Applications of Deep Learning Approaches in Speech Recognition: A Survey

Sameer I. Ali Al-Janabi and Ali Azawii Abdul Lateef

Abstract Automated speech recognition (ASR) appeared to be a driving force for a variety of machine learning (ML) techniques, include to ubiquitously utilized discriminative learning, Bayesian learning, hidden Markov model, adaptive learning, and structured sequence learning. Although machine learning utilize ASR as a large scale, it can reasonable application to thoroughly test viability for a given procedure and to motivate unused issues emerging from intrinsically consecutive and discourse energetic nature. Also, although ASR is accessible commercially for a few applications used in this research through the limitation and research gaps that the researcher try to access high accuracy of these systems. The advance technology from new ML techniques appears incredible guarantee to progress the literature review in ASR innovation. This study gives reader with a diagram of present-day ML methods as used within the relevant and current as significant for ASR future systems and research. The study goal is to promote advanced cross-pollination between ML and ASR communities more than has hither to occurred.

Keywords Machine learning · Automated speech recognition · Speech identification · Speech to text

1 Introduction

Speech is the foremost characteristic and viable strategy of communication between human creatures. Speech identification aimed to decipher speech to text [1]. It may be a standard classification issue where discourse signals got to be mapped to or recognized as words. Therefore, it is not conceivable to work with discourse reports

S. I. Ali Al-Janabi (✉)
Collage of Islamic Science, University of Anbar, Anbar, Iraq
e-mail: isl.samir.ia2012@uoanbar.edu.iq

A. A. A. Lateef
Human Resources Department, University of Anbar, Anbar, Iraq
e-mail: Aliazawii@uoanbar.edu.iq

A. K. Bashir et al. (eds.), *Proceedings of International Conference on Computing and Communication Networks*, Lecture Notes in Networks and Systems 394,
https://doi.org/10.1007/978-981-19-0604-6_17

in case they are recorded as sound signals. Hence, discourse acknowledgment has gotten to be a vital zone of research [2, 3].

There are numerous challenges which make real-time speech recognition a difficult issue. Different possible pronunciations are a few of these challenges. There is a considerable misfortune in precision when we move from a controlled exploratory setup to genuine life circumstances. Despite this, automated speech recognition system has copious utilization in correspondence, human–machine interfacing and control of machines among others.

2 Speech Recognition Techniques

Recently, there has been a broad applications of deep learning approaches and neural networks to perform speech recognition leading to critical new outcomes. The slant started two decades back, when modern comes about were accomplished utilizing hybrid ANN-HMM schemes. These schemes appropriated the utilize of neural network (NN) with as it was one layer of covered up units having nonlinear actuation capacities to foresee probabilities over well states from brief windows of acoustic coefficients [4]. The Neural Network is effective approach that can speak to complex nonlinear capacities but at that time, not one or the other the computation control nor the preparing calculations that were accessible, were progressed sufficient for preparing NN with numerous covered up layers. So, cross-breed ANN-HMM schemes might not supplant the exceptionally fruitful combination of HMMs with acoustic models based on Gaussian mixtures.

Malla et al. proposed a system which can recognizing feeling within the discourse from the speech signals. This system done based on the most recent studies within speech emotion recognition (SER) field schemes using neural network convolutional related with the issue and give an ideal solution. The details of framework proposed are dataset, stage of extraction feature, and classification task that assist a help within the usage and assessing the framework. This framework will help the conclusion clients in emotion acknowledgment from discourse flag and making AI more vigorous by utilizing neural organize convolutional, encouraging a colossal nearness within the future system [5].

Dhande and Shaikh studied how the epochs playing a vital part within preparing databases. The epochs number chooses whether the information over trained or not. Results depend on database prepare. Speech recognition broadly utilized application in these days. Deep learning based on speech recognition has changed the viewpoint of the world to see at the innovation. The proposed system is based on speech recognition with deep learning approach where there are sound files and content transcripts within the datasets. The sound records are prepared with the acknowledgment show, and transcript contents are prepared by a dialect demonstrate. The dataset that is used in this architecture is made up of pieces of sounds taken from three diverse situation, to be specific clean, white clamor and persistent noise [6].

Tarunika et al. used *k*-nearest neighbor and deep neural network for recognition of emotion from speech particularly terrifying state of intellect. The field of applications of the framework is primarily concerned over the healthcare sectors. The establishment of this inquire has primary firm applications in field of palliative care. Beneath most exact result, the signals of caution are made by cloud. Numerous crude information is collected beneath extraordinary accentuation methods. After that the acoustic voice signals are changed over to wave shape, discourse level highlight extraction feeling classification, existing database acknowledgment, alarm flag creation through cloud is the grouping of steps to be followed [7].

Yousefi and Hansen proposed a block-based CNN design to address discourse covering modeling in streams sound with outlines as brief as 25 ms. The proposed engineering is strong for: (i) shifts in arrange enactments dispersion due to changing in arrange parameters amid preparing, (ii) nearby varieties from input highlights caused by extraction highlight, natural commotion, or room interference. Moreover, examine substitute input highlights counting ghostly greatness, MFCC, MFB, and pyknogram impact on both computational time and classification execution [8].

Tzirakis et al. presented a modern method for persistent emotion recognition from discourse. The proposed system comprised from a convolutional neural arrange (CNN), which extricates highlights from crude flag, and stacked on beat a two-layer long short-term memory (LSTM), to consider the relevant data within information. In terms of concordance relationship coefficient, our show essentially outflanks the state-of-the-art strategies for RECOLA database [9].

Arif and Puji developed the framework of existing speech recognition Indonesian that has a precision and still not great for unconstrained speech recognition. The framework is prepared utilizing HMM-GMM acoustic show. In this ponder, unconstrained discourse information collected in Indonesian for duration 14 h and discourse acknowledgment framework execution was progressed by supplanting acoustic show with a neural network-based demonstrate. The utilized neural networks topology are time delay neural network, deep neural network, and convolutional neural network [10].

Zakiah and Lestari propose advancement iterative acoustic models by using an extra unlabeled speech corpus. They used unlabeled information for revamp acoustic models by utilizing segment's translations created by already directed created ASR. For more urge solid translation, we utilized four ASRs with four sorts of profound learning-based acoustic models (CNN, TDNN, DNN, and LSTM) and chosen fragments with reliable transcripts given by models or fragments with completely understanding names [11].

Nugroho et al. discussed gender voice identification for Javanese individuals who are handled utilizing mel recurrence cepstral coefficient extracted features, at that voice classification point is done utilizing deep learning technique combined with singular value decomposition strategy in decreasing information delivered measurements. The dataset used for building this approach divided into two parts: the first part is 70% of dataset used for training model and the second part is the remaining 30% of the dataset used for testing model information the comes about of the inquire for appear profound learning method's precision is (97.78%) higher than calculated

relapse strategy (95.56%) and SVM (93.33%). Discourse acknowledgment investigate appears profound learning, and SVD strategy can be utilized for performing discourse acknowledgment with high precision degree 93.33% [12].

Agrawal and Ganapathy offer a deep variational model-based method for learning modulation filters. They formulate filter learning problem in a deep unsupervised generative modeling framework, in which variational autoencoder convolutional filters capture voice modulations significant. In combined spectro-temporal domain, the spectrogram properties for voice identification for process and train are used two-dimensional modulation filters and deep variational networks, respectively. Several voice recognition studies are carried out on a series of challenges that include reverberation (REVERB Challenge), noise addition with reverberation (REVERB Challenge) (CHiME-3), noise addition with artifacts channel (Aurora-4). The modulation filter learning framework beats baseline properties and a range of current noise-resistant front ends in these suggested tests (average relative improvements of over the baseline features 7.5 and 20% in Aurora-4 and CHiME-3 databases, respectively). In addition, the proposed method has been demonstrated to be beneficial in semi-supervised automatic voice recognition systems. By employing 30% labeled training data, for example, on the Aurora-4 database, a relative improvement of 25% over the baseline system was discovered [13].

The metaheuristic algorithm pigeon inspired optimization (PIO) technique was introduced by Waris and Aggarwal used to optimize weight matrix for DNN model. This heuristic method is used to optimize the weight matrix. DNN training time is reduced because of this, and the system's recognition rate improves. The weight matrix optimization result is tested on phoneme recognition TIMIT database [14].

Liu et al. offer a two-module deep representation learning system that is local–global aware. To learn local representation, for example, time frequency CNN (TFCNN) is one module includes a multi-scale CNN. Framework with dense connections for several blocks is another module to learn deep and shallow global knowledge. Each block in this structure is a fully functional CapsNet that has been enhanced by a new routing algorithm [15].

A convolutional neural network (CNN) architecture is proposed by Saheaw et al. In order to compare it with long short-term memory (LSTM), the Thai language speech dataset turn-on and off by seven types from electrical applications. The process of reducing noise and silence from the front and the back audio is completed by 14 classes. According to tests findings, the proposed long short-term memory has best accuracy [16].

Han et al. studied and quickly explained the principles and categories, methodologies, and applications of transfer learning, as well as the application of speech emotion identification, before noting the important areas that require more investigation [17].

Table 1 shows the summary of most work in literature review with the used methods and its achievements.

Table 1 Summary of deep learning approaches applied in speech recognition

Authors	Approach	Year
Malla et al. [5]	CNN model which can recognizing feeling within the discourse from the speech signals. Typically done based on the most recent studies within speech emotion recognition (SER) field schemes using neural network convolutional related with the issue and give an ideal solution	2020
Dhande and Shaikh [6]	Studied how the epochs playing a vital part within preparing databases. The epochs number chooses whether the information over trained or not. Results depend on database prepare	2019
Tarunika et al. [7]	Used *k*-nearest neighbor and deep neural network for recognition of emotion from speech particularly terrifying state of intellect. The field of applications of the framework is primarily concerned over the healthcare sectors	2018
Yousefi and Hansen [8]	Proposed a block-based CNN design to address discourse covering modeling in streams sound with outlines as brief as 25 ms. The proposed engineering is strong for: (i) shifts in arrange enactments dispersion due to changing in arrange parameters amid preparing, (ii) nearby varieties from input highlights caused by extraction highlight, natural commotion, or room interference	2021
Tzirakis et al. [9]	Presented a modern method for persistent emotion recognition from discourse. The proposed system comprised from a convolutional neural arrange (CNN), which extricates highlights from crude flag, and stacked on beat a two-layer long short-term memory (LSTM), to consider the relevant data within information	2018
Arif and Puji [10]	Developed the framework of existing speech recognition Indonesian that has a precision and still not great for unconstrained speech recognition. The framework is prepared utilizing HMM-GMM acoustic show	2020
Zakiah and Lestari [11]	Propose advancement iterative acoustic models by using an extra unlabeled speech corpus. They used unlabeled information for revamp acoustic models by utilizing segment's translations created by already directed created ASR	2020
Nugroho et al. [12]	Discussed gender voice identification for Javanese individuals who are handled utilizing mel recurrence cepstral coefficient extracted features, at that voice classification point is done utilizing deep learning technique combined with singular value decomposition strategy in decreasing information delivered measurements	2019

(continued)

Table 1 (continued)

Authors	Approach	Year
Agrawal and Ganapathy [13]	Offer a deep variational model-based method for learning modulation filters. They formulate filter learning problem in a deep unsupervised generative modeling framework, which variational autoencoder convolutional filters capture voice modulations significant	2019
Waris and Aggarwal [14]	Metaheuristic algorithm pigeon inspired optimization (PIO) technique that used to optimize weight matrix for DNN model. This heuristic method uses to optimize the weight matrix	2018
Liu et al. [15]	Offer a two-module deep representation learning system that is local–global aware. To learn local representation, for example, time frequency CNN (TFCNN) is one module includes a multi-scale CNN	2020
Saheaw et al. [16]	The Thai language speech dataset turn-on and off by seven types from electrical applications using a CNN architecture contrasted to LSTM	2020
Han et al. [17]	Study and quickly explain the principles and categories, methodologies, and applications of transfer learning, as well as the application of speech emotion identification	2019

3 Challenges

High reliability and stable detection are still difficult to achieve due to the intricacy of speech recognition system. The following are the key reasons: (1) The context of speech, such as the speaking scene, the speaker's manner of speaking, and the speaker's age, gender, and speaking behaviors, all influence human audio generation. (2) Speech data gathering is difficult, and it must account for ambient noise. (3) Emotion is a personal experience, and there is no proper statement of emotion. (4) The capacity of the human that defines the data to perceive emotion has an impact on the annotation of the emotion data. Annotation is time-consuming since it depends on the whole display of speech information. As a result, the lot of public speaking emotion corpora that have been annotated is restricted.

4 Conclusions and Future Works

The field of deep learning has seen quick advance and lead to critical enhancements in different areas. In this survey, we have given a brief instructional exercise and outline of deep learning procedures and models within the domain of speech recognition. Recently, acoustic models based on CNNs and DBNs have effectively supplanted Gaussian blends and have been illustrated to work very well for expansive lexicon assignments. Additionally, there has been the thought of killing preparing stages,

utilizing one unified neural organize to attain end-to-end discourse acknowledgment. To this conclusion, RNNs are presently being tested with but require much computation control for preparing. The utilization of RNNs for acoustic system inside a hybrid DNN-HMM framework as compared to the utilization of RNNs for end-to-end speech recognition utilizing CTC misfortune work and a dialect demonstrate has had blended responses. Deep learning holds the control to work with crude inputs and learn wealthy representations whereas disposing of difficult handling stages. With quick progression of computational advances, deep learning will as it was developed within the future.

In the future, greater datasets will be used to test deeper CNN models for speech analysis. We believe that using the raw signal, we can achieve superior results for various speech analysis tasks. When creating a new model, however, we must stick to the core principles of kernel size and pooling size.

References

1. A. Kumar, S. Verma, H. Mangla, A survey of deep learning techniques in speech recognition, in *Proceedings of IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2018* (2018), pp. 179–185
2. E. Trentin, M. Gori, A survey of hybrid ANN/HMM models for automatic speech recognition. Neurocomputing **37**(1–4), 91–126 (2001)
3. L. Deng, X. Li, Machine learning paradigms for speech recognition: an overview. IEEE Trans. Audio Speech Lang. Process. **21**(5), 1060–1089 (2013)
4. H. Bourlard, N. Morgan, *Connectionist Speech Recognition—A Hybrid Approach* (1994)
5. S. Malla, A. Alsadoon, S.K. Bajaj, A DFC taxonomy of speech emotion recognition based on convolutional neural network from speech signal, in *CITISIA 2020—IEEE Conference on Innovative Technologies in Intelligent Systems and Industrial Applications, Proceedings* (2020)
6. G. Dhande, Z. Shaikh, Analysis of epochs in environment based neural networks speech recognition system, in *Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019*, no. Icoei (2019), pp. 605–608
7. K. Tarunika, R.B. Pradeeba, P. Aruna, Applying machine learning techniques for speech emotion recognition, in *2018 9th International Conference on Computing, Communication and Networking Technologies* (2018), pp. 1–5
8. M. Yousefi, J.H.L. Hansen, Block-based high performance CNN architectures for frame-level overlapping speech detection. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 28–40 (2021)
9. P. Tzirakis, J. Zhang, W. Schuller, *End-to-End Speech Emotion Recognition Using Deep Neural Networks* (Department of Computing, Imperial College London, London, UK Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany, 2018), pp. 5089–5093
10. D.A. Rahman, Indonesian spontaneous speech recognition system using deep neural networks (2020), pp. 2020–2022
11. I. Zakiah, D.P. Lestari, Iterative deep learning-based acoustic models using transcription agreement from multi-models automatic speech recognitions, in *2020 7th International Conference on Advanced Informatics: Concepts, Theory and Applications. ICAICTA 2020* (2020), pp. 1–5
12. K. Nugroho, E. Noersasongko, Purwanto, Muljono, H.A. Santoso, Javanese gender speech recognition using deep learning and singular value decomposition, in *Proceedings—2019 International Seminar on Application for Technology of Information and Communication: Industry 4.0: Retrospect, Prospect, and Challenges, iSemantic 2019* (2019), pp. 251–254

13. P. Agrawal, S. Ganapathy, Modulation filter learning using deep variational networks for robust speech recognition. IEEE J. Sel. Top. Signal Process. **13**(2), 244–253 (2019)
14. A. Waris, R.K. Aggarwal, Optimization of deep neural network for automatic speech recognition, in *Proceedings of the International Conference on Inventive Research in Computing Application. ICIRCA 2018*, no. Icirca (2018), pp. 524–527
15. J. Liu, Z. Liu, L. Wang, L. Guo, J. Dang, Speech emotion recognition with local-global aware deep representation learning, in *ICASSP, IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, May 2020, vol. 2020 (2020), pp. 7174–7178
16. W. Saheaw, S. Jaiyen, A. Hanskunatai, Thai voice recognition for controlling electrical appliances using long short-term memory, in *2020 IEEE 7th International Conference on Industrial Engineering and Applications. ICIEA 2020* (2020), pp. 697–700
17. Z. Han, H. Zhao, R. Wang, Transfer learning for speech emotion recognition, in *Proceedings of 5th IEEE International Conference on Big Data Security on Cloud, BigDataSecurity 2019, 5th IEEE International Conference on High Performance and Smart Computing, HPSC 2019, 4th IEEE International Conference on Intelligent Data and Security, IDS 2019* (2019), pp. 96–99