International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

AUTOMATIC SPEECH RECOGNITION USING DEEP NEURAL NETWORKS

1st Shiwani Singhal
Computer Science Engineering
Chandigarh University
Mohali, Punjab
shiwanisinghal212@gmail.com

2nd Muskan Deswal
Computer Science Engineering
Chandigarh University
Mohali, Punjab
deswalmuskan2002@gmail.com

3rd Er. Shafalii Sharma
Computer Science Engineering
Chandigarh University
Mohali, Punjab
shafalii.e13752@cumail.in

Abstract - This research concentrates on the progression and present status of automatic speech recognition systems powered by deep neural networks. It discusses model architectures, training approaches, evaluating model efficacy, and recent advancements specific to deep neural networks applied in automatic speech recognition models. It considers the challenges faced in crafting these speech recognition models, such as data scarcity and the necessity for adaptability. Our exploration traces the evolution of automatic speech recognition through deep neural networks, presenting valuable insights aimed at propelling the domain of speech recognition for diverse applications, spanning from smart devices to healthcare.

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

I. INTRODUCTION

The Automatic Speech Recognition model plays a pivotal role in converting spoken language to text and enabling seamless interactions between humans and machines. This evolution has revolutionized communication systems by empowering voice-controlled devices, language translation, and accessibility tools for those with hearing impairments. ASR models have extensive applications across various sectors, including healthcare, entertainment, and many more.

The core objective of research in Automatic Speech Recognition utilizing Deep Neural Networks is to enhance the accuracy, efficiency, and dependability of speech recognition models. Constructing deep neural network architectures aids in accurately capturing intricate speech patterns, contextual cues, noise reduction, and subtleties.

II. LITERATURE REVIEW

Several research initiatives investigating automatic speech recognition through deep neural networks have notably progressed communication and human-machine interaction. Prior to commencing this ASR study utilizing DNNs, extensive reviews of the literature were conducted, encompassing the developments, methodologies, challenges, and future trajectories within this domain. The evolution of ASR has seen a significant transition from rule-based models to statistical methods and the integration of neural networks. Ensuring the accuracy of models, gauged through word and character error rates, remains a critical focus. An end-to-end methodology utilizing recurrent neural networks and attention mechanisms has

has been introduced for automatic speech recognition. The LAS model has significantly influenced the evolution of speech recognition models. Progress in natural language processing and diverse fields has substantially contributed to advancements in machine translations.

III. CHARACTERSTICS OF ASR

Automatic speech recognition models have three main dimensions for characterization: dependence, vocabulary semantics, and speech continuity. They can either be speaker-dependent, necessitating training for each speaker or speaker-independent, using various speech examples to recognize new speakers. In terms of speech continuity, there are four different types of systems.

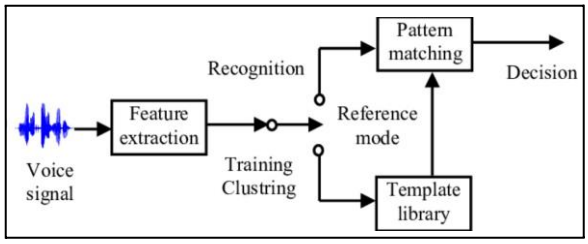


Figure: Structure of ASR system

This comprises isolated word recognition systems, connected word recognition systems, continuous speech recognition systems, and word spotting systems. Automatic speech recognition models can face different types of errors, such as insertion errors, substitution errors, and deletion errors.

IV. RESEARCH METHODOLOGIES

Creating a mathematical model for an Automatic Speech Recognition (ASR) system using Deep Neural Networks (DNNs) involves fundamental components. ASR systems typically consist of three key parts: an acoustic model, a language model, and a decoding algorithm.

A. Input Representation:

Let X,

$$X=\{x_1,x_2,\ldots,x_T\}$$

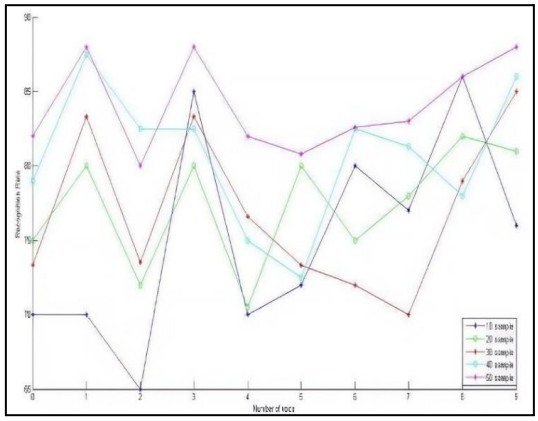
be the input speech signal, where x_t is the feature vector at time t.

B. Feature Extraction:

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

$$F(X)=\{f_1,f_2,\ldots,f_T\}$$

be the feature sequence.



C. Neural Network Architecture:

Define a deep neural network with L layers. Let,

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

represent the weights and biases of layer l, respectively.

The output of layer l is given by

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

where, g is the activation function.

D. Input Layer:

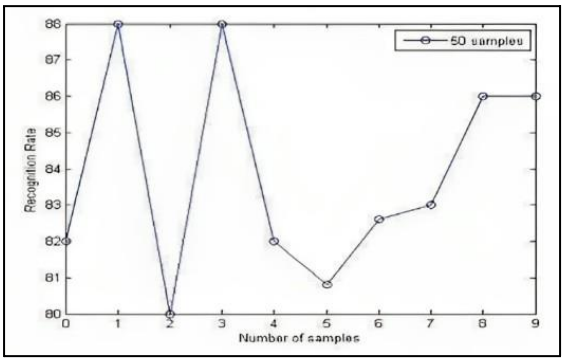
The feature sequence is taken as the input layer **F(X)** as its input.

E. Output Layer:

The output layer generates probabilities representing the likelihood of each phoneme or subword unit, known as posterior probabilities. If there are N units, the output is

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

where y_i is the probability of unit i.



F. Training:

Define a training dataset

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

Minimize the cross-entropy loss function:

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

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

G. Inference:

When provided with a new input sequence X, the ASR system utilizes the trained neural network to predict the sequence of units.

H. Decoding:

A decoding algorithm is employed to convert the sequence of predicted units into the final recognized text.

V. APPLICATIONS OF ASR

A. Voice assistants:

Voice assistants employ automatic speech recognition to convert spoken language into text, enabling interaction with machines through voice commands. Deep neural networks play a crucial role in enhancing the accuracy of voice recognition and understanding natural speech patterns.

B. Call centres:

In call centres, automatic speech recognition is used to boost customer service quality and operational efficiency. It automates the transcription of conversations between customers and agents, enabling prompt service delivery.

C. Language translations:

The ASR model converts spoken language into text and utilizes deep neural network architectures to enhance the text for translation. Machine translation models are then applied to interpret speech in various languages. This cooperation between ASR and DNNs allows for immediate translation.

VI. CHALLENGES FACED

A. Data scarcity:

The restricted data accessible for automatic speech recognition models creates a challenge due to the lack of diverse training data. This limitation impacts the system's accuracy in precisely transcribing speech into text, especially when handling various accents, languages, and speaking styles.

B. Robustness to noise:

In speech processing, maintaining accuracy despite background noise or disruptions signifies the system's robustness. Achieving this involves employing methods like noise reduction and robust feature extraction to enhance the precision and accuracy of speech recognition in noisy environments.

C. Model complexity:

The size and complexity of neural network architectures play a role in the intricacy and evolution of model development and implementation. More intricate models often encompass a higher number of parameters and intricate components, potentially necessitating larger datasets and heightened computational resources.

VII. RESULTS OF THE RESEARCH

Progress in automatic speech recognition has significantly boosted the accuracy and robustness of DNN-driven ASR models. These models have showcased better word error rates and character error rates than conventional systems.

Figure 7.1 The STT demonstration

Enhancements in deep neural network structures have enabled their capacity to manage vast datasets and intricate tasks, thus improving the system's scalability. Furthermore, there have been advancements in DNN architectures and methodologies. These DNN-based ASR models find utility across sectors like healthcare, education, and smart devices.

VIII. FUTURE DIRECTIONS

In the domain of automatic speech recognition using deep neural networks, there's extensive exploration yet to be done. Data augmentation plays a vital role, especially in low-resource ASR situations characterized by limited training data. Techniques such as introducing noise, adjusting speed, or creating synthetic data significantly contribute to enhancing ASR models. It is crucial to underscore the significance of robustness and effective noise management within the realm of automatic speech recognition. A robust ASR model must possess the ability to precisely transcribe speech, even when confronted with noisy environments.

IX. CONCLUSION

In conclusion, this research paper thoroughly investigates automatic speech recognition systems utilizing deep neural networks. It explores the transition from traditional ASR approaches to the substantial impact of DNN-based models, marking a transformative shift in speech recognition. The analysis of architectural paradigms and training strategies highlights significant improvements in the accuracy and adaptability of ASR models. Despite these advancements, challenges like scarce data, noise resilience, and speaker variability are recognized.

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 signal processing (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' speech recognition, 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.