Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 169-176, Nisan 2020

Derleme Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 169-176, April 2020

Review Article

Ses Tanıma için Derin Öğrenme Mimarileri Üzerine Derleme*

Yeşim Dokuz [1, †], Zekeriya Tüfekçi [2]

[1] Computer Engineering Department, Nigde Omer Halisdemir University, Nigde, Turkey (ORCID: 0000-0001-7202-2899)
[2] Computer Engineering Department, Cukurova University, Adana, Turkey (ORCID: 0000-0001-7835-2741)

(Conference Date: 5-7 March 2020)

(**DOI:** 10.31590/ejosat.araconf22)

ATIF/REFERENCE: Dokuz, Y., & Tüfekçi, Z. (2020). A Review on Deep Learning Architectures for Speech Recognition. *European Journal of Science and Technology*, (Özel Sayı), 169-176.

Öz

Derin öğrenme, çeşitli algoritmalar kullanarak çok sayıda işlem katmanından oluşan derin mimariler yardımıyla veri kümelerinin modelini çıkarmaya çalışan makine öğrenmesi alanının bir alt alanıdır. Derin öğrenme mimarilerinin başarılı uygulamaları ve popülerliğinden dolayı, derin öğrenme sistemleri ses tanıma alanında da kullanılmaya başlanmıştır. Araştırmacılar bu mimarileri ses tanıma ve ses tanımanın uygulamalarında, örneğin ses duygu tanıma, ses etkinliği tespiti ve konuşmacı tanıma ve doğrulama, ses girdileri ve çıktıları arasındaki modellerin daha iyi kurulması ve ses tanıma sistemlerinin hata oranlarının düşürülmesi amaçlarıyla kullanmışlardır. Literatürde, ses tanıma sistemleri için derin öğrenme mimarilerini kullanan çok sayıda çalışma yapılmıştır. Literatürde yapılmış olan çalışmalar ses tanıma ve uygulamaları için derin öğrenme mimarilerinin kullanılmasının pek çok ses tanıma alanın için fayda sağladığını ve hata oranlarını düşürerek daha iyi performans elde edilmesini sağladığını göstermiştir. Bu çalışmada, ilk olarak, ses tanıma probleminden ve ses tanıma adımlarından bahsedilmiştir. Daha sonra, derin öğrenme tabanlı ses tanıma için yapılmış olan çalışmalar incelenmiştir. Özellikle, derin öğrenme mimarilerinden olan Derin Sinir Ağları (DSA), Evrişimli Sinir Ağları (ESA) ve Özyinelemeli Sinir Ağları (ÖSA) ve bu mimarilerden üretilmiş olan hibrit yaklaşımlar değerlendirilmiş ve bu mimarilerin ses tanıma ve ses tanımanın uygulama alanlarındaki kullanımları ile ilgili literatürdeki çalışmalar değerlendirilmiştir. Sonuç olarak, hata oranları ve ses tanıma performansı açısından tüm mimariler arasında en yaygın olarak kullanılan ve en güçlü derin öğrenme mimarisinin ÖSA olduğu gözlemlenmiştir. ESA ise diğer bir başarılı derin öğrenme mimarisidir ve ses tanıma performansı ve hata oranları açısından ÖSA ile yakın sonuçlar üretmektedir. Ayrıca, hibrit derin öğrenme mimarilerinin de gittikçe yaygın hale geldiği ve ses tanıma hata oranlarını düşürebildiği gözlemlenmiştir.

Anahtar Kelimeler: Ses tanıma, Derin Öğrenme, DSA, ESA, ÖSA, Hibrit mimariler

A Review on Deep Learning Architectures for Speech Recognition

Abstract

Deep learning is a branch of machine learning that uses several algorithms which tries to model datasets by using deep architectures with many processing layers. With the popularity and successful applications of deep learning architectures, they are being used in speech recognition, as well. Researchers utilized these architectures for speech recognition and its applications, such as speech emotion recognition, voice activity detection, and speaker recognition and verification to better model speech inputs with outputs and to reduce error rates of speech recognition systems. Many studies are performed in the literature that use deep learning architectures for speech recognition systems. The literature studies show that using deep learning architectures for speech recognition and its applications provide benefits for many speech recognition areas and have ability to reduce error rates and provide better performance.

* This paper was presented at the ***International Conference on Access to Recent Advances in Engineering and Digitalization (ARACONF 2020)***.

† Corresponding Author: Nigde Omer Halisdemir University, Engineering Faculty, Computer Engineering Department, Nigde, Türkiye, ORCID: 0000-0001-7202-2899, ytorun@ohu.edu.tr

In this study, first of all, we explained speech recognition problem and the steps of speech recognition. Then, we analyzed the studies related to deep learning based speech recognition. In particular, deep learning architectures of Deep Neural Networks, Convolutional Neural Networks, and Recurrent Neural Networks and hybrid approaches that use these architectures are evaluated and the literature studies related to these architectures for speech recognition and the application areas of speech recognition are investigated. As a result, we observed that RNNs are the most utilized and powerful deep learning architecture among all of the deep learning architectures in terms of error rates and speech recognition performance. CNNs are other successful deep learning architectures and have closer results with RNN in terms of error rates and speech recognition performance. Also, we observed that new deep architectures that use either hybrid of DNNs, CNNs, and RNNs or other deep learning architectures are getting attention and have increasing performance and could reduce error rates in speech recognition.

Keywords: Speech recognition, Deep learning, DNNs, CNNs, RNNs, Hybrid architectures.

1. Introduction

Speech recognition is the task of processing audio files and converting text transcription of the input audio files (Yu and Deng, 2016). Speech recognition gained attention with the availability of high performance computing systems, presence of more data for training speech recognition systems, and the effective use of new computer science methods and algorithms for speech recognition, such as deep learning architectures. With the use of deep learning architectures, speech recognition achieved massive performance, and the speech recognition systems became a part of people's daily lives.

Deep learning is a branch of machine learning that uses a set of algorithms that attempt to model high-level abstractions in data by using a deep graph with multiple processing layers, composed of multiple linear and non-linear transformations (Dahl et al., 2011; Yu and Deng, 2016). Deep learning provides automatic selection and ranking of features in the datasets with efficient algorithms. Deep learning architectures had tremendous success in speech recognition, image processing, natural language processing, and sequence prediction.

Deep learning has several deep architectures, such as Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) which are widely utilized to be used in speech recognition systems (Yu and Deng, 2016). DNNs and CNNs are feed-forward architectures that contain multiple layers of transformations and nonlinearity with the output of each layer that are feeding subsequent layer. RNNs is a recurrent architecture that has both forward pass which transfers information to subsequent layers and recurrent pass that processes past information and current input together.

With the successful studies that use deep learning architectures for speech recognition, deep learning gained much attention in speech recognition domain. Researchers investigated the use of DNNs, CNNs, and RNNs for both acoustic modelling, and also for end-to-end speech recognition systems. In the literature, the studies consider using deep learning architectures for end-to-end speech recognition have impact and provided better performance. Also, researchers consider proposing hybrid strategies that are combinations of deep learning architectures.

In this study, we analyzed the studies related to deep learning based speech recognition. First, mostly utilized and popular deep learning architectures of DNNs, CNNs, and RNNs are explained, and then the literature studies related to these architectures for speech recognition are investigated. The modifications of each architecture for achieving better speech recognition performance are researched, and also the applications of speech recognition, such as speech emotion recognition, voice activity detection, and speaker verification, are researched for each architecture. Also, hybrid architectures that use more than one deep learning architecture for speech recognition are analyzed.

The rest of this study is organized as follows. Section 2 presents the speech recognition problem and steps of speech recognition. Section 3 presents deep learning architectures of DNNs, CNNs, and RNNs. Section 4 presents the studies related to these architectures and also hybrid architectures. Section 5 presents the discussion about the literature studies.

2. Speech Recognition

Speech recognition is the task of producing a text transcription of the audio signal from a speaker (Yu and Deng, 2016). Speech recognition gained attention recently, with the help of computation power of computing devices, presence of more data, and increasing success rates of speech recognition systems. Speech recognition has several possible applications, such as voice search, personal digital assistance, smart home environments, and mobile communications.

Formally, speech recognition problem can be explained as given in Equation (1) (Yu and Deng, 2016). Given a sequence of t vectors of acoustic information $X = x_1 \dots x_t$ that we assume encodes a sequence of T words $w = w_1 \dots w_t$. The aim of speech recognition is to find the best transcription hypothesis w according to some learned scoring function $s(w, X)$:

$$\widehat{w} = \underset{w}{argmax} \quad s(w, X) \quad (1)$$

A typical speech recognition system consists of four modules, namely, signal processing and feature extraction, acoustic model, language model, and hypothesis search (Yu and Deng, 2016). Signal processing and feature extraction module takes audio signal as input and removes noises, converts the signal to feature domain, and extracts features from the audio. The acoustic model takes features as input and phonetic knowledge and generates an acoustic model score for the variable-length feature sequence. The language model estimates the probability of a hypothesized word sequence by using the correlation between words in a training

corpus. The performance of language model could be improved with providing domain knowledge to the model. The hypothesis search component combines acoustic model and language model scores and the hypothesized word sequence, and outputs the word sequence with the highest score as the recognition result. Basic speech recognition flow is presented in Figure 1 with respect to these four modules.

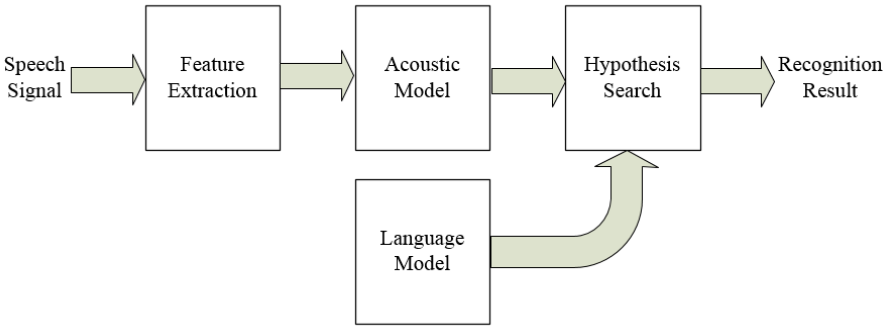

Figure 1: Basic flowchart of speech recognition

3. Deep Learning Architectures

In this section, we introduced Deep Neural Networks, Convolutional Neural Networks, and Recurrent Neural Networks which are highly preferred and successful deep learning architectures in speech recognition and its applications.

3.1. Deep Neural Networks

Deep Neural Networks (DNN) are a type of Artificial Neural Networks (ANN) with multiple layers between input and output layers (Dahl et al., 2011; Dahl et al., 2013). The main aim of DNN is to find a proper mathematical modelling for a given input to obtain the output. In exploration of mathematical explanation, DNN considers both linear and non-linear relations of input and hidden vectors to achieve the desired output. Figure 2 presents an example of a DNN structure.

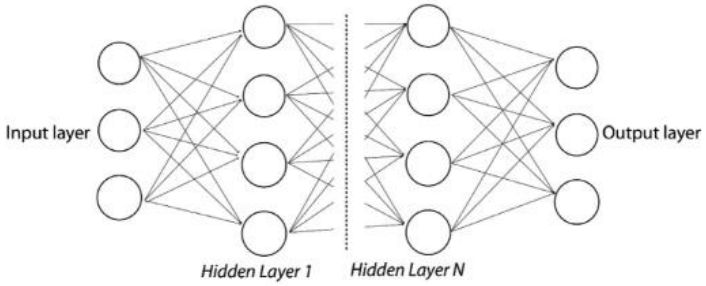

Figure 2 Example of a DNN structure

DNN are powerful for modelling complex non-linear relationships of input-output pairs. Because the DNN have multiple hidden layers and have many neurons in the layers, the DNN have the ability to model and process multiple features. DNN are strong alternatives to ANN, however have two main challenges. First of all, the error is propagated to first layers and the effect of the error to other layers is minor. Second, the learning process of DNN is slow due to complex and high-order matrix multiplications.

3.2. Convolutional Neural Networks

Convolutional Neural Networks (CNN) are a type of deep learning architectures that is specialized to be mostly used in analysing visual datasets (Abdel-Hamid et al., 2014). In CNN, layers are utilized to perform one specific job, i.e. convolution, or sub-sampling, and then the network is connected to a fully connected deep architecture to produce an output. In each convolution and sub-sampling layers, the higher level features are extracted from the input images and the output becomes more accurate. A sample CNN architecture is presented in Fig. 2.

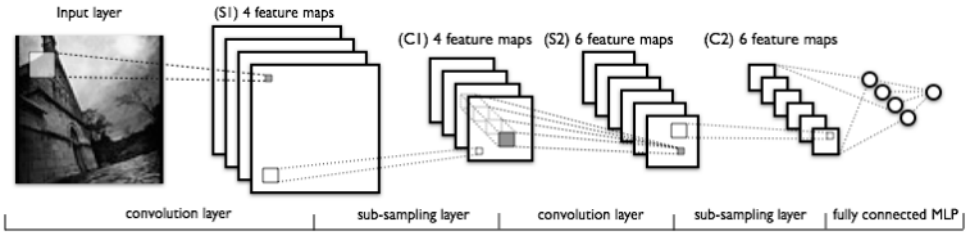

Fig. 2 Example of a CNN architecture

CNN are successful architectures in image analysis and processing, because they have the ability to exploit spatial and temporal correlation in the datasets. Convolution layer explores useful local features from the input data, and sub-sampling layer gets the features and summarizes the results. With the help of these steps, CNN have the ability to extract features automatically.

CNN have some challenges based on their operation and result extraction. First of all, the process of CNN is black-box and the interpretation and explanation are hard. Second, selecting appropriate hyper-parameter values are important for the performance of CNN. Third, efficient training of CNN requires powerful computational resources. Fourth, CNN perform poorer performance with noisy and un-labelled datasets.

3.3. Recurrent Neural Networks

Recurrent Neural Networks (RNN) is a type of deep learning architectures which is capable of handling large sequential inputs (Graves et al., 2013). Main idea behind RNN is to extract outputs of current time step based on current input and previous inputs with weighted manner. this approach is beneficial for several tasks which needs information about previous inputs, such as speech recognition, natural language processing. The weights of input-to-hidden, hidden-to-hidden, and hidden-to-output do not change along the network.

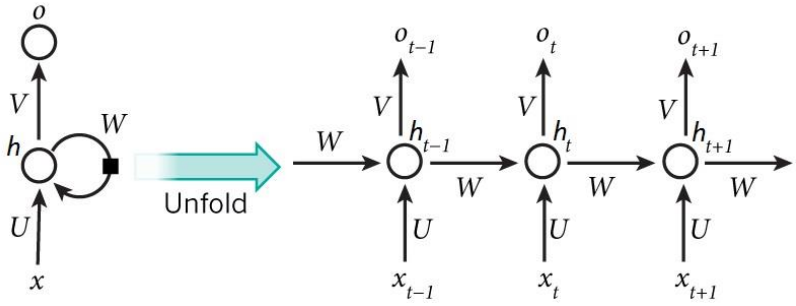

Fig. 3 An RNN (left) and unfolded over time (right)

Fig. 3 presents an RNN architecture which is unfolded over three time instances. x, h and o are input, hidden state and output vectors and U, W and V are weights of input, previous hidden states and current hidden state values, respectively. h_t and o_t are calculated based on (1) and (2). σ_h and σ_o are activation functions of hidden state and output vectors which regulate effect of input and hidden state instances. b_h and b_o are biases of hidden state and output vectors.

$$h_t = \sigma_h(U_{xt} + W_{ht-1} + b_h) \quad (1)$$

$$o_t = \sigma_o(V_{ht} + b_o) \quad (2)$$

There are several things to note in RNN architecture. First is, hidden state of the nodes, h_t in this case, is the memory of the network which passes information through time steps. Second, U, W and V are same for all time steps of the network. Third, there is no need to provide output for every time steps.

4. Deep Learning Architectures for Speech Recognition

In this section, the studies related to deep learning architectures that are used on speech recognition are presented. First, each deep learning architecture is investigated, and then hybrid approaches that combine more than one deep learning architectures are presented.

4.1. Deep Neural Networks for Speech Recognition

Deep neural networks are used in speech recognition systems as an alternative approach mainly for acoustic modelling. With the successful application of DNNs in speech recognition, the studies have emerged that use DNNs for improving accuracy of speech recognition systems and obtaining better performance.

Dahl et al. (2011) proposed a context-dependent pre-trained DNN approach for acoustic model of speech recognition and compared with Gaussian Mixture Model (GMM) and DNN-HMM approach outperformed GMM-HMM approach. Yu et al. (2013) proposed Kullback–Leibler divergence (KLD) regularization adaptation technique for context-dependent DNN-HMM for better speech recognition performance. Seltzer et al. (2013) investigated the performance of DNN-based acoustic models and proposed three methods to improve accuracy for noise robust speech recognition. Jaitly et al. (2012) proposed to use deep belief networks (DBN) for pre-training DNNs for DNN-HMM hybrid speech recognition systems and reported that proposed approach outperformed GMM-HMM baseline. Dahl et al. (2013) investigated the behaviour of DNNs using rectified linear units (ReLU) and dropout and reported that using ReLU and dropout improves performance of speech recognition systems.

In the literature, some studies focus on applications of speech recognition using DNNs. Han et al. (2014) proposed to use DNNs to extract high level features for speech emotion recognition. Lalitha et al. (2019) proposed a DNN based perceptual speech feature extraction approach for emotion recognition. Lei et al. (2014) proposed a framework for speaker recognition using i-vector models and DNNs. Snyder et al. (2016) proposed an end-to-end speaker verification system that consists of DNNs that take variable length speech segments and maps into a speaker embedding, and they reported that proposed system outperformed i-vector based baseline in

equal error-rate (EER) by 13%, average. Variani et al. (2014) investigated the use of DNNs for a small footprint text-dependent speaker verification. Zen et al. (2013) proposed a speech synthesis scheme that is based on DNNs and they reported that DNN based speech synthesis system outperform HMM based baseline.

4.2. Convolutional Neural Networks for Speech Recognition

Convolutional neural networks are one of popular deep architectures for speech recognition due to their ability to reduce spectral variations and to model spectral correlations of speech signals. CNNs use spectrogram of speech signals that is represented as an image and recognize speech based on these features. CNNs are proposed as an alternative to DNN based speech recognition and achieved better performance with its architecture (Sainath et al., 2013).

Sainath et al. (2013) explored configurations of CNN, such as convolution layer count, optimal number of hidden units, best pooling strategy, and best input feature type, for obtaining better performance than DNN for large vocabulary continuous speech recognition. Sercu et al. (2016) proposed a very deep CNN approach that consists of 14 weight layers and applied multilingual speech recognition task on the generated deep architectures. Abdel-Hamid et al. (2014) proposed a CNN based speech recognition system to reduce error rate using limited-weight-sharing scheme to better model speech features. Qian et al. (2016) proposed a very deep CNN model for noise robust speech recognition and investigated best configurations for proposed deep CNN model for noise robust speech recognition. Zhang et al. (2017) proposed an end-to-end speech recognition system that is based on CNNs with Connectionist Temporal Classification (CTC) approach without using a recurrent layer for obtaining computationally efficient model and competitive results. Palaz et al. (2015) investigated the use of CNNs to large vocabulary speech recognition which takes raw speech signals as inputs and the proposed approach outperformed classical DNN based speech recognition system.

Several studies consider using CNNs for applications of speech recognition. Mao et al. (2014) utilized CNN for learning affect-salient features for speech emotion recognition using two-phase learning. Badshah et al. (2017) used CNN for extracting discriminative features for speech emotion recognition using spectrograms of input speech signals. Thomas et al. (2014) utilized CNNs as acoustic models for speech activity detection (SAD) in mismatched acoustic conditions using noisy radio communication channels data. Swietojanski et al. (2014) used CNNs for large vocabulary distant speech recognition using the data from single distant microphone and multiple distant microphones and CNN outperformed DNN and GMM in terms of Word Error Rate (WER). Fu et al. (2016) proposed two signal-to-noise-ratio (SNR) aware algorithms for modelling CNN for speech enhancement and the proposed model outperformed DNN for denoising performance. Torfi et al. (2018) proposed a 3D CNN model for adaptive feature learning for text independent speaker verification.

4.3. Recurrent Neural Networks for Speech Recognition

Recurrent neural networks are the mostly preferred and utilized deep learning architecture due to their ability to model sequential data, including speech recognition. RNNs could model long-term dependencies between features of input datasets and produce output based on past observations. This approach is beneficial for speech recognition tasks, because in speech recognition, the output of a frame is dependent on past frames of observations. RNNs and Long-Short Term Memory (LSTM) RNNs, which is an improved and modified version of RNNs, have the best performance for speech recognition tasks over all deep learning architectures and are preferred among other alternatives.

Graves et al. (2013) investigated the use of deep LSTM RNNs for speech recognition to achieve state-of-the-art results and reported that their deep LSTM model achieved best phoneme error rate. Graves et al. (2013b) investigated the use of deep bidirectional LSTM architecture as an acoustic model to NN-HMM hybrid speech recognition system and achieved equal performance with previous studies. Graves and Jaitly (2014) proposed an end-to-end speech recognition system that do not require phonetic representation using a combination of LSTM and CTC objective function. Sak et al. (2015) proposed techniques that improve performance of LSTM RNNs as acoustic models for LVSR and resulted that stacking frames and reducing frame rate provides more accurate models and faster decoding. Li and Wu (2015) proposed a deep LSTM to obtain performance improvement and applied on large vocabulary telephone speech recognition task and resulted that deep LSTM strategy provide better performance. Miao et al. (2015) proposed an end-to-end speech recognition system that uses weighted finite-state transducers (WFSTs) and bidirectional LSTM deep architecture. Sak et al. (2014) proposed a distributed training for LSTM using stochastic gradient descent on a cluster of machines and reported that their proposed system outperformed DNN. Lu et al. (2016) proposed an efficient learning rate schedule method that improves the accuracy of large vocabulary speech recognition.

Many studies focus on using RNNs for applications of speech recognition. Mirsamadi et al. (2017) investigated using RNNs for automatically extracting emotion-related features for speech emotion recognition by using both short-time frame-level emotional features and temporal aggregation of such features. Maas et al. (2012) investigated the use of deep recurrent auto encoder neural network for noise reduction in automatic speech recognition. Weninger et al. (2015) proposed an LSTM RNN framework that are trained by an optimal speech reconstruction objective for speech enhancement in noise robust speech recognition. Weninger et al. (2014) investigated the use of LSTM RNNs on training, network architecture and representation of features for regression based single-channel speech separation. Hughes and Mierle (2013) proposed a multi-layer RNN model for voice activity detection that outperforms larger baseline GMM with a hand-tuned state machine (SM) system. Sun et al. (2015) investigated the use of Deep Bidirectional LSTM RNN (DBLSTM RNN) for voice conversion which is able to model temporal correlations between speech frames. Zen and Sak (2015) proposed a unidirectional LSTM with recurrent output layers for low-latency speech synthesis system.

4.4. Hybrid Approaches for Speech Recognition

Although one deep learning architecture is sufficient for gaining good performance for speech recognition systems and applications, some studies consider using a hybrid of two or more deep learning architectures to achieve better performance. Trigeorgis et al. (2016) proposed a framework that combines CNNs and LSTMs to automatically learn best feature representation from raw speech signals for speech emotion recognition. Lim et al. (2016) investigated the use of concatenated architecture from CNNs and RNNs for extracting better features than hand-crafted features for speech emotion recognition. Zhao et al. (2018) proposed an end-to-end CNNs and RNNs based model for catching local variations in both time and frequency domains for speech enhancement. Hori et al. (2017) proposed an end-to-end model using CNNs as encoder and LSTMs as language model for speech recognition which reduced the error rate. Chan et al. (2015) utilized RNNs and DNNs for increasing speech recognition performance on embedded devices by building a large RNNs acoustic model and pass this model to DNNs for speech recognition. Chen et al. (2018) proposed a 3D attention-based CRNN deep learning architecture which takes MFCC with deltas and delta-deltas as input for speech emotion recognition. Wu et al. (2016) proposed a deep model in which CNNs and DNNs extract visual cues and acoustic features, and BiLSTMs model higher level dependencies among features and visual information. Sainath et al. (2015) combined CNNs, LSTMs, and DNNs into a unified deep learning system, which is named as CLDNN, for taking advantage of each architecture and achieved better performance than LSTM which is considered as strongest architecture of these three alternatives in speech recognition. Wang et al. (2019) proposed CNN-BLSTM-CTC deep learning hybrid model for Mandarin speech recognition. They employed CNN for learning of local speech features, BLSTM for learning past and future dependencies, and CTC for decoding purposes and claim that their proposed method outperformed best existing model.

5. Discussion

In this paper, we reviewed the studies that consider using deep learning architectures for speech recognition. First, the most utilized deep learning architectures of Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) are presented, and then the studies that use these networks for speech recognition are investigated. When the investigated studies are evaluated, it is observed that there are many studies present for each deep learning architecture, especially those that use RNNs and LSTMs. DNNs are preferred as a hybrid method for HMMs and CNNs are utilized when spectrograms are used as input for speech features, while RNNs are utilized when using raw speech signals, such as MFCC features. However, hybrid architectures are getting attention recently, which has more potential and achieves better performance in speech recognition tasks with respect to using only one deep architecture. Using a hybrid deep architecture provides the utilization of benefits of each deep learning architecture which results better performance, but also requires much hard work for training such architectures. As a result, when the investigated studies are examined for speech recognition, new deep architectures that use either hybrid of reviewed architectures or other deep learning architectures are getting attention.

As applications of speech recognition, speech emotion recognition is the uttermost studied application, which tries to discover emotions in the input speech data. Also, speech enhancement, speech separation, voice activity detection, and speaker identification and verification are other widely studied applications. Deep learning architectures have many benefits and are successfully utilized on the applications of speech recognition.

References

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, 22(10), 1533-1545.

Badshah, A. M., Ahmad, J., Rahim, N., & Baik, S. W. (2017, February). Speech emotion recognition from spectrograms with deep convolutional neural network. In 2017 international conference on platform technology and service (PlatCon) (pp. 1-5). IEEE.

Chan, W., Ke, N. R., & Lane, I. (2015). Transferring knowledge from a RNN to a DNN. arXiv preprint arXiv:1504.01483.

Chen, M., He, X., Yang, J., & Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters, 25(10), 1440-1444.

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2011). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on audio, speech, and language processing, 20(1), 30-42.

Dahl, G. E., Sainath, T. N., & Hinton, G. E. (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8609-8613). IEEE.

Fu, S. W., Tsao, Y., & Lu, X. (2016, September). SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement. In INTERSPEECH 2016, San Francisco, USA (pp. 3768-3772).

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). IEEE.

Graves, A., Jaitly, N., & Mohamed, A. R. (2013b, December). Hybrid speech recognition with deep bidirectional LSTM. In 2013 IEEE workshop on automatic speech recognition and understanding (pp. 273-278). IEEE.

Graves, A., & Jaitly, N. (2014, January). Towards end-to-end speech recognition with recurrent neural networks. In International conference on machine learning, Beijing, China (pp. 1764-1772).

Han, K., Yu, D., & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In Fifteenth annual conference of the international speech communication association INTERSPEECH 2014, Singapore.

Hori, T., Watanabe, S., Zhang, Y., & Chan, W. (2017). Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM. arXiv preprint arXiv:1706.02737.

Hughes, T., & Mierle, K. (2013, May). Recurrent neural networks for voice activity detection. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 7378-7382). IEEE.

Jaitly, N., Nguyen, P., Senior, A., & Vanhoucke, V. (2012). Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition. In Thirteenth Annual Conference of the International Speech Communication Association INTERSPEECH 2012, Portland, OR, USA.

Lalitha, S., Tripathi, S., & Gupta, D. (2019). Enhanced speech emotion detection using deep neural networks. International Journal of Speech Technology, 22(3), 497-510.

Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014, May). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1695-1699). IEEE.

Li, X., & Wu, X. (2015, April). Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4520-4524). IEEE.

Lim, W., Jang, D., & Lee, T. (2016, December). Speech emotion recognition using convolutional and recurrent neural networks. In 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA) (pp. 1-4). IEEE.

Lu, L., Zhang, X., & Renais, S. (2016, March). On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5060-5064). IEEE.

Maas, A., Le, Q. V., O'neil, T. M., Vinyals, O., Nguyen, P., & Ng, A. Y. (2012). Recurrent neural networks for noise reduction in robust ASR. INTERSPEECH 2012, Portland, OR, USA.

Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. IEEE transactions on multimedia, 16(8), 2203-2213.

Miao, Y., Gowayyed, M., & Metze, F. (2015, December). EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. In 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) (pp. 167-174). IEEE.

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017, March). Automatic speech emotion recognition using recurrent neural networks with local attention. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2227-2231). IEEE.

Palaz, D., Doss, M. M., & Collobert, R. (2015, April). Convolutional neural networks-based continuous speech recognition using raw speech signal. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4295-4299). IEEE.

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

Sainath, T. N., Mohamed, A. R., Kingsbury, B., & Ramabhadran, B. (2013, May). Deep convolutional neural networks for LVCSR. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8614-8618). IEEE.

Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015, April). Convolutional, long short-term memory, fully connected deep neural networks. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4580-4584). IEEE.

Sak, H., Senior, A., Rao, K., & Beaufays, F. (2015). Fast and accurate recurrent neural network acoustic models for speech recognition. INTERSPEECH 2015, Dresden, Germany.

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. the INTERSPEECH 2014, Singapore.

Seltzer, M. L., Yu, D., & Wang, Y. (2013, May). An investigation of deep neural networks for noise robust speech recognition. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 7398-7402). IEEE.

Sercu, T., Puhrsch, C., Kingsbury, B., & LeCun, Y. (2016, March). Very deep multilingual convolutional neural networks for LVCSR. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4955-4959). IEEE.

Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., & Khudanpur, S. (2016, December). Deep neural network-based speaker embeddings for end-to-end speaker verification. In 2016 IEEE Spoken Language Technology Workshop (SLT) (pp. 165-170). IEEE.

Sun, L., Kang, S., Li, K., & Meng, H. (2015, April). Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4869-4873). IEEE.

Swietojanski, P., Ghoshal, A., & Renals, S. (2014). Convolutional neural networks for distant speech recognition. IEEE Signal Processing Letters, 21(9), 1120-1124.

Thomas, S., Ganapathy, S., Saon, G., & Soltau, H. (2014, May). Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2519-2523). IEEE.

Torfi, A., Dawson, J., & Nasrabadi, N. M. (2018, July). Text-independent speaker verification using 3d convolutional neural networks. In 2018 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016, March). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5200-5204). IEEE.

Variani, E., Lei, X., McDermott, E., Moreno, I. L., & Gonzalez-Dominguez, J. (2014, May). Deep neural networks for small footprint text-dependent speaker verification. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4052-4056). IEEE.

Wang, D., Wang, X., & Lv, S. (2019). End-to-End Mandarin Speech Recognition Combining CNN and BLSTM. Symmetry, 11(5), 644.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., & Schuller, B. (2015, August). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In International Conference on Latent Variable Analysis and Signal Separation (pp. 91-99). Springer, Cham.

Weninger, F., Hershey, J. R., Le Roux, J., & Schuller, B. (2014, December). Discriminatively trained recurrent neural networks for single-channel speech separation. In 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP) (pp. 577-581). IEEE.

Wu, Z., Sivadas, S., Tan, Y. K., Bin, M., & Goh, R. S. M. (2016). Multi-modal hybrid deep neural network for speech enhancement. arXiv preprint arXiv:1606.04750.

Yu, D., Yao, K., Su, H., Li, G., & Seide, F. (2013, May). KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 7893-7897). IEEE.

Yu, D., & Deng, L. (2016). Automatic Speech Recognition A Deep Learning Approach. Springer.

Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C. L. Y., & Courville, A. (2017). Towards end-to-end speech recognition with deep convolutional neural networks. arXiv preprint arXiv:1701.02720.

Zhao, H., Zarar, S., Tashev, I., & Lee, C. H. (2018, April). Convolutional-recurrent neural networks for speech enhancement. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2401-2405). IEEE.

Zen, H., Senior, A., & Schuster, M. (2013, May). Statistical parametric speech synthesis using deep neural networks. In 2013 ieee international conference on acoustics, speech and signal processing (pp. 7962-7966). IEEE.

Zen, H., & Sak, H. (2015, April). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4470-4474). IEEE.