research and with a long history, allows computers to capture and accurately identify spoken words through a microphone. This process facilitates the exchange of information using acoustic signals. Essentially, it involves the scientific exploration of efficient communication between humans and computers. [1] A breakthrough in integrating deep neural networks with other models a decade ago resulted in a notable improvement in automated speech recognition (ASR) accuracy. In the recent past, a significant and revolutionary change took place as the hybrid modeling shifted towards to end-to-end modeling. E2E works for the seamless transformation of given input into a sequence of simultaneous tokens of output in which everything is present in a single neural network. This eliminates the necessity for the traditional automatic speech recognition (ASR) components that have been in use for decades. [2] [3] [4] E2E learning offers enormous advantages by handling noisy surroundings, different accents, and various languages, which make overall communication skills better. As automatic speech recognition (ASR) systems advances, the applications of speech technology expanded eventually. Meeting diverse needs and high expectations necessitates as well as ensuring the performance of automatic speech recognition (ASR) components in challenging acoustic conditions is improving with time. These conditions may involve making it proficient enough so that it can detect or manage environmental noise and overlapping speech on its own. The evolution toward large, versatile, multi-domain, and multilingual ASR models introduces complexity in handling noise and variations.

[5] End to End modeling in speech recognition can be defined as the development of a single neural network input audio data or signals into corresponding text, eliminating the need of intermediate representation or modules. Deep speech is an end-to-end speech recognition model which is capable of taking English text from audio speech as input and. It firstly converts the input speech into spectrograms then applies neural networks (CNN and RNN) on it and finally produces the text output using CTC.

The expansion of deep learning has led to great advancements in speech-to-text technologies, which have helped in the evolution of automatic speech recognition. Deep learning models can automatically learn hierarchical features from raw audio signals. Deep learning facilitates end-to-end modeling, where a single neural network can be trained to map input audio signals directly to text outputs without the need for intermediate representations. Deep neural networks like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models, are capable of learning patterns and representations from sequential data. The significance of deep learning in speech-to-text systems lies in its ability to automatically learn relevant features, model complex relationships in sequential data, handle temporal dependencies, and improve overall system robustness. These advancements have contributed to more accurate, adaptable, and scalable speech recognition systems, making them increasingly valuable in various applications, from virtual assistants to transcription services.

## II. LITERATURE SURVEY

[3] State-of-the-art speech recognition models demand extensive transcribed data, especially end-to-end (E2E) models, compared to traditional hybrids. Semi-supervised

learning (SSL) mitigates this by utilizing abundant unlabelled data. Common SSL methods include self-training, pre-training with restricted Boltzmann machines, entropy minimization, and graph-based approaches. Recently, transfer learning from pre-trained language models has excelled in NLP tasks. We got an idea from this and created a method to improve speech recognition. Our method learns about sound patterns from data that doesn't have labels and then uses this knowledge to help understand speech that does have labels but not as much. We use certain techniques to learn these patterns and then train the system to recognize speech better.

[4] wav2vec-U is a tool that makes speech recognition models without needing labels. It looks at small pieces of speech and learns from them. Even though it doesn't have labels, it works just as well as models that do. It works for many languages, even those with fewer resources. But turning text into phonemes (speech sounds) for different languages is hard since not all languages have tools for it. [7] In the future, we could make tools for more languages or find ways to do this without labelled data. Also, improving how we break down speech and handle different sizes of speech parts could make the method even better.

[5] Attention-based end-to-end speech recognition systems directly convert spoken words into written text. These systems usually need labelled pairs of audio and text, which can be expensive to get. Using only text to train these systems adds more complexity and requires additional resources. MUTE is a different method that includes text directly when training without needing any audio or outside models. It's efficient because it only needs one pass during the final step. MUTE does better than other methods on LibriSpeech datasets and doesn't require an extra language model. Future work includes refining MUTE's adaptation methods, integrating with audio-only techniques, and extending its applicability to different model architectures and tasks. MUTE presents a promising approach to incorporate language-level information effectively into end-to-end speech recognition models.

[6] Pre-training is crucial for learning universal features from large datasets and benefits tasks with limited data. Methods like BERT, BART, and wav2vec2.0 have become standard for various speech and language tasks. Integrating information from both speech and text data enhances performance. We introduce STPT, a framework combining speech and text in pre-training for speech-to-text tasks. It leverages self-supervised speech and text tasks to improve representation learning. Our approach achieves state-of-the-art results, outperforming existing systems in tasks like speech-to-text translation and automatic speech recognition, demonstrating its effectiveness in leveraging both modalities for better performance.

[7] Automatic speech recognition (ASR) has evolved from traditional hybrid models to end-to-end (E2E) models, which directly translate speech to text using a single network. E2E models offer advantages such as streamlined pipeline, compactness, and improved accuracy. However, practical factors like streaming and latency still favor hybrid models in commercial systems. Recent trends show a shift towards E2E models with advanced encoder structures like RNN-T and Transformer, aiming at efficient and multilingual applications. Adaptation remains crucial for E2E models to replace hybrids, particularly in domain and speaker adaptation. Overall, while E2E models excel academically, commercial deployment hinges on addressing practical challenges.

[8] Advances in deep learning have enhanced automated speech recognition (ASR) systems, widely used in virtual assistants, dictation, translation, and accessibility tools. Concerns about racial bias in ASR systems echo broader issues in machine learning applications. Analysis of five commercial speech-to-text tools reveals disparities, largely attributed to acoustic model performance gaps with African American Vernacular English. While regional differences in data collection pose limitations, evidence suggests AAVE speech influences results.[11] Future research should explore error rates among white and black speakers from the same region to further understand and address biases in ASR technology.

[9] Deep neural networks, like convolutional neural networks (CNNs), improve speech recognition by breaking down the process into smaller parts. CNNs, which are good at sorting things into groups, are used in a new method for recognizing uncommon speech, like Punjabi Gurbani recitation. This method uses specific parts of the sound and connections between them to avoid making mistakes. It also looks at the details of the sound to understand it better. Unlike other methods, CNNs can handle noisy and distorted sounds better. The method uses different layers to process the sound and make sense of it, helping to make speech recognition better.

[10] Traditional speech-to-text translation involves separate ASR and MT models, leading to time delays and parameter redundancy. Recent end-to-end models show promise but struggle without transcriptions. Prior methods use pretraining or multi-task learning, yet fail to fully utilize information exchange between tasks. A new learning model does both recognizing and translating at the same time, making them better by sharing information. A special attention layer helps predict using outputs from both tasks. Also, waiting a bit before translating makes it better. [13] Lots of tests on different languages show that this way of doing things works well for translating speech.

[11] Emotion recognition (ER) in human-computer interaction remains challenging but crucial for effective communication. Speech emotion recognition (SER) systems, focusing on audio input, have evolved from traditional spectral feature-based methods to end-to-end deep neural models. It suggested a new speech emotion recognition (SER) model trained using multi-task learning (MTL), which gets really good results on the IEMOCAP dataset. It uses a pretrained tool called wav2vec-2.0 to help pick out important features from the speech. The model learns both SER and automatic speech recognition (ASR) tasks at the same time. [15]They did some tests to show that learning both tasks together helps, and they looked at how ASR affects SER. Also, the model can transcribe speech as a bonus.

## III. Methodology

For our research, we used the LJ Speech Dataset, a publicly available collection comprising 13,100 short audio clips offering a single speaker studying passages from 7 non-fiction books. This dataset is comprised of schooling speech synthesis models, encompasses various texts published between 1884 and 1964, all inside the public

domain. Each audio clip, starting from one to ten seconds, is meticulously paired with a transcription, providing a complete period of about 24 hours. The recordings, captured in 2016-17 by way of the LibriVox venture, provide a rich useful resource for textual content-to-speech packages. The different length of clips in the dataset makes it great choice for training strong and flexible voice synthesis models in scientific research that uses deep learning.

In our studies, we embraced the evolution in speech recognition technology, especially the transition from conventional hybrid modelling to give up-to-stop (E2E) modelling.

One high-quality deep learning model that we implemented in our experiments is Deep Speech 2. This E2E speech recognition model demonstrates the functionality to transform English audio speech without delay into corresponding textual content without relying on conventional ASR components. The model begins by means of transforming input speech into spectrograms after which applies an aggregate of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) is applied on it with ReLU as an activation function. The very last step involves the usage of Optimizer and Evaluation matrix to produce the preferred textual output.

DeepSpeech2 represents a powerful instance of ways of deep learning, in particular CNNs and RNNs, which can be used for simplifying modelling process in speech recognition. This approach lets in for an extra direct and green mapping from audio inputs to textual outputs, contributing to advanced accuracy and adaptability in diverse acoustic conditions, consisting of noisy environments, accents, and numerous languages. The usage of deep learning models like DeepSpeech2 emphasizes our commitment to remaining at the forefront of advancements in the era of speech recognition.

In our research, we have use ReLU as an activation function which stands for rectified linear unit which addresses the problem of vanishing gradients and adds the characteristic of nonlinearity to a deep learning model. It explains the portion of its argument that is positive. In deep learning, it is among the most widely used activation functions.

TABLE I. MATRIX OF DATASET

| | |
|---|---|
| ***Total Clips*** | 130100 |
| ***Total Words*** | 222715 |
| ***Total Characters*** | 1308678 |
| ***Total Duration*** | 23:55:17 |
| ***Mean Clip Duration*** | 6.57 sec |
| ***Min Clip Duration*** | 1.11 sec |
| ***Max Clip Duration*** | 10.10 sec |
| ***Mean Words per Clip*** | 17.23 |
| ***Distinct Words*** | 13821 |

### *A. Training Metrics*

Connectionist Temporal Classification Loss*:* CTC loss is an essential metric used at some point of the schooling segment of give up-to-stop fashions like DeepSpeech2. It measures the dissimilarity among the anticipitated an goal sequences, assisting the version modify its parameters to reduce the discrepancy and improve accuracy.

The ground truth of word sequence

Acoustic frames

$$\text{lctc} = -\log P(\boldsymbol{S}|\boldsymbol{X})$$

$$P(\boldsymbol{S}|\boldsymbol{X}) = \sum_{c\epsilon A(S)} \mathrm{P}\,(\boldsymbol{C}|\boldsymbol{X})$$

Sum over all possible path (eg. cc&&tt)

$$\mathrm{P(C|X)} = \prod_{t=1}^{T} y(c_t, t)$$

joint probability of a path (eg. cc&aa&tt)

Optimizer: We applied Adam which stands for adaptive moment estimation as an optimizer for the model. The optimizer, known as Adam, adjusts the learning rate for each neural network weight by utilising estimates of the first and second moments of the gradient.

Mathematically

$$w_{t+1} = w_t - \alpha m_t$$

Where

$$m_t = \beta m_{t-1} + (1-\beta)[\delta L/\delta w_t]$$

### *B. Evaluation Metrics*

Word Error Rate(WER*)*: WER is a broadly used metric for comparing the accuracy of speech reputation structures. It quantifies the differences among the expected and reference transcriptions in phrases of phrases. A decrease WER indicates higher accuracy.

Word error rate can then be computed as:

*S* is the number of substitutions

*D* is the number of deletions

*I* is the number of insertions

*C* is the number of correct words

*N* is the number of words in the reference (N=S+D+C)

$$\boldsymbol{WER} = (S + I + D)/N$$

Character Error Rate (CER): CER is another vital metric that measures the accuracy of person characters within the expected and reference transcriptions. Similar to WER, a decrease CER suggests higher performance. The formula to calculate CER is as follows: CER = [ (i + s + d) / n ]*100.

### *C. Learning Architecture*

We have used a deep learning model similar to DeepSpeech2 to create the required model for our speech recognition system. It is a set of various different speech learning models that are designed for automatic speech recognition tasks. It uses word rate error as the main evaluation metric. The main aim of our model is to detect

speech and convert it into text as accurately as possible using deep learning techniques.

The various techniques used by DeepSpeech2 are: -

1. It makes use of CNNs, which are famous for capturing patterns in data, for the task of feature detection. As its layers extract important acoustic features from the audio and extract them for use.

2. In addition, CNN layers extract a sequence of flattened vectors for processing by a recurrent neural network. It does this to retrieve contextual or relevant data and create output probabilities.

3. It uses CTC, which stands for Connectionist Temporal Classification Loss, as it allows the model to learn directly from the audio waveform and helps the network generate output directly from the waveform created from the audio.

4. It also includes batch normalization layers, which result in faster merging of the text and also improve the performance of the model.

5. During inference, it uses a very powerful tool, i.e., beam search decoding, which generates the most optimal transcription of the input audio. The beam search technique explores multiple output sequences and selects the most optimal one for the output, because of which the accuracy is also improved.

6. And at the end, to prevent the model from overfitting, it uses data augmentation techniques such as time stretching, frequency masking, and noise injection.

## IV. Experimental Results

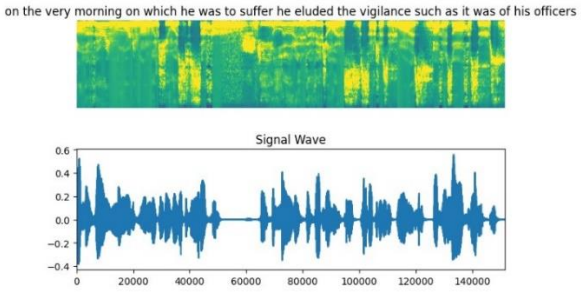


Fig. 1. Visualization of Data

TABLE II. Parameter Table

| Layer | Output Shape | Parameters |
|---|---|---|
| Input (Input Layer) | (None, None, 193) | 0 |
| expand dim (Reshape) | (None, None, 193, 1) | 0 |
| conv_1 (Conv2D) | (None, None, 97, 32) | 14,432 |
| conv_1_bn (Batch Normalization) | (None, None, 97, 32) | 128 |
| conv_1_relu (ReLU) | (None, None, 97, 32) | 0 |
| conv_2 (Conv2D) | (None, None, 49, 32) | 236,544 |
| conv_2_bn(Batch Normalization) | (None, None, 49, 32) | 128 |
| conv_2_relu (ReLU) | (None, None, 49, 32 | 0 |
| reshape (Reshape) | (None, None, 1568) | 0 |
| bidirectional_1 (Bidirectional) | (None, None, 1024) | 6,395,904 |
| dropout (Dropout) | (None, None, 1024) | 0 |
| bidirectional_2 (Bidirectional) | (None, None, 1024) | 4,724,736 |
| dropout_1 (Dropout) | (None, None, 1024) | 0 |
| bidirectional_3 (Bidirectional) | (None, None, 1024) | 4,724,736 |
| dropout_2 (Dropout) | (None, None, 1024) | 0 |
| bidirectional_4 (Bidirectional) | (None, None, 1024) | 4,724,736 |
| dropout_3 (Dropout) | (None, None, 1024) | 0 |
| bidirectional_5 (Bidirectional) | (None, None, 1024) | 4,724,736 |
| dense_1 (Dense) | (None, None, 1024) | 0 |
| dense_1_relu (Dense) | (None, None, 1024) | 0 |
| dropout_4 (Dropout) | (None, None, 1024) | 0 |
| dense (Dense) | (None, None, 32) | 32,800 |

a. Total Params: 26,658,480

b. Trainable params: 26,628,352

c. Non-Trainable params: 128

TABLE III. EPOCH TABLE

| Number of Epochs | Word Error Rate (approx.) |
|---|---|
| 10 | 54% |
| 20 | 45% |
| 30 | 36% |
| 40 | 22% |
| 50 | 17% |

As mentioned previously, the model used in our speech-to-text detection model is similar to DeepSpeech2, so similarly, the results are regulated by calculating the word error rate. Word-error rate is the measuring technique used to measure the accuracy of speech recognition models. Basically, it is the total number of errors divided by the total number of words. So, the word-error-rate-out model is giving 17% error rate for 50 epochs (an epoch is one complete traversal of a training dataset through an algorithm).

## V. CONCLUSION

In concluding our research, we initially emphasized and extended the program's versatility and potential impact. As we transitioned into the development phase, our focus sharpened on creating efficient speech recognition software. Our research has a significant advancement in Natural Language Processing and Machine Learning. Though the utilization of deep learning techniques and training our model on significant amount of Speech data we demonstrate remarkable accuracy and efficiency in translating spoken words into text. Our model has a potential for a wide range of applications in creating useful tools for the individuals with disabilities by helping to understand spoken words more accurately in real time which helps them in daily task like web searching, form filling etc. and our model also enables improved communication and productivity in various domain. Our model has the ability of handling diverse accents and noisy environments. In comparison to other model present in the industry our model provides a remarkable 17% error rate which is better than some and competitive to many others..

## REFERENCES


[1] G. Sable, M. Rupali, S. Chavan, and G. S. Sable, "An Overview of Speech Recognition Using HMM International Journal of Computer Science and Mobile Computing An Overview of Speech Recognition Using HMM," 2013. [Online]. Available: https://www.researchgate.net/publication/335714660

[2] A. Narayanan, J. Walker, S. Panchapagesan, N. Howard, and Y. Koizumi, "Mask scalar prediction for improving robust automatic speech recognition," Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.12092

[3] *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[4] J. Li *et al.*, "Recent Advances in End-to-End Automatic Speech Recognition," *APSIPA Trans Signal Inf Process*, vol. 11, p. 8, 2022, doi: 10.1561/116.00000050_supp.

[5] S. Dua *et al.*, "Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural Network," *Applied Sciences (Switzerland)*, vol. 12, no. 12, Jun. 2022, doi: 10.3390/app12126223.

[6] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised Speech Recognition." [Online]. Available: https://github.com/pytorch/fairseq/tree/

[7] C. Le *et al.*, "ComSL: A Composite Speech-Language Model for End-to-End Speech-to-Text Translation." [Online]. Available: https://github.com/nethermanpro/ComSL.

[8] P. Wang, T. N. Sainath, and R. J. Weiss, "Multitask Training with Text Data for End-to-End Speech Recognition," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.14318

[9] Y. Tang *et al.*, "Unified Speech-Text Pre-training for Speech Translation and Recognition," Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.05409

[10] "Racial disparities in automated speech recognition," vol. 117, no. 14, pp. 7684–7689, 2020, doi: 10.1073/pnas.1915768117/-/DCSupplemental.y.

[11] M. E. Matre and D. L. Cameron, "A scoping review on the use of speech-to-text technology for adolescents with learning difficulties in secondary education," *Disability and Rehabilitation: Assistive Technology*, vol. 19, no. 3. Taylor and Francis Ltd., pp. 1103–1116, 2024. doi: 10.1080/17483107.2022.2149865.

[12] Y. Liu *et al.*, "Synchronous Speech Recognition and Speech-to-Text Translation with Interactive Decoding." [Online]. Available: www.aaai.org

[13] Z. Weng and Z. Qin, "Robust Semantic Communications for Speech Transmission," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.05187

[14] S. A. El-Moneim, M. A. Nassar, M. I. Dessouky, N. A. Ismail, A. S. El-Fishawy, and F. E. Abd El-Samie, "Text-independent speaker recognition using LSTM-RNN and speech enhancement," *Multimed Tools Appl*, vol. 79, no. 33–34, pp. 24013–24028, Sep. 2020, doi: 10.1007/s11042-019-08293-7.

[15] Y. Zhang *et al.*, "Identifying depression-related topics in smartphone-collected free-response speech recordings using an automatic speech recognition system and a deep learning topic model," *J Affect Disord*, vol. 355, pp. 40–49, Jun. 2024, doi: 10.1016/j.jad.2024.03.106.