AUTOMATIC SPEECH RECOGNITION USING DEEP NEURAL NETWORKS

1st Shivam Kushwaha Computer Science Engineering Chandigarh UniversityMohali, India
kshivam1002@gmail.com
2nd Piyush Deep Computer Science Engineering Chandigarh UniversityMohali, India
piyush.deep02@gmail.com
3rd Mohd Muaz Computer Science Engineering Chandigarh UniversityMohali, India
mmuaz1721@gmail.com
4th Er. Shafalii Sharma Computer Science Engineering Chandigarh University Mohali, India
hafalii.e13752@cumail.in

Abstract - This research paper revolves around the evolution and the present scenario of Automatic Speech Recognition systems using Deep Neural Networks. It includes the designs, techniques of training, the evaluations of the model performance, and the emerging trends that are specific to deep neural networks embedded in automatic speech recognition models. This research also incorporates the challenges faced while building and deploying the speech recognition model including limited data availability and adaptability. We have examined how deep neural networks have transformed automatic speech recognition and this research provides valuable insights to improve the speech recognition technology across a number of applications from healthcare to smart devices.

INDEX TERMS - Automatic Speech Recognition, Deep Neural Networks, Language Modeling, Robustness to Noise, Speech Modelling

I. INTRODUCTION

A. Significance of the model

The Automatic Speech Recognition model plays a vital role in converting spoken language into text. This also enables the seamless interaction between humans and machines. This has evolved the communication system by empowering the devices which are voice-controlled, translation the languages, and accessibility tools for the hearing impaired. ASR models have a large number of applications including healthcare, entertainment,

education, enhancing productivity and customer services.

B. Objectives of the research

The primary aim of the research in Automatic Speech Recognition with Deep Neural Networks is to improve the precision, effectiveness, and reliability of the speech recognition models. The creation of deep neural network structures helps to accurately grasp complex speech patterns, contexts, noise avoidance, and nuances. This allows the speech recognition models to understand different languages, accents and environmental factors with higher accuracy. This boosts the performance of the model through innovative neural network designs and data augmentation approaches.

II. LITERATURE REVIEW

There have been a number of researches on automatic speech recognition using deep neural networks which have resulted in

significant advancements in communication and human-machine interaction. A number of literature reviews have been reviewed before conducting this research on automatic speech recognition using DNNs. We have reviewed the evolutions, methodologies, challenges and, future directions in this specific field. The history of ASR has shown a significant shift from rule-based models to statistical models and adoption of the neural networks. Deep Neural Networks have the ability to model complex patterns that have emerged as of great significance in ASR research. They can handle large datasets and

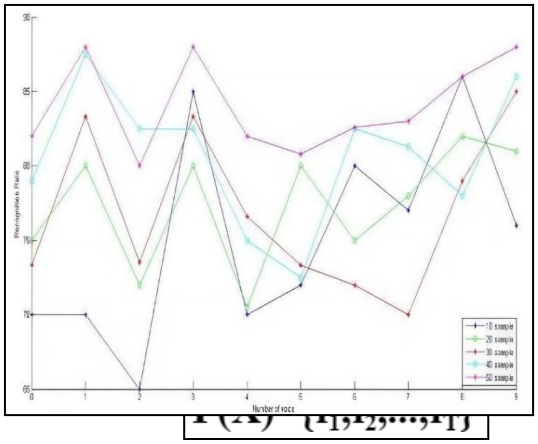
their hierarchical representations.

The model accuracy also remains vital including the word error rate and character error rate. An end-to-end framework has been introduced for automatic speech recognition, particularly recurrent neural networks and attention mechanisms. The LAS model also served as a source of inspiration for speech recognition models. The revolution in natural language processing and diverse fields also showed great contributions to machine translations. The ASR models have been enhanced with the help of convolutional neural networks. The main objectives also lie in presenting an ASR model that integrates acoustic modelling and language modelling into a single recurrent neural network architecture.

III. CHARACTERSTICS OF ASR

Automatic speech recognition models have three main dimensions that can help to characterize them. This involves dependence, vocabulary semantics and speech continuity. The speech recognition models can be speaker-dependent in which the system has to be trainedfor every single speaker or can be speaker-independent in which the training database contains a number of speech examples from different speakers which helps the system to recognize the new speaker. According to the speech continuity, there are four types of systems.

Figure: Structure of ASR system



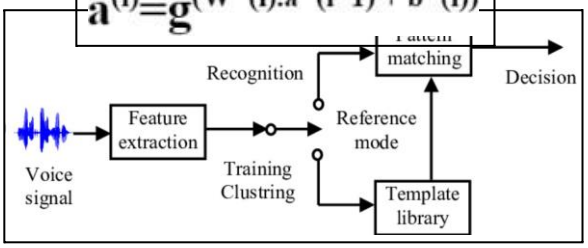
This includes isolated word recognition systems, connected word recognition systems, continuous speech recognition systems, and word spotting systems. The vocabularies are used to train the models. These vocabularies can be small or large-sized vocabularies. There are a number of errors that can occur in the automatic speech recognition models such as insertion errors which occur when the system perceives noise as a speech unit, substitution errors which occur when the recognizer incorrectly identifies an utterance, deletion errors that can occur when the model ignores an utterance. The errors can be direct, intent, and indirect.

IV. RESEARCH METHODOLOGIES

A mathematical model for a DNN-based Automatic Speech Recognition system is crucial to demonstrate the working of the model. ASR systems consist of three components - an acoustic model, a language model, and a decoding algorithm. Here, we have illustrated the mathematical model incorporating all the steps necessary for the working of the machine learning model.

A. Input Representation:

$$a^{(l)}=g(W^{(l)}.a^{(l-1)} + b^{(l)})$$



Let X,

be the input speech signal, where x_t is the feature vector at time t.

B. Feature Extraction:

Extract the features from the raw signal,e.g., using Mel-Frequency Cepstral Coefficients.
Let be the feature sequence.

C. Neural Network Architecture:

Define a deep neural network with Llayers. Let,

$W^{(l)}$ and $B^{(l)}$

represent the weights and biases of layer l, respectively.

The output of layer l is given by

where g is the activation function.

D. Input Layer:

The feature sequence is taken as theinput layer **F(X)** as its input.

E. Output Layer:

The output layer produces posteriorprobabilities for each phoneme or subword unit.
If there are N units, the output is

where y_i is the probability of unit i.

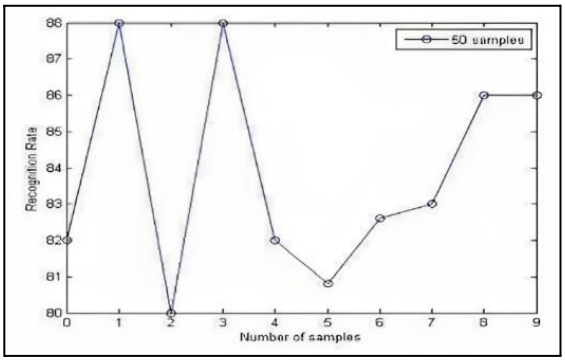
F. Training:

Define a training dataset.

Minimize the cross-entropy loss function:

where $y^{(i)}$ is the predicted probability.

$$Y = \{y_1, y_2, \ldots, y_N\}$$



$$\{(X^{(1)},Y^{(1)}),(X^{(2)},Y^{(2)}),\ldots,(X^{(m)},Y^{(m)})\}$$

$$J(W,b)=(-1/m)\ (\sum(m)(i=1))\ (\sum(N)(j=1))\ (y_j^{(i)})\ (\log(y_j^{(i)}))$$

G. Inference:

Given a new input sequence X, the ASR system predicts the sequence of units using the trained neural network.

H. Decoding:

A decoding algorithm is used to map the sequence of predicted units to the final recognized text.

V. APPLICATIONS OF ASR

A. Voice assistants:

The assistants use automatic speech recognition technology to translate spoken language into text. This helps to interact with the machines with the voice commands. The deep neural network here helps to improve voice recognition accuracy and understanding of the natural networks. They produce realistic speech using the DNN networks which enhances the user experience.

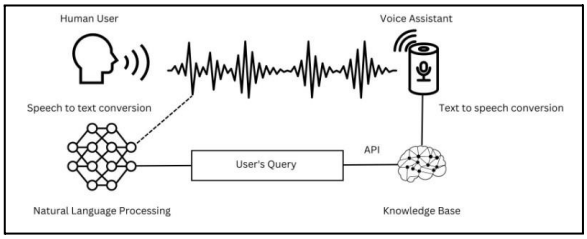


Figure: Working of voice assistants

B. Call centres:

Automatic speech recognition is also used in call centres to improve customer service and the efficiency of the operation. It makes the transcription of conversations between the customers and the agents automatic facilitating the real-time services.

C. Language translations:

ASR model converts the spoken language into text and deep neural network architectures improve this text for translation. Machine translation models are used to translate the speech into different languages. The synergy between ASR and DNNs helps to translate in real time.

D. Security and Authentication:

Speech recognition and deep neural networks play a vital role in voice-based authentication. ASR verifies the user's spoken language and DNN verifies the voice patterns for the biometric authentication. This combination helps to verify the user in applications, for example, phone locks, home-control systems, and voice assistants.

VI. CHALLENGES FACED

A. Data scarcity:

The data that is available for automatic speech recognition model is limited which creates a problem as it lacks the training data and the diversity. This affects the precision of the system in accurately converting the speech to the text which includes different accents, languages, and speaking styles which ultimately declines the performance of the speech recognition models.

B. Robustness to noise:

In speech processing, robustness to the noise means the system is capable of sustaining the performance with accuracy instead of the presence of background noises or disturbances. This incorporates the methods that can be utilized such as noise reduction and resilient extraction of features to improve the precision and accuracy of speech recognition in noisy settings.

C. Model complexity:

The size and complexity of the neural network architecture networks result in the complexity and advancements in model designing and deployment. The more complex the model, the more it includes a number of parameters and detailed aspects of the model. They may require larger amounts of data and computational resources.

D. Scalability:

Expanding the automatic speech recognition systems using deep neural networks may face some hurdles when it is applied across hardware and managing increased data sizes. Problems may include the optimization requirements, distributed computing resources, and maintaining real-time efficiency.

VII. RESULTS OF THE RESEARCH

The research in the field of automatic speech recognition has shown significant advancements in the accuracy and robustness of the DNN-based ASR model. The DNN-based ASR model has resulted in word error rate and character rate which made it

outperform the traditional systems. It has also shown robustness to various background noises and other environmental factors.

Figure 7.1 The STT demonstration

The deep neural network architectures have been optimized. They can handle large datasets and complex tasks which improves the scalability of the system. The enhancements have also been made in the DNN architectures and methodologies. DNN-based ASR models have seen a number of applications in the industry such as healthcare, education, and smart devices.

Although a number of improvements have been made, still achieving robustness to all types of noise and environmental factors is an ongoing focus area. The advancements are still ongoing for supporting multilingual language models. Therefore, the research outcomes in automatic speech recognition models have shown remarkable progress in practical deployment, robustness, and improved accuracy.

VIII. FUTURE DIRECTIONS

There is still a lot to explore in the field of automatic speech recognition using deep neural networks. Data augmentation is very important in scenarios of low-resource ASR where there is a limited data set for training the model. Techniques like adding noise, altering the speed, or generating synthetic data are very helpful for the ASR models.

Furthermore, robustness and noise handling is very important. Robust ASR models can transcribe speech even in noisy environments with accuracy. Noise reduction techniques and acoustic modelling help to develop more reliable models. As we get deeper into the work of neural networks, it will help to improve the customization and personalization of the ASR.

IX. CONCLUSION

To conclude, this research paper has provided an overall exploration of deep neural network-based automatic speech recognition systems. It mentions the shift of technologies from the traditional ASR methodologies to the vital role played by the DNN-based automatic speech recognition model. This has revolutionized the field of speech recognition. Dissection of the architectural paradigms of the speech recognition model and the training strategies have shown remarkable improvements in the accuracy and versatility of the ASR model. This research also has a number of challenges including scarcity of data, noise resilience, and speaker variability. The ASR continues to evolve making the future of ASR systems promising. It not only includes the present state of the field but also the progressand future innovations.

REFERENCES

[1] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G. and Chen, J., 2016, June. Deep speech 2: End-to-end speech recognition in English and Mandarin. In International conference on machine learning (pp. 173-182). PMLR

[2] Chan, W., Jaitly, N., Le, Q. and Vinyals, O., 2016, March. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In 2016 IEEE international conference on acoustics, speech and signalprocessing (ICASSP) (pp. 4960-4964). IEEE.

[3] Cui, X., Goel, V. and Kingsbury, B., 2015. Data augmentation for deep neural network acoustic modelling. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(9), pp.1469-1477.

[4] Du, J., Wang, Q., Gao, T., Xu, Y., Dai, L.R. and Lee, C.H., 2014. Robust speech recognition with speech-enhanced deep neural networks. In the Fifteenth annual conference of the International Speech Communication Association.

[5] Espana-Bonet, Cristina, and José AR Fonollosa. "Automatic speech recognition with deep neural networks for impaired speech." In Advances in Speech and Language Technologies for Iberian Languages: Third International Conference, IberSPEECH 2016, Lisbon, Portugal, November 23-25, 2016, Proceedings 3, pp. 97-107. Springer
International Publishing, 2016.

[6] Fantaye, T.G., Yu, J. and Hailu, T.T., 2020. Investigation of automatic speech recognition systems via the multilingual deep neural network modelling methods for a very low-resource language, Chaha. Journal of Signal and Information Processing, 11(1), pp.1-21.

[7] Fendji, J.L.K.E., Tala, D.C., Yenke, B.O. and Atemkeng, M., 2022. Automatic speech recognition using limited vocabulary: A survey. Applied Artificial Intelligence, 36(1), p.2095039.

[8] Gulati, A., Qin, J., Chiu, C.C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. and Pang, R., 2020. Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100.

[9] Han, K., He, Y., Bagchi, D., Fosler-Lussier, E. and Wang, D., 2015. Deep neural network-based spectral feature mapping for robust speech recognition. At the Sixteenth annual conference of the International Speech Communication Association.

[10] Iosifova, O., Iosifov, I., Sokolov, V.Y., Romanovskyi, O. and Sukaylo, I., 2021. Analysis of automatic speech recognition methods. Cybersecurity Providing in Information and Telecommunication Systems, 2923, pp.252-257.

[11] Mukhamadiyev, A., Khujayarov, I., Djuraev, O. and Cho, J., 2022. Automatic speech recognition method based on deep learning approaches for Uzbek language. Sensors, 22(10), p.3683.

[12] Nassif, A.B., Shahin, I., Attili, I., Azzeh, M. and Shaalan, K., 2019. Speech recognition using deep neural networks: A systematic review. IEEE Access, 7, pp.19143-19165.

[13] Palaz, D. and Collobert, R., 2015. Analysis of CNN-based speech recognition system using raw speech as input (No. REP_WORK). Idiap.

[14] Pardede, H.F., Yuliani, A.R. and Sustika, R., 2018. Convolutional neural network and feature transformation for distant speech recognition. International Journal of Electrical and Computer Engineering, 8(6), p.5381.

[15] Qian, Y., Bi, M., Tan, T. and Yu, K., 2016. Very deep convolutional neural networks for noise robust speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(12), pp.2263-2276.

[16] Sarma, M., 2017. Speech recognition using deep neural network trends. International Journal of Intelligent Systems Design and Computing, 1(1-2), pp.71-86.

[17] Sim, K.C., Qian, Y., Mantena, G., Samarakoon, L., Kundu, S. and Tan, T., 2017. Adaptation of deep neural network acoustic models for robust automatic speech recognition. New Era for Robust Speech Recognition: Exploiting Deep Learning, pp.219-243.

[18] Serizel, R. and Giuliani, D., 2014. Deep neural network adaptation for children's and adults' speech recognition. Deep neural network adaptation for children's and adults' speechrecognition, pp.344-348.

[19] Soundarya, M., Karthikeyan, P.R. and Thangarasu, G., 2023, March. Automatic Speech Recognition trained with Convolutional Neural Network and predicted with Recurrent Neural Network. In 2023 9th International Conference on Electrical Energy Systems (ICEES) (pp. 41-45). IEEE.

[20] Toledano, D.T., Fernández-Gallego, M.P. and Lozano-Diez, A., 2018. Multi-resolution speech analysis for automatic speech recognition using deep neural networks: Experiments on TIMIT. PloS one, 13(10), p.e0205355.

[21] Tong, S., Garner, P.N. and Bourlard, H., 2017. An investigation of deep neural networks for multilingual speech recognition training and adaptation (No. CONF, pp. 714-718).

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

[23] Weng, C., Yu, D., Seltzer, M.L. and Droppo, J., 2015. Deep neural networks for single-channel multi-talker speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(10), pp.1670-1679.

[24] Yao, Kaisheng, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong. "Adaptation of context-dependent deep neural networks for automatic speech recognition." In 2012 IEEE Spoken Language Technology Workshop (SLT),pp. 366-369. IEEE, 2012.

[25] Yu, D., Siniscalchi, S.M., Deng, L. and Lee, C.H., 2012, March. Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4169-4172). IEEE.