WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Automatic Speech Recognition Using Deep Learning

Pitchika Devisreesai

CSE, GMRIT, Rajam, India
devisreesaipitchika@gmail.com

ABSTRACT—

Speech recognition is a technology that enables a computer or a device to identify and process spoken language, converting it into text or commands. This involves analysing the sound waves of spoken words and breaking them down into phonetic units. Speech recognition is commonly used in virtual assistants and voice-controlled applications. Deep transfer learning is a technique in machine learning where a model trained in one task is used to solve it in different way with less data. It helps to perform better on new tasks without needing the large data or time to train. There are different methods like CNN, seq2seq and end to end models. End-to- end is the best model in automatic speech recognition. It is helping to enhance the efficiency in automatic speech recognition. By studying this model to avoid the unnecessary y intermediate steps, it is possible to achieve high performance with less data ,leading to be more efficient in automatic speech recognition systems.

Keywords— Automatic speech recognition, data scarcity, deep transfer learning, language modelling, neural networks

Introduction

Automatic Speech Recognition (ASR) technology is a procedure that transcribes speech into written language. It has become almost essential in services such as virtual assistants like Alexa, Google, and voice transcription, voice-controlled smart devices. ASR improves the interface of human- computer interaction and can allow users to interact with a computer without having to type. This technology employs progressive signal processing and neural networks to understand spoken words correctly ignoring any form of noise and accent variations on words Recognizing spoken words is a complex process involved in the ASR starting with the audio signal processing.

In this phase, the system records the utterance of the speaker, removes the noise background from it and normalizes the audio for further analysis. Speech activity must be separated from undesired noise at this stage, and this make noise reduction and filtering a vital step. This stage brings an audio signal which is fit for deeper analysis.This follows after satisfactory audio cleaning.Following cleaning, feature extraction occurs. Here the ASR system decomposes the audio into pitch, energy and frequency, so that it can develop a simplified model of the sound.

As these audio signals are reduced into these base aspects, ASR can easily target aspects of how speech is distinct in words. These can be said to be like a “fingerprint” for the sounds so that the ASR system can quantify the many distinct phonemes, that is the switches of sound that comprise each word.The next procedure is both the acoustic and language modeling. The acoustic model aligns the acoustic features of the stimuli to a given phoneme. Specifically, these models are trained on huge databases of human vocal performance which also provides for different ways of pronouncing the same word, different accents and different speaking styles. Besides, the language model has a capacity of estimating probabilities of word sequences in accordance to grammatical rules and context. Using both of these models allows ASR systems to

not only depict sound alone but also how the words expected to be grouped in a particular sentence.Finally decoding fuses all these features predicting the probable words sequence in a correct way. This process is done through deep learning to ensure the ASR system is ever adaptable to new knowledge. And when all these steps are accomplished in unison, it brings high levels of accuracy in the transcription through ASR.

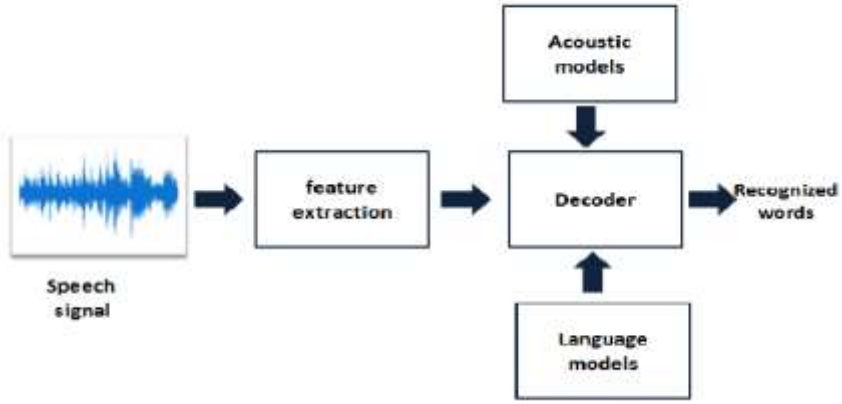

Fig.1.Architecture of ASR

and is slowly becoming significantly important tool in areas such as accessibility services and hands-free solutions. Voice assistants have extremely changed how drivers operate their vehicles by enabling them to operate a variety of systems with spoken commands rather than manual inputs.

Simultaneously developments in speech recognition systems such as the Explain simple system are making it possible to Explain across languages in real time while concentrating on the spectral and temporal aspects of speech.

To increase recognition Precision and Representation Productivity these systems make use of a variety of supervised and semi- supervised Representations that have been trained on extensive lexicons and a wide range of voice Information including in languages like Arabic and Mandarin.

Full supervised self- supervised and decrepit supervised representations are examples of new advancements inch language methoding notably once transaction with Complex languagesprosodic qualities and the power to set to disparate language layouts accents and speeds were better away methods such as arsenic care mechanisms and lstm representationscombining amp great number of unlabeled Information with amp mean number of labeled Information has been successful contingent away semi-supervised acquisition techniques such as option pseudo- labellingtechniques that Improve audio-visual character language credit systems are decent further available and good.

These innovations aim to improve communication by focusing on human auditory Methoding Representations and leveraging pre-trained deep learning Representations to identify speech emotions very importantly improving the Operator Encounter in everyday interactions.

Literature survey

An Arab driver's assistance gadget uses speech recognition to improve accessibility and speed up vehicle control, which improves human-vehicle interaction [1]. This study emphasises how voice-activated gadgets might encourage safer driving environments.Similar to this, it created a calculator-assisted explanation system for multilingual conferences that relies on automatic speech recognition, showing how technology may facilitate real-time communication and comprehension across linguistic barriers [2].Deep learning techniques have greatly improved speech methodology by enhancing nerve-related speech detection and overcoming the challenges posed by background noise. A time-frequency attention module was also created by them. This access illustrates how learning sounds enhances the precision and clarity of language [3].They also offered Banspeech and a number of global benchmarks for Bangla language proficiency; emphasising the essential information sets for enhancing the robustness of ASR systems in harsh environments and optimising their performance across several languages and demanding situations [4] is a crucial search centre. Through the introduction of an online hybrid CTC/attention-based ASR structure, they demonstrated the system's potential for real-time healthcare applications by enabling end-to-end recognition [5].

The growing interest in employing prosodic elements in ASR representations to improve their effectiveness in more natural conversational settings is best illustrated by this student [6] Due to developments in ASR, people with speech impairments have also been provided with assistive technologies. To make it easier for operators who have a problem with language, the dysarthric language revolutioniser, which is based on the amp sound learning, was developed [7]. They proposed a teaching strategy for speech enhancement based on knowledge distillation, and they demonstrated that it was beneficial in improving the robustness of ASR systems against noise [8]. Likewise, they investigated the application of long short-term memory (LSTM) webs for automatic speech recognition, providing enhanced efficacy in complex acoustic settings [9]. Using a large amount of labelled data, they developed a novel pseudo-labelling technique [10].

This search area is crucial for applications such as real-time communication systems or arsenic league settings, where a large number of people are speaking in public at once [11]. Weak supervision uses in specialised fields were advanced by their investigation of weakening human-based features for content-based speech recognition [12]. In the medical field, ASR technology has also proven to be quite useful, particularly in the diagnosis and treatment of speech disorders [13]. They developed a full target speaker system that includes speaker activity detection, which is essential for improving performance in difficult environments [14]. It pointed out the benefits of integrating the visual and audio data by enhancing the audio-visual speech recognition through a multi-level distortion measure [15].

They extend the use of ASR in rehabilitation by providing a method to improve dysarthric speech recognition using a sequence-to-sequence model [16].For enhancing the speech improvement representations and multitask methoding, they introduced combined speech-text embeddings and proposed a single-channel speech separation system for verifying individual speakers in the overlapping audio streams [17]. Identifying the speaker in overlapping speaking situations has presented many difficulties. manipulation of the neurological system in this way Webs centre on the detection and spacing of individual speakers in multi-talker settings [18]. It enhances the high accuracy in recognising emotion states [19]. It develops to predict the speech risk and severity [20].

Using speech augmentation approaches, they have attempted to improve segmentation and recognition precision in speech recognition for patients with dysarthria [21].They looked at how well speech information could be transferred and adjusted for dysarthric speaking recognition [23].Using human hearing representations significantly improved the resilience of ASR systems against noise [24]. Prosody, the rhythm and repetition of language, is a crucial aspect of nursing. Using a role model to make false claims Nursing students would be involved in the process to help separate inflection representations in language reconstruction, which is an important step in creating outputs that seem more natural. The pretrained speech representations and EEG data were merged to improve The usefulness of cross-disciplinary approaches in noisy settings emphasises their importance in ASR research [25].

Methodology

Automatic speech recognition converts spoken language into text by analyzing audio signals. It involves preprocessing audio, extracting meaningful features, and applying models to map these features to language. Advanced ASR systems leverage neural networks and language models to improve accuracy, handling diverse accents, noise levels, and linguistic contexts

End to end

The transformation of an audio signal into text is preceded by a set of separate stages, providing for certain unique functionality. These modules are the feature extractor, an acoustic model, a pronunciation model, a language model and a decoder. All these components are developed independently and need a lot of language knowledge and resources to construct a good ASR system. In an E2E ASR model, however, all of these functions are learned in one model that directly maps an audio signal to text meaning that it is far less fragmented and much easier to train.In conventional ASR systems, the feature extraction block is the initial step in the signal processing chain. This module accepts the raw audio data and converts it to feature space, commonly Mel-spectrograms or Mel Frequency Cepstral Coefficients (MFCCs) that retain all the speaker's phonetic information. Such features include; the frequency of the audio signals which are used to differentiate structurally, many phonetic sounds. Feature extraction has the effect of simplifying the input data to a more manageable form for the acoustic model to analyse. This process however has to be done efficiently with a considerable amount of engineering in a way that the extracted features are meaningful and relevant to the rest of ASR pipeline.

After feature extraction, the acoustic model analyses the extracted features in a bid to identify the phonetics present in the audio. In this model, it's common that it is trained in a way that it learns to identify small sound chunks which would represent a phoneme. The acoustic model provides probabilities of phonetic units for each segment of audio, which in effect determines what sounds are present. But this is only half the story; phonemes cannot form words on their own, they need other components.

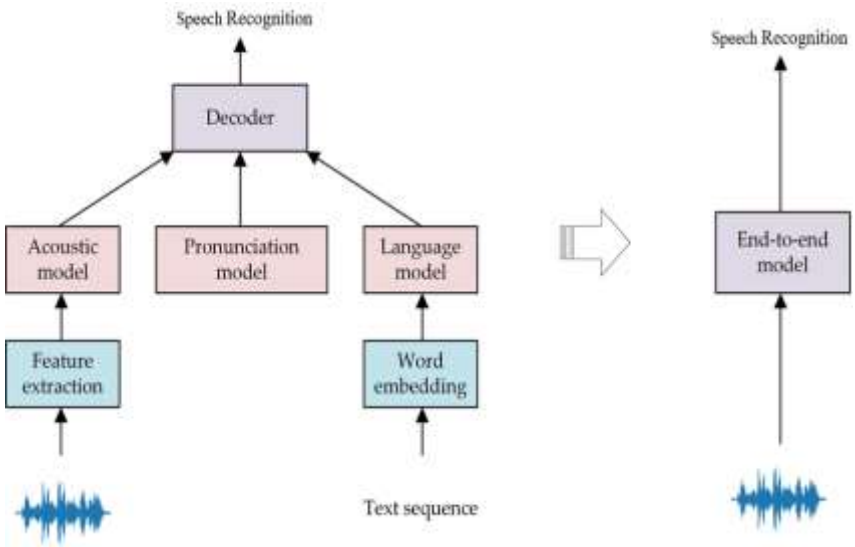

Fig.2.methodology of end-end

In order to form these phonemes into words that can be understood the system uses a pronunciation model. The resulting pronunciation model connects sequences of phonemes to words, according to the way phonemes group together in different languages. For instance, it will also understand that if the English phonemes are selected as "k", "x", "t" it means "cat". This step involves knowledge in phonetics and phonemics or pronunciation patterns, a

dictionary or/and lexicon which must be developed by professional linguistic personnel.The next part is the language model, increasing the layer of comprehension by estimating which word sequence is probable adhering to context and syntactic structure. While the pronunciation model ensures that selected phonemes can be combined into words, the language model facilitates to ensure that selected words belong to a particular context. For example, considering the underlying words or partial phrase such as 'the cat' will have the language model estimate the probability of the word that will follow it to be 'sat' and not "dog". In this component it is essential in avoiding making errors in the way we choose to write either in complicated phrases or even ambiguous words. Language models are used to gain an understanding of the frequency and context of use of words and phrases in text but also add additional layers of complication and computation.The more conventional ASR system is completed with a decoder that integrates the results from the acoustic model, the pronunciation model, and the language model. This is a complex multiple stage process that is not linear but dovetails through several modules and each module is trained with a different objective.While these systems can be very reliable these systems are difficult to develop maintain and update with new languages or dialects as each component is unique and often requires its own expertise and optimization. Furthermore, since each component is developed independently, problems in one module can affect the other modules, and hence the entire system is more susceptible to errors.In contrast, the E2E ASR model, shown on the right side of the diagram, integrates all these functions into a single, unified model that learns to map audio input directly to text output. Instead of relying on separate acoustic, pronunciation, and language models, the E2E model is trained as a single neural network that can learn complex relationships between sounds and text through large amounts of paired audio and transcription data. Using architectures like Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), or Transformers, the E2E model can recognize patterns in speech and directly generate transcriptions. This approach not only simplifies the system but also makes it more adaptable to diverse languages and environments. Since the model learns from raw data, it can generalize across accents, dialects, and noisy backgrounds, reducing the need for handcrafted linguistic resources. E2E models offer a streamlined and more robust solution for modern speech recognition tasks

It helps to improve accuracy, handling diverse accents, noise levels, and linguistic contexts *CNN*

Recently, the models built using the convolutional neural network (CNN) and Long Short Term Memory (LSTM) networks are preferred that are well suited for the sequence & time series data such as video sequences. This in combination of both CNNs and LSTMs helps the model to capture spatial and temporal features of the data sets. The architecture starts with Data Preprocessing, where data set for input is first preprocessed before going for training the model and making predictions. Some of the common techniques performed under preprocessing phase include normalization, resizing and data augmentation of the material in the data set, which in one way or another makes the data set better hence improving the performance of models.

Recently, the models constructed using the convolutional neural network (CNN) and Long Short Term Memory (LSTM) networks are preferred that are suitable for the sequence & time series data such as video sequences. This in combination of both CNNs and LSTMs enables the model to capture spatial and temporal features of the data sets. The architecture begins with Data Preprocessing which heavily involves working on the data set for input before it is used for training the model and making predictions. Some of the common techniques that are performed under preprocessing phase include normalization, resizing and data augmentation of the material in the data set, which in one way or another makes the data set better hence improving the performance of models.

Data pre-processing is an important aspect of deep learning because of its ability to make data amenable to being utilised in training algorithms. Normalisation brings data in a certain range which makes the algorithm more fair in terms of the weights and brings faster convergence. Scaling changes the size of images or sequences to a standard size for the purpose of processing multiple items at once. The processes like cropping, flipping or rotation are applied to transform the existing data into further data thus increasing the size of the data set artificially. It also prevent overfitting besides making the model more resistant as it is trained to identify different representations of the same data set. In aggregate, these preprocessing operations form a more stable input to the successive levels of the model. The input data that have gone through preprocessing are passed to a number of convolutional that work at learning spatial features. Convolutional layers are very useful for image and spatial data since they work with local patterns and between them hierarchal structure.

In convolution operation, different filters are passed cross the input and produce feature maps which are important features.After the convolution, the authors perform batch normalization on these feature maps. This technique stands for the normalization of output of the convolutional layers and helps to learning process have a stability and a higher speed of the training process and the probability of over learning is small.

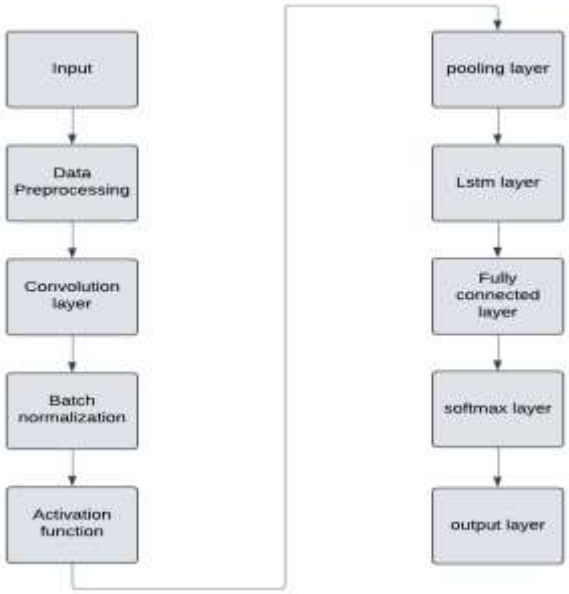

Fig.3.flowchart of cnn

Batch normalization therefore scales the outputs allowing the model to converge faster than traditional networks making it useful in the architecture. After features have been obtained from the spatial information and the features have been multiplied by the weight matrix, the features mapped through a normalization process undergo activation and the most commonly applied is Rectified Linear Unit (ReLU). The activation functions enable non linearity within the model so that the model can learn sophisticated relations within the data. Even after activation, the feature maps also perform pooling operation commonly known as max-pooling, which again downsamples it. This operation preserves the important features of the data as well as rejecting the unimportant features. Pooling truly reduces the parameters needed which in turn reduces the calculation and also minimizes over fit problem since the model is made simpler. These then are passed to LSTM layers in a format appropriate for sequential processing of feature maps of a lesser dimensionality. After the convolution and pooling layers, the succeeding data undergoes LSTM layers because the network is intended for sequential data analysis. LSTMs in particular are good at capturing temporal characteristics and within them temporal dependencies so the model will be able to keep some sequence information throughout sequences and make predictions of temporal patterns. This characteristic is most useful where there are associations such as between video activities and times in time series. The LSTM layers allow the network to remember useful information of previous time steps which advancing its prediction ability. Lastly, the output of the LSTM layers is connected to a fully connected layer That is if the network is used for classification the final layer will contain a softmax activation function. The last of these generates probabilities within the possible classes and returns the class that has the highest probability. This architecture of CNNs and LSTMs is therefore useful in performing both temporal and spatial analysis of sequential data and is therefore useful in many applications.

A speech-to-text system designed to transcribe Nepali audio into text. The process begins with raw audio input, which is initially converted into a spectrogram using the Short-Time Fourier Transform (STFT). A spectrogram represents the frequency content of the audio signal over time, allowingthe system to visualize the audio's pitch and intensity patterns. This is crucial because it helps transform the sound into a form that is easier to analyze for patterns relevant to speech, creating a time-frequency representation that captures essential speech features.After generating the spectrogram,

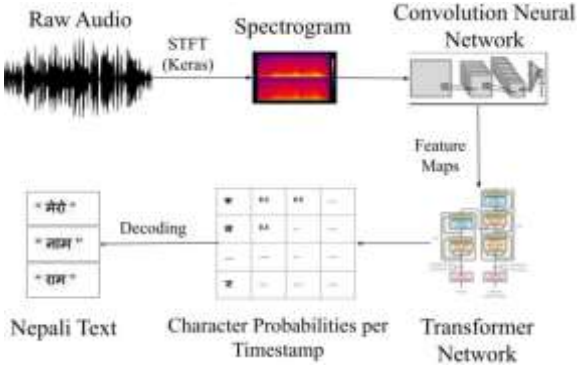

Fig.4.Methodology of cnn

it is passed through a Convolutional Neural Network (CNN). CNNs are highly effective at detecting local patterns in images, making them well-suited for analyzing spectrograms. In this context, the CNN extracts feature maps from the spectrogram, identifying essential characteristics of the sound that

represent linguistic components, like phonemes (the smallest units of sound in language). These feature maps serve as condensed and structured representations of the original audio, capturing significant details while reducing noise and irrelevant information.

The extracted features are then processed by a Transformer Network, which is adept at handling sequential data and understanding contextual relationships. The Transformer uses attention mechanisms to analyze the feature maps, producing character probabilities for each timestamp—essentially predicting which Nepali character is most likely being spoken at each moment. This probabilistic output is then decoded to form coherent words and phrases in Nepali. The final result is a transcription of the spoken audio in Nepali text, achieved by combining CNNs for feature extraction and Transformers for sequence prediction, creating an efficient end-to-end speech-to-text pipeline.

Seq-seq

Seq2Seq means Sequence to sequence models, a type of neural network that found really useful in handling natural language processing tasks such as translation, text summarization and chatters. Their core structure relies on two key components: an encoder and a decoder. Encoder takes the input sequence, which converts the desired dimension n to a nearly fixed dimension d, also referred to as the context vector, which provides the relevancy of the input data. In this case, the resulting context vector is combining with the decoder in order to produce an output sequence. Such architecture is useful when input and output are in sequences of different lengths, as for the translation, where the encoder-decoder structure changes the input sequence from one domain (the source language) to another (the target language).The encoder is the first component of Seq2Seq model and its aim is to map input sequence to a fixed size vector that comprises information about the input sequence. Basically, the encoder takes a set of tokens in a given sentence and processes these tokens one at a time using the help of a RNN, LSTM or the GRU. Every token changes the encoder's hidden state that stays aligned with the information acquired on previous tokens of the sequence. In the input sequence at the last state of a sequence, the last hidden state is used as context vector which encompasses all the input sequences and feeds them to the decoder. This setup is most suitable at least for short to moderately long sequences where the sequence's defining characteristics can be compactly encoded.

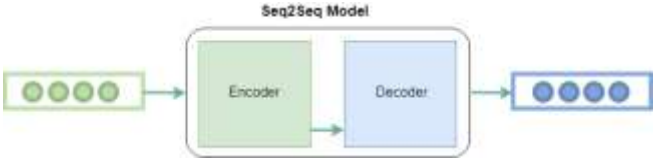

Fig.5.methodology of seq2seq

It then utilizes the context vector that the encoder came up with, in order to give out the output sequence. As with the encoder, the decoder is commonly an RNN, LSTM or GRU, but also can be a CNN. The decoder takes one token at a time and its generation depends on the context vector and the tokens which are generated so far. In each stage the decoder computes a distribution over the output tokens and picks the best one to append to the sequence. This process goes on until all the result sequence is produced because iteration is done in this computation. In training session special tricks such as the teacher forcing can be used to enhance learning rate while in the decoding session, the true preceding token is used instead of the one generated by the model to prevent the propagating of errors.There is one issue that can be attributed to this vanilla Seq2Seq configuration – the architectural design that encodes all contextual information regarding the input sequence into a single, comparatively small and fixed-sized vector. When input sequences are short, this compression is quite good, but as the length of the sequences increases, the model may fail to 'remember' considerable detail and hence the accuracy reduces. This limitation has led to emergence of attention mechanisms to enable the decoder to attend to the relevant parts of the input sequence during generation. In contrast to the more rigid fixed vector of context, states that have an extension in the position domain, which lets decoder adjust weights for input sequence parts, devote sections in text and letting the decoder pay attention to a different token.The other improvements were to the attention mechanisms which essentially helped in retaining the context when completing the sequence generation. The decoder is able to control different aspects of the input sequence with respect to the type of the output token currently being produced. This dynamic access, in fact, helps to manage the complex dependencies, specifically in the tasks like translation where and particular input tokens should be closely connected with and correspond to the output tokens even if these tokens are located far from each other in the sequence. Pior developments were driven from the improvements in the attention based mechanisms within the Seq2Seq models, and the Transformer models expanded on these structures to manage and process even longdistance dependencyIn addition to translation, Seq2Seq models have shown success in other NLP and speech-related tasks that involve variable-length input and output sequences. Speech recognition, for example, benefits from the Seq2Seq framework's flexibility, as audio sequences can vary significantly in length. In these applications, evaluation metrics such as Word Error Rate (WER) or Character Error Rate (CER) are commonly used to assess model accuracy. These metrics help fine-tune model performance and improve accuracy, addressing the challenges inherent in variable-length sequences. Despite the improvements that attention mechanisms bring to Seq2Seq models, longer, more complex texts still pose challenges in maintaining coherence and context, an area that architectures like the Transformer are better equipped to handle.

Results and discussion

This study evaluates the effectiveness of three distinct models-CNN,seq2seq and end to end automatic speech recognition systems using metricssuchasaccuracy,precision,recall,F1 score. The results reveal that end to end ASR systems consistently outperform both cnn and seq2seq models across all metrics that achieving the accuracy levels above 95%.The seq2seq model demonstrates commendable performance particularly in precision and F1 score. These findings underscore the superiority of end to end ASR systems in generalization and prediction capabilities and making them the most effective choice for accurate and reliable speech recognition.The robustness of end to end suggests that they are well-suited for diverse linguistic

contexts and It enhances the applications in real -world scenarios.This positions end to end ASR systems as the preferred solution for applications requiring the high accuracy such as voice-activated systems and transcription services which is useful for the computer-human communication. particularly end to end models,convolution neural networks and sequence-to-sequence.It is to enhance the accuracy and efficiency.end to end models aredirectly mapping the audio inputs to text outputs,while cnn extract the high level features from audio signals and enables the recognition of complex speech patterns. It handles the variable- length sequences and improve the accuracy.

Comparison among method

Model	***Accuracy***	***Precision***	***Recall***	***F1_score***
Cnn	90%	91%	89.90%	88.20%
Seq2seq	95.50%	94.20%	93.80%	90.99%
End to End	95%	92%	94%	92%

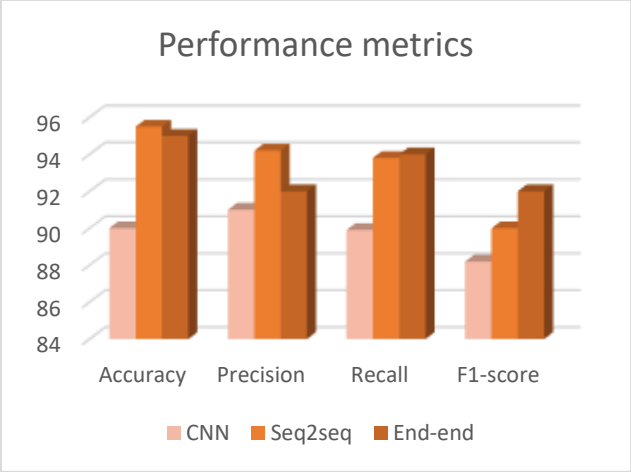

Fig .6.graphical representation of various performance metrics

Conclusion

Research on Automatic speech recognition (ASR) has increasingly emphasized the integration of deep learning techniques,particularly end to end models,convolution neural networks and sequence-to-sequence.It is to enhance the accuracy and efficiency.end to end models are directly mapping the audio inputs to text outputs,while cnn extract the high level features from audio signals and enables the recognition of complex speech patterns.It handles the variable-length sequences and improve the accuracy.Compared to traditional ASR methods,these advanced approaches demonstrate superior performance in addressing the challenges of diverse speech patterns,accents and contextual variations,significantly enhancing the capabilities and reliability of ASR systems across various applications.

References

Jaradat, G. A., Alzubaidi, M. A., & Otoom, M. (2022). A novel Human-Vehicle Interaction assistive device for Arab drivers using speech recognition.*IEEEAccess*,*10*,127514127529https://doi.org/10.1109/access.2022.3226539

Liu, J., Liu, C., Shan, B., & Ganiyusufoglu, Ö. S. (2024b). A computer-assisted interpreting system for multilingual conferences based on automaticspeechrecognition.*IEEEAccess*,*12*,6749867511.https://doi.org/10.1109/access.2024.3400014

Zhang, Q., Qian, X., Ni, Z., Nicolson, A., Ambikairajah, E., & Li, H. (2022). A Time-Frequency attention module for neural speech enhancement. *IEEE/ACM Transactions on Audio Speech and LanguageProcessing*,*31*,462475https://doi.org/10.1109/taslp.2022.3225649

Samin, A. M., Kobir, M. H., Rafee, M. M. S., Ahmed, M. F., Hasan, M., Ghosh, P., Kibria, S., & Rahman, M. S. (2024). BanSpeech: a multi-domain Bangla speech recognition benchmark towards robust performanceinchallengingconditions.*IEEEAccess*,1.https://doi.org/10.1109/access.2024.3371478

Huang, E. H., Wu, C., & Lin, H. (2021). Combination and comparison of sound coding strategies using Cochlear implant simulation with Mandarin speech. *IEEE Transactions on Neural Systems and RehabilitationEngineering*,*29*,24072416.https://doi.org/10.1109/tnsre.2021.3128064

Qu,Leyuan, et al. "Disentangling Prosody Representations with Unsupervised Speech Reconstruction." *IEEE/ACM Transactions on Audio,Speech,andLanguageProcessing*,vol.32,1Jan.2024,pp.3954,https://doi.org/10.1109/taslp.2023.3320864.

Shahamiri, S. R. (2021). Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System. *IEEETransactionsonNeuralSystemsandRehabilitationEngineering*,*29*,852861.https://doi.org/10.1109/tnsre.2021.3076778

Geon Woo Lee, et al. "Knowledge Distillation-Based Training of Speech Enhancement for Noise-Robust Automatic Speech Recognition." *IEEEAccess*,vol.12,no.72707,1Jan.2024,pp.11,https://doi.org/10.1109/access.2024.3403761

Oruh, Jane, et al. "Long Short-Term Memory Recurrent Neural NetworkforAutomaticSpeech Recognition." *IEEEAccess*, vol.10,no.30069,2022,pp.11,https://doi.org/10.1109/access.2022.3159339.

Zhu, Han, et al. "Alternative Pseudo-Labeling for Semi-Supervised Automatic Speech Recognition." *IEEE/ACM Transactions on Audio Speechand Language Processing*, vol.31,no.3320,1Jan.2023,pp.33203330,https://doi.org/10.1109/taslp.2023.3306709

Miao, Haoran, et al. "Online Hybrid CTC/Attention End-To End AutomaticSpeechRecognitionArchitecture." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 1452, 2020, pp. 1–1, https://doi.org/10.1109/taslp.2020.2987752.

"Rahdar, Amir, et al. "Serial Weakening of Human-Based Attributes Regarding Their Effect on Content-Based Speech Recognition." *IEEE Access*,vol.11, no. 24394, 2023, pp. 24394–24406, https://doi.org/10.1109/access.2023.3255982.

Tran, Van-Thuan, and Wei-Ho Tsai. "Speaker Identification in Multi-Talker Overlapping Speech Using Neural Networks." *IEEE Access*, vol.8,no.134868,2020,pp.134868134879,https://doi.org/10.1109/access.2020.3009987.

Moriya, Takafumi, et al. "Streaming End-To-End Target-Speaker Automatic Speech Recognition and Activity Detection." *IEEE Access*, vol.11,no.13906,2023,pp.11,https://doi.org/10.1109/access.2023.3243690.

Chen, Hang, et al. "Optimizing Audio-Visual Speech Enhancement Using Multi-Level Distortion Measures for Audio-Visual Speech Recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, no. 2508, 1 Jan. 2024, pp. 1–14, https://doi.org/10.1109/taslp.2024.3393732.

Shahamiri, S. R., Lal, V., & Shah, D. (2023). Dysarthric Speech Transformer: A Sequence-to-Sequence dysarthric Speech Recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*,*31*,34073416.https://doi.org/10.1109/tnsre.2023.3307020

Jin, Rong, et al. "Speaker Verification Based on Single Channel Speech Separation." *IEEE Access*, vol. 11, no. 112631, 1 Jan. 2023, pp. 112631–112638, https://doi.org/10.1109/access.2023.3287868.

Azim, M. A., Hussein, W., & Badr, N. L. (2023). Using Character-Level Sequence-to-Sequence model for word level text generation to enhance Arabic speech recognition. *IEEE Access*, *11*, 91173–91183. https://doi.org/10.1109/access.2023.3302257

Oliveira, J., & Praca, I. (2021). On the Usage of Pre-Trained Speech Recognition Deep Layers to Detect Emotions. IEEE Access, 9, 9699–9705. https://doi.org/10.1109/access.2021.3051083

Kashyap, B., Pathirana, P. N., Horne, M., Power, L., & Szmulewicz, D. J. (2023). Machine Learning-Based scoring System to predict the risk and severity of ataxic speech using different speech tasks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 4839– 4850. https://doi.org/10.1109/tnsre.2023.3334718

Gonzales, M. G., Corcoran, P., Harte, N., & Schukat, M. (2024). Joint Speech Text embeddings formulti task speech processing. *IEEEAccess*,1https://doi.org/10.1109/access.2024.3473743

Takashima, Y., Takashima, R., Takiguchi, T., & Ariki, Y. (2019). Knowledge transferability between the speech data of persons with dysarthria speaking different languages for dysarthric speechrecognition.*IEEEAccess*,*7*,164320–164326. https://doi.org/10.1109/access.2019.2951856

Yasin, I., Drga, V., Liu, F., Demosthenous, A., & Meddis, R. (2020). Optimizing speech recognition using a computational model of human hearing: effect of noise type and efferent time constants. *IEEE Access*, *8*, 56711–56719. https://doi.org/10.1109/access.2020.2981885

Zhou, J., Duan, Y., Zou, Y., Chang, Y., Wang, Y., C Lin, C. (2023). Speech2EEG: Leveraging pretrained speech model for EEG signal recognition.*IEEETransactionsonNeuralSystemsandRehabilitationEngineering*,*31*,2140–2153. https://doi.org/10.1109/tnsre.2023.3268751

Yasin, I., Drga, V., Liu, F., Demosthenous, A., & Meddis, R. (2020). Optimizing speech recognition using a computational model of human hearing: effect of noise type and efferent time constants. *IEEE Access*, *8*, 56711–56719. https://doi.org/

10.1109/access.2020.2981885