Volume: 5, Issue: 3
Page: 1-14
2021

International Journal of
Science and Business
Journal homepage: ijsab.com/ijsb

IJSAB International

Deep Learning Convolutional Neural Network for Speech Recognition: A Review

Kazheen Ismael Taher & Adnan Mohsin Abdulazeez

Abstract:
In the last few decades, there has been considerable amount of research on the use of Machine Learning (ML) for speech recognition based on Convolutional Neural Network (CNN). These studies are generally focused on using CNN for applications related to speech recognition. Additionally, various works are discussed that are based on deep learning since its emergence in the speech recognition applications. Comparing to other approaches, the approaches based on deep learning are showing rather interesting outcomes in several applications including speech recognition, and therefore, it attracts a lot of researches and studies. In this paper, a review is presented on the developments that occurred in this field while also discussing the current researches that are being based on the topic currently.

IJSB
Literature review
Accepted 24 January 2021
Published 28 January 2021
DOI: 10.5281/zenodo.4475361

Keywords: *Machine Learning, Speech Recognition, Convolutional Neural Networks, Deep Learning, Word Error Rate (WER)*

About Author (s)

Kazheen Ismael Taher (corresponding author), Information Technology Department, Akre Technical College of Informatics, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq. E-mail: kajeen.ismael@gmail.com
Professor Adnan Mohsin Abdulazeez, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq. E-mail: adnan.mohsin@dpu.edu.krd

1. Introduction

Machin Learning is now used in all fields of computer work where algorithms are developed and performance is enhanced (Abdulqader et al., 2020; Chaudhary & Vasuja, 2019; Zeebaree et al., 2019a; Abdulazeez et al., 2020; Jahwar & Abdulazeez, 2020; Maulud & Abdulazeez, 2020). Learning from unbalanced data sets has been a key challenge in machine learning in recent years and is also used in many implementations, such as information security, engineering, remote sensing, biomedicine, and transformation (Abdulqader et al., 2020; Muhammad et al., 2020; Anuradha & Reddy, 2008) industries. Classification, regression, and band techniques include supervised learning approaches where the focus variable is categorical in classification and tends to decline (Zeebaree et al., 2019b; Sethi & Mittal, 2019). A huge number of samples (tuples) and a few attributes are contained in ML datasets. Microarray technology is very distinct from the normal databases for ML (Abdulazeez et al., 2020; Zeebaree et al., 2017). Help vector machines are one of the supervised learning approaches used for classification Support Vector Machine (SVM). The neural network is used for data training. The algorithm for lazy learning is K-Nearest-Neighbor (K-NN).

Deep learning (DL) is a part of ML. Unlike ML, DL networks are not directly used to retrieve and classify functions. Without engaging the outside observer, the hidden layers of the deep learning network do all of these indirectly by itself (Zebari et al., 2020; G. et al., 2018; Song, 2020). DL is a theory that comes from the artificial neural network, which uses multiple layers of processing units to transform and extract features. Each layer's input is the output of the previous layer. DL consists of many architectural styles such as Deep Neural Network (DNN), CNN, and Recurrent Neural Network (RNN). Where, like several layers, DNN corresponds to a Multi-Layer Perceptron (MLP). In speech recognition, CNN is often successive and consists of convolutional layers, pooling layers, and non-linear layers. Meanwhile, by cyclic relation, RNNs are established (Ahmed & Brifcani, 2019; Obaid et al., 2020; Zeebaree et al., 2020; Ahmed & Abduallah, 2017).

Transcription of human expression into spoken words is the target of Automated Speech Recognition (ASR). Due to different speaker characteristics, different voice patterns, uncertain ambient sounds, and so on, it is a very difficult activity because human speech signals are highly variable. Also, ASR requires the translation of variable-length speech signals into sequences of words or phonetic symbols of variable length (Abdel-Hamid et al., 2014). DL algorithms have almost all been used to further develop computer skills to understand what people can do, including voice recognition. In particular, voice, being the primary medium of contact between human beings, has gained a great deal of attention from the introduction of artificial intelligence over the past five decades (Nassif et al., 2019; Anasuya & Katti, 2009; Adeen et al., 2020). A raw input signal sequence is given to the convolutional neural network, separated into frames, and a score for each class, for each frame, is output. The network architecture, followed by a classification stage, consists of multiple filter stages. A filter stage includes a convolutional layer, followed by a non-linearity (tanh()) and a temporal max-pooling layer (Palaz & Collobert, 2015). Recent developments in ML have made it possible to train systems in an end-to-end way, i.e. systems where every step is concurrently trained, taking into account all the other steps and the entire system's final mission. It is generally referred to as deep learning, primarily because, opposed to classical "shallow" systems, such architectures are normally composed of several layers (supposed to have an increasing degree of abstraction) (Palaz et al., 2015).

CNN consists of one or more pairs of convolution and max-pooling layers (Zeebaree et al., 2018; Omar et al., 2020; Hasan & Abdulazeez, 2020; LeCun et al., 1998; Abdel-Hamid et al., 2012; Abdulazeeza et al., 2020). A convolution layer uses a series of filters that process small local input sections where these filters are repeated in the entire input space. By taking the full filter activation from different positions within a given window, a max-pooling layer produces a lower resolution version of the convolution layer activation. This adds in-variance and tolerance for the localization of small variations in the location of pieces of objects. To process more complex sections of the data, higher layers use wider filters that work on lower resolution inputs. To identify the total inputs, the top-related layers ultimately merge inputs from all locations. This hierarchical organization delivers strong results in tasks for image processing (Abdel-Hamid et al., 2012; LeCun et al., 2004a).

The effect of CNNs was first verified in the classification of photos (LeCun et al., 1995; LeCun et al., 1998; Abdel-Hamid et al., 2012; Qian et al., 2016) and subsequently extended to speech recognition (Graves et al., 2013; Abdel-Hamid et al., 2012; Sainath et al., 2015; Abdel-Hamid et al., 2014; Swietojanski et al., 2014). Most of the previous CNN speech recognition methods used only up to 2 convolutional layers, and (Zeebaree et al., 2017; Qian et al., 2016) attempted to raise to 3 convolutional layers but achieved impaired efficiency. The computer vision community (Szegedy et al., 2015; Bargarai et al., 2020; Zhang et al., 2017) has recently found that by using a significantly increased number of convolutional layers with carefully built topology, the efficiency of image classification can be enhanced. Specifically, using very basic building blocks, such as convolutional layers with 3 × 3 filters and 2 × 2 pooling layers, VGGNet (Simonyan & Zisserman, 2014; Qian et al., 2016) is built and displays amazing efficiency.

In this paper, the theory of speech recognition and CNN are presented in Section 2, whereas the related work is summarizing in Section 3. Section 4 presences the discussion and comparison, where Section 5 concludes the paper.

2. Speech Processing

Speech processing is the study of speech signals and the methods of signal processing. Signals are typically processed in a digital representation, so speech processing should be viewed as a special case of digital signal processing applied to speech signals. Speech processing concerns the acquisition, manipulation, handling, delivery, and output of speech signals. The input is called recognition of speech and the output is called the synthesis of speech. Speech signals can provide us with knowledge of various kinds. All sorts of data are (Nassif et al., 2019):

- Recognition of speech, which offers details on the content of voice signals.
- A recognition of the speaker that holds details on the identity of the speaker.
- Recognition of feeling, which offers details about the mental state of the speaker.
- Health identification, which provides information about the health status of the patient.
- Language understanding, which includes specifics of the language spoken.
- Accent detection, which creates data regarding the accent of the speaker.
- Age identification, which offers information about the age of the speaker.
- Sex identification, which holds details on the gender of the speaker.

Automatic speaker recognition can be identified as the process of recognition of an unknown speaker based on information encoded in the speech signal using a computer (computer). Recognition of speakers is broken into two parts: recognition of speakers and authentication of speakers (authentication). The protocol for deciding which of the registered speakers refers to that assertion shall be referred to as the identifying component of the speaker. This part can be found in public buildings or newspapers. These include, but are not limited to, district or other government departments, radio stations, insurance companies, or recorded conversations (Reynolds, 2002; Furui, 1991; Nassif et al., 2019).

Automatic identification of gender is the method of recognizing whether a speaker is male or female. Generally, automated gender identification results in high precision without much intervention, since the results of this form of recognition are binary (either male or female) (Vogt & André, 2006; Bocklet et al., 2008; Nassif et al., 2019). Automatic gender identification is apparent in the implementations of call centers in some traditional cultures, where automatic dialog systems with the ability to identify gender are preferred over those lacking that ability. The block diagram system is shown in Figure 1:

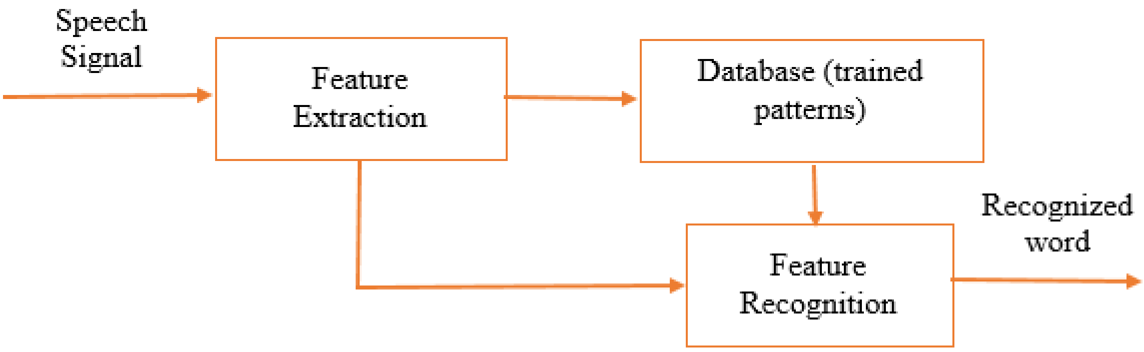

Figure 1: speech recognition block diagram system (V.Chapaneri, 2012)

3. Speech Recognition Techniques

A computer can "hear," interpret," understand and "act upon" spoken information, the purpose of speech recognition. Bell Laboratories, Davis, Biddulph, and Balashek first sought to establish a single speaker's single-digit recognition device in the early 1950s (Klevans & Rodman, 1997). The purpose of automatically identifying speakers is the study, extraction, and identification of speech identity information. In a four-stage the speaker recognition system can be considered to function (Gaikwad et al., 2010):

1. Research: Analysis
2. Extraction of function
3. Customization
4. Check

4. Feature Extraction Technique

The elimination of the language function in an issue of categorization concerns the reduction of input vector dimensions while retaining the discriminatory power of the signal. We know that the amount of training and test vectors required for classification problems increases with the input parameter, as we are aware of fundamental learning in the speaker recognition and verification method so that a speech signal can be extracted by function (Gaikwad et al., 2010; Laine, 2017; J. Ahmed & Brifcani, 2015).

5. Convolutional Neural Network

CNN (LeCun et al., 2004b)is a DNN with linked networks. It is a DNN. A CNN conducts raw data convolution processes (for example, sensor values) and is one of the most sophisticated and well-researched strategies for the profound learning of images, word modeling, voice recognition, and recent human activity recognition based on smartphone and wearable sensors. The CNN model typically consists of a convolutional layer, pool layer, and fully interconnected layer. These layers are stacked into deep architecture to automatically remove functionality from raw sensor data (Ordóñez & Roggen, 2016; Nweke et al., 2018).

5.1 Convolutional Neural Network Architecture

CNN architecture shows the highest performance by manipulating two-dimensional data formats such as images, video, and so on. CNN's spatially local association occurs from the introduction of a local pattern of communication between neighboring layer neurons. CNN differs from Neocognitron because it never needs a decrease in the number of weights, thereby removing the need for a special attribute extraction algorithm that is commonly used in traditional learning algorithms. CNN architecture is specified in a typical neural multilayer network. (Sornam et al., 2017).

.

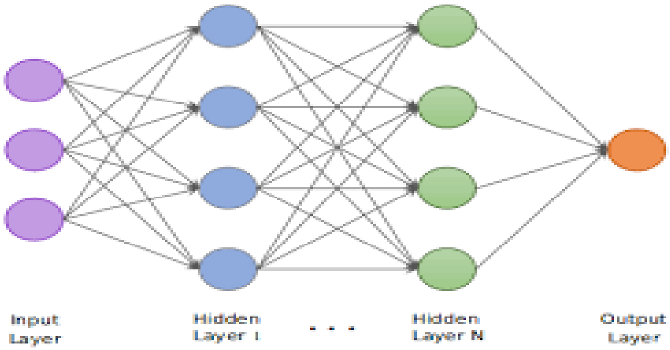

Figure 2: Connectivity Pattern of Neurons (Sornam et al., 2017)

There are input, output, and opaque layers in between in the CNN architecture. Hidden layers execute a role called the identification of features. The Linear Unit for Convolution, Pooling, and Rectifier makes the network learn quicker and work more effectively. If the function detection has been completed, CNN's architecture progresses towards classification. The CNN part is shown in the figure below. The next layer to the last is a connected layer that outputs the vector of the K dimension. K is the number of output groups that can be estimated by the network (Sornam et al., 2017).

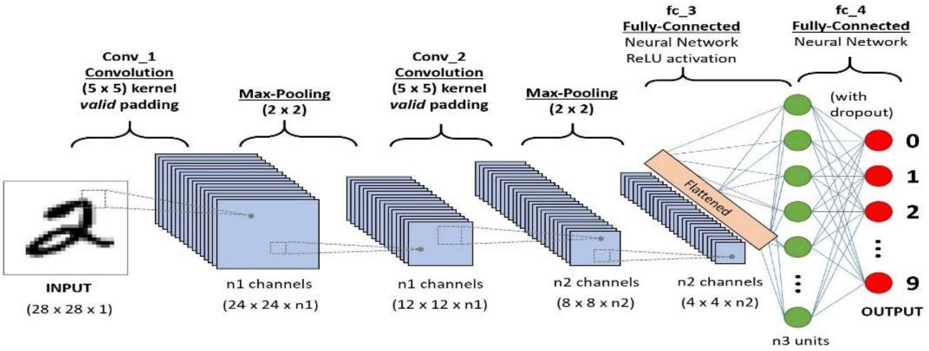

Figure 3: CNN architecture (Sornam et al., 2017)

5.2 CNN Layers

5.2.1 Convolutional layer

Convolutional layers within CNN contain a variety of filters that can be trained, often kernels. Each filter is automatically strides over the whole input, and an output function map is generated for each filter. When the same filter is added to the whole input space, it can be used in the input at many points. Convolutional layers use relatively small filter sizes, which allow local characteristics to be sensed that are critical for data like images and sound. The function map of the previous layer contains filters of stacked convolution layers that combine the features of a higher order (Jansson, 2018; Palaz & Collobert, 2015).

On two-dimensional sources, such as the height and width of images, the most common convolution layer is used. Since the 1-dimensional raw audio waveform inputs are used for this work, one-dimensional filters are used in convolutional layers.

5.2.2 Pooling Layer

Whilst convolutional layers' sense local characteristics from their data, a pooling layer fuses semantically related characteristics only with the maximum value (max pooling) or normal (average pooling). This decreases the sensitivity of feature modifications and distortions by reducing the feature map size. The pooling layers are used after one or more convolutional layers, which minimize the input by the deepening of the network constantly. Each decrease in the input size and the grouping operation also decreases the measurement required in the network (LeCun et al., 2015; LeCun et al., 1998).

6. Related Work

Several surveys have been performed in the field of speech recognition. (Morgan, 2011; Nassif et al., 2019) performed a speech recognition analysis with the assistance of discriminatory qualified feed-forward networks. The key aim of the analysis was to present the papers that utilized several layers of processing before Markov model-based encoding word sequences (Nassif et al., 2019). Anvarjon et al. (2020) suggested a new simple, efficient Speech Emotion Recognition (SER) model with low computational complexity and high accuracy of recognition. The proposed method uses the CNN approach to learn the characteristics of deep frequency. Two benchmarks, including the immersive emotional dyadic motion capture (IEMOCAP) and the Berlin Emotional Expression Database (EMO-DB). Speech databases tested the proposed SER model and achieved 77.01 percent and identified 92.02 percent of the results. The researchers claim that it is possible to improve the recognition thresholds and even integrate them with other methods of deep learning. However, The National Research Foundation of Korea which is financed by the Korean Government through the Ministry of Science and ICT sponsored this process. Adebowale et al. (2020) proposed Intelligent Phishing Detection System (IPDS) gave an outstanding 93.28 percent accuracy in classification. Based on the behavioral characteristics gathered from previous data sets, the scheme is able to filter malicious websites. The IPDS was able to react with great agility in real time and could check a URL before loading on the user's device in 25 s. Overall, in terms of time, CNN performed better, but it was marginally less efficient on average than LSTM. The analysis is an extension of our previous work that considered how best to merge image, text and frame characteristics with a deep learning algorithm (LSTM CNN) to construct a combined phishing detection system solution.

Bingol & Aydogmus (2020) suggested DNN Pruning deep Neural Network (p-DNN) was compared to classic algorithms for classification. A KUKA KR Agilus KR6 R900 sixx robot manipulator was used to apply the built HRI framework. After hearing the word "KUKA " (the brand name of the robot producer) before any conversation, the robot starts listening. Researchers plan to develop image processing algorithms in future studies that can handle more complicated industrial procedures, such as welding. The study was carried out with a low-grade camera costing about $2, deemed to be a reasonable outcome. The contribution of this thesis is that an industrial robot uses this software to conveniently configure individuals who are not robotics experts and know Turkish. The importance of WAR suggests that, on a broad scale, the mechanism is not dependent on speakers. Bird et al. (2020) proposed the classification methods of the Artificial Neural Network and Secret Markov Model for Human Speech Recognition in the English Phonetic Alphabet. Subjects from the United Kingdom and Mexico record a collection of audio clips and the recordings are converted into a static statistical dataset. A deep neural network with evolutionary optimized topology achieves

90.77 percent phoneme classification accuracy. The multi-objective secularization approaches provided, in which the use of real-time resources is often taken into account for solution fitness, far more optimal solutions have generated that train much faster than the forecast approach. The solutions obtained are much more difficult than the HMM, which requires roughly 248 seconds to practice on strong hardware. Also, models that needed less computational resources and yet outperformed HMM were extracted via a multi-objective algorithm relative to the Hidden. Markov model. A more fine-tuned minimization of resources (ʎ2) should be explored in more studies since a value of 0.9 proved to be too drastic and yielded poor results, and so more pair weights should be explored to this end. revealed that hyper-heuristically optimizing the topology of an Artificial Neural Network led both native and non-native English speakers to a high classification capability of the MFCC data. Fantaye et al. (2020) proposed the Amharic and Chaha datasets as well as on the benchmark datasets published under the Intelligence Advanced Research Projects Operation (IARPA) Babel Program's restricted language packages (10-h) The authors conclude that the optimal neural network architectures and training strategies are discussed. Two Ethiopian languages were used to test the performance of these sophisticated neural network models. All the suggested advanced neural network acoustic models were found to be successful for speech recognition systems. They will also investigate new acoustic models of neural networks that combine the CNN model with advanced RNN models, such as LSTM and GRU. They used minimal language packages of the IARPA Babel databases in the research. He said the authors are involved in investigating the feasibility of their proposed neural Network models for the complete language packages. The Chinese Government Scholarship completely supports this initiative.

Additionally, Zoughi et al. (2020) proposed a methodology that further lowers the error rate of identification. In some speech recognition activities, the proposed approach decreases the absolute error rate by 7 percent relative to state-of-the-art methods, they say. Suggested Adaptive Windows Convolutional Neural Network (AWCNN) against both intra- and inter-speaker variants. Also suggested new residual learning, which leads to greater use of deep-layer information and allows better control over transferring input data, it says. suggested speech recognition system can be used for many artificial and professional applications. Deep MRes network has a lower error rate and, on the validation, set, more generalization. the proposed process, the number of parameters increases relative to CNN and Res. They concluded that in industrial and commercial ASR goods, the proposed approach can be used and applied effectively. The method proposed is a route to a powerful acoustic model. Moreover, Huang et al. (2019) suggested an approach to the identification of spontaneous speech emotion considering verbal and nonverbal sounds of speech. There was a segmentation process for decomposing input speech into prosodic sentences, nonverbal intervals, and parts of silence. Then, each section is represented by concatenating emotional features and sound features derived from the CNN-based models as feature vectors. The LSTM-based sequence-to-sequence model generated the emotional sequence as a result, given the sequence of feature vectors. Experimental findings on the identification of seven emotional states in the NNIME (Chinese immersive multimodal emotion corpus of the NTHU-NTUA) revealed that the proposed approach obtained a 52.00 percent detection accuracy that surpassed conventional methods. Finally, the findings revealed that CNN-based features and LSTM-based emotion model are helpful for emotion recognition relative to the standard frame-based emotion recognition technique. Jing et al. (2019) proposed two approaches in speech recognition based on CNN to boost speech accuracy. In the pooling layer of the CNN model, a dynamic adaptive pooling (DA-pooling) algorithm is proposed. A dropout technique based on sparseness is proposed in a full-connected layer to overcome conventional dropout hiding neuron nodes randomly. Results show that the CNN-based speech recognition added DA- pooling algorithm and enhanced dropout sparseness strategy will increase the spars nesses of speech recognition. It will decrease the sophistication of the model, speed up the training of the model, and increase the RR of Speech Recognition, authors say. They say in the future, under noisy circumstances, they plan to investigate the efficiency of speech Recognition in noisy circumstances. Liu et al. (2019) worked on the Long Short Term Memory (LSTM) to estimate the coefficients in WPE, which compensates for the drawbacks and enhances speech recognition efficiency. The findings of the experiment on the CHiME-5 dataset reveal that the best model of the proposed approach has a 2.1 percent reduction in the absolute Word Error Rate (WER) relative to the baseline design. A novel method of voice deriver beration for far-field speech recognition was introduced in this article. To address the shortcomings of the previous approach, this technique blends LSTM and WPE. For BNF, the TDNNLSTM acoustic model is used in the back end. The upgraded system's WER is 73.78 percent, which corresponds to an actual decrease of 2.1 percent relative to the base-line system. Meng et al. (2019) suggested anew ADRNN (dilated CNN with residual block and BiLSTM based on the attention mechanism) architecture is presented to apply for speech emotion recognition. In the spontaneous emotional expression of the IEMOCAP database, it is higher than 64.74 percent of previous state-of-the-art approaches. Also, on Berlin EMODB with speaker-dependent and speaker-independent tests respectively, suggested networks that obtain identification accuracy of 90.78 percent and 85.39 percent, which are higher than previous work. Concluded that the algorithm is incredibly successful and can be used to identify speech emotions with 63.84 percent accuracy in the end.

Deepak & Ameer (2019) suggested classification method adopts the principle of deep transfer learning and extracts features from brain Magnetic Resonance Imaging (MRI) images using a pre-trained GoogLeNet. To identify the extracted functions, validated classifier models are implemented. Experimented follows a five-fold cross-validation procedure at patient level, on fig share's MRI dataset. A mean classification accuracy of 98 percent is recorded by the proposed scheme, outperforming all state-of-the-art approaches. The Area Under the Curve (AUC), accuracy, recollection, F-score and precision are other success metrics used in the analysis. Moreover, by testing the scheme with less training sets, the paper tackles a realistic factor. They fined of the research show that when the supply of medical images is reduced, transfer learning is a valuable technique. Also presented an empirical debate on misclassifications. Sajjad et al. (2019) suggested a new multi-grade brain tumor classification system based on the CNN. Using a deep learning method, tumor regions from an MR image are segmented. Data augmentation is used to successfully train the proposed method, eliminating the lack of data issue. On both augmented and original data, the proposed device is experimentally tested and findings indicate its persuasive success relative to current approaches, say the authors. The experimental findings indicate the efficacy of the proposed CNN-based CAD method to support the radiologist in taking a specific judgment on the four-grade classification of brain tumors.

Zhao et al. (2019) introduced 1D and 2D CNN LSTM networks to classify the emotions in speech. Looked at how local similarities and global contextual knowledge can be learned from raw audio clips and log-Mel spectrograms. Authors concluded that the acquisition of greater precision in speech emotion detection is not the finish and it is worth exploring any new network architecture or algorithms that can learn more general characteristics or can train a superior predictive model. 2D CNN LSTM network achieves 95.33 percent and 95.89 percent recognition accuracies on Berlin EmoDB of speaker-dependent and speaker-independent experiments, respectively, which compare favorably with the 91.6 percent and 92.9 percent accuracy of conventional approaches; and 89.16 percent and 52.14 percent yield-recognition accuracies on the IEMOCAP speaker database, that is far better than the accuracy achieved by DBN and CNN of 73.78 percent and 40.02 percent. Solanki & Pandey (2019) proposed a deep convolution neural network architecture is used to identify instruments in polyphonic music. The network is focused on fixed-length music and calculates an arbitrary number of instruments. The study finds an exceptional outcome of 92.80 percent precision after the 60 epochs. Used neural network convolution of eight layers for instrument recognition. It is based on audio data from a variable-length audio signal. Recognition of the music instrument using a deep convolution neural network is done in the proposed work. The machine receives feedback in the sampled audio signal form that further transforms the form of the Mel spectrogram. Conducted using eight layers of deep neural network convolution. The activation function of ReLu was found to perform better in this work than in others. Lalitha et al. (2019) focused on investigating the productive efficiency of emotion recognition features of perceptual dependent expression. The Perceptual-based speech patterns bear useful speaker knowledge. DNN with three hidden layers resulted in classifying seven emotions of the Berlin speech emotion corpus. The obtained emotions are equivalent to state-of-the-art approaches. Many of the speech emotion companies have imbalanced datasets which is one of the factors for decreasing success in emotion recognition. The potential course of study may be to experiment to perceptual speech features on multi-corpus-acted and natural datasets. Other deep architectures may be implemented for distinguishing emotions with perceptual features. It may further pursue multimodal emotion recognition and cross-corpus emotion recognition for Cross-Corpus emotion identification.

Nagajyothi & Siddaiah (2018) The article, an Airport inquiry framework based on ASR is presented. The framework for the Telugu language has been natively developed. Based on the most commonly asked questions in an airport inquiry, the database is established. CNN has been used for training and analyzing the database due to its high efficiency. The key trait of weight connectivity, local connectivity, and polling outcome is a framework by preparation, resulting in superior test results. In contrast to standard methods, studies undertaken on wideband speech signals result in substantial changes in the system's efficiency. Zhang et al. (2018) introduced a mechanism for automatic modulation detection to detect radio signals in a communication environment. They introduced a deep CNN network and a deep LSTM network for radio signal Modulation detection. Also, they suggested the representation of modulated signals by IQ-FOC data, where raw IQ data and signal Fourth Order Cumulates are combined. In future work, they planned to operate a wireless communication device to use realistic signals to capture and construct our dataset. The dataset used is generated by GNU radio with a hierarchical block of the GNU Radio Dynamic Channel Model. The data representation presented helps our CNN and LSTM models to make 8% changes in our research dataset. The accuracy of recognition of all deep learning-based recognition algorithms is limited in extremely noisy scenarios. In highly noisy scenarios, the current study explores how to increase the precision of identification can be increased. Zhang et al. (2018) discussed how the difference in speech signals can be bridged by using Deep Convolutional Neural Network (DCNN). It used log Mel-spectrogram channels as the DCNN input, analogous to the RGB image representation. The Discriminant Temporal Pyramid Matching (DTPM) technique aggregates the studied segment-level characteristics. DTPM combines temporal pyramid matching and optimal Lp-norm pooling followed by the linear SVM for emotion classification. The positive findings of their DCNN model and the DTPM technique are shown by experimental results on four public datasets such as EMO-DB, RML, ENTERASE05, BAUM-1s i.e., concluding that further fine-tuning of the target emotional speech datasets greatly promotes the efficiency of appreciation. Consequently, comparing to the state of the arts, their approach will produce promising results.

Now we will discuss the comparison among twelve recent researches on speech recognition, as shown in the below table 1.

Table 1: Summary of Literature review related to speech recognition.

Ref	Dataset	Method	Record & frequencies	WER	Time Rate	Language	Accuracy
(Anvarjon et al., 2020)	Speech-Spectrogram EMOCAP EMO-DB	CNN SER	16 kHz	1.8 Validation loss IEMOCAP 2.0 Validation loss EMO-DB	3120 s for IEMOCAP 1260 s for EMO-DB	-	77.01% for IEMOCAP 92.02% for EMO-DB
(Adebowale et al., 2020)	web pages (Frame, Text, and image)	CNN LSTM IPDS	10,000 images	The report showed zero errors	25 s. 173 m, 10s	Natural language	92.55% for CNN 92.79% for LSTM 93.28% for IPDS
(Bingol & Aydogmus, 2020)	speech word (KUKA KR)	p-DNN	48 kHz	78.50%	1.00s	Turkish	90.37%
(Fantaye et al., 2020)	telephone speech, scripted recordings, and far-field recordings	Kneser–Ney SGD RMSprop	-	WER reductions of 0.97–22.96% for all target languages	10-h	25 languages	0.21–15.59%, 0.1–6.99%, and 4.10–24.92%) 0.5–42.79% 1.33–23.8%
(Zoughi et al., 2020)	TIMIT, FARSDAT, Switchboard, and CallHome MNIST	AMRes back-propagation algorithm	4.0 GHZ	8% absolute-error Error-rate over CD-DNN&DNN-CNN by 7.1% and 4.4% on Switchboard (Hub500 SWB)	3.5s	Bigram language	2.3% and 0.9% over CNN-RNN and RNN LSTM on CallHome (Hub500 CH)
(Liu et al., 2019)	CHiME-5	SGD	-	2.1% WER		-	73.78%
(Passricha & Aggarwal, 2019)	speech signal acoustic data	CNN, CNN-BLSTM, DNN, LSTM-RNN.	-	1.1% WER-of-each-method CNN 18.9 CNN-BLSTM 17.8 DNN 19.8 RNN 19.3	1.6s	English	5.8% over best-performing CNN and 10% over a DNN
(Métais et al., 2019)	OSAC corpus	Adam CRF k-means clustering	6, 532 patients 2.2 million records 1.1 million records,	around 65% 5% mean-squared-error than-when using LS.	3.5 – 38 min	Arabic	79.500 for GloVe-CNN 83.000 for GloVe-LSTM
(Huang et al., 2019)	NNIME (The NTHU-NTUA Chinese interactive multimodal emotion corpus)	CNN	recording of 44 subjects who	-	silence threshold (dB), minimum silence interval (second), and minimum sounding interval (second).	Chinese	52.00%
(Ghosal & Kolekar, 2018)	GTZAN-Music	GMM K-NN SVM	22050Hz	-	Each tracks being 30s	-	94.2%
(Solanki & Pandey, 2019)	Audio Signal IMRAS dataset is used	CNN	6705 western musical recording data	-	985 min	-	92.8%
(Korvel et al., 2018)	111-word experiment	CNN	69,075 utterances	Spectrum=0.0951 Cepstrum=1.4278 Chromagram=0.9845 Mel-scale cepstrum=0.4019	The time of the word “vyriausyb˙e” (“government”) =0.8s	English	0.99 for spectrum 0.91 for mel-scale-Cepstrum 0.76 for-chromagram, 0.64 for-cepstrum

7. Comparison and Discussion

This paper reviews some current CNN algorithms on speech recognition. This indicates that all the major categories are selected. Table 1 shown above corresponds the explanation of each column. In each field, the following insights can be found. Each paper used a different dataset such as (Anvarjon et al., 2020) which used Speech-Spectrogram. (Bingol & Aydogmus, 2020) used speech word dataset and (Korvel et al., 2018), utilized a 111-word dataset. As for their methodologies, several tools have been developed by utilizing the machine learning algorithms (and specifically deep learning algorithms) to conduct these experiments. Some of the researchers such as (Zoughi et al., 2020) have revealed that the frequencies in their work are recorded at 4.0 GHZ. (Métais et al., 2019), at the same time, reveals that their recorded dataset was recorded from 6, 532 patients with 1.1 million records and approximately 2.2 million utterances. Therefore, it appears to us that the researches have extracted a lot of results from the data used by analyzing it in different methods. From the Table 1, it's obvious that CNN appeared as the best method among all the methods to reduce WER, such as (Fantaye et al., 2020), where WER reduced by 0.97–22.96% for all the target languages. As we know, each work needs time, for that reason, it is better to reduce Time Rate (TR) as much as we can. In this review, many researchers scheduled time for each and used different methods to change the time. For instance, (Adebowale et al., 2020) used three methodologies (CNN TR=25 s, LSTM TR=173 m and IPDS TR=10s). The same researcher used different languages such as English, Arabic, and Turkish and his results showed impressive results in accuracy.

8. Conclusions

In this study, we have studied some papers about speech recognition based on CNN. Then we have reviewed a number of works that were experimented in the field and a comparison among them was discussed. It has been noticed that the results of the speech recognition based on CNN are much better than DNN. CNN significantly decreases the complexity of the model and has less WER. We expect that the findings of this analysis will enable prospective scholars to recognize new and important research areas that have not yet been explored, as well as to highlight some of the weaknesses in current studies.

References

Abdel-Hamid, O., Mohamed, A., 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.

Abdel-Hamid, O., Mohamed, A., Jiang, H., & Penn, G. (2012). *APPLYING CONVOLUTIONAL NEURAL NETWORKS CONCEPTS TO HYBRID NN-HMM MODEL FOR SPEECH RECOGNITION*. 4.

Abdel-Hamid, O., Mohamed, A., Jiang, H., & Penn, G. (2012). Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4277–4280.

Abdulazeez, A. M., Sulaiman, M. A., & Qader, D. (2020). *Evaluating Data Mining Classification Methods Performance in Internet of Things Applications*. *1*(2), 15.

Abdulazeez, A., Salim, B., Zeebaree, D., & Doghramachi, D. (2020). *Comparison of VPN Protocols at Network Layer Focusing on Wire Guard Protocol*.

Abdulazeeza, A. M., Nahmatwllab, L. L., & Qader, D. (2020). Pipelined Parallel Processing Implementation based on Distributed Memory Systems. *International Journal of Innovation*, *13*(7), 12.

Abdulqader, D. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). *Machine Learning Supervised Algorithms of Gene Selection: A Review*. *62*(03), 13.

Adebowale, M. A., Lwin, K. T., & Hossain, M. A. (2020). Intelligent phishing detection scheme using deep learning algorithms. *Journal of Enterprise Information Management*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/JEIM-01-2020-0036

Adeen, I. M. N., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). *Systematic Review of Unsupervised Genomic Clustering Algorithms Techniques for High Dimensional Datasets*.

Ahmed, J. A., & Brifcani, A. M. A. (2015). A new internal architecture based on feature selection for holonic manufacturing system. *International Journal of Mechanical, Aerospace, Industrial, Mechatronic and Manufacturing Engineering*, *2*(8), 1431.

Ahmed, O., & Brifcani, A. (2019). Gene Expression Classification Based on Deep Learning. *2019 4th Scientific International Conference Najaf (SICN)*, 145–149.

Ahmed, O. M., & Abduallah, W. M. (2017). A Review on Recent Steganography Techniques in Cloud Computing. *Academic Journal of Nawroz University*, *6*(3), 106–111.

Anasuya, M. A., & Katti, S. K. (2009). Speech recognition by machine: A review. *International Journal of Computer Science and Information Security*, *6*(3), 181–205.

Anuradha, B., & Reddy, V. V. (2008). ANN for classification of cardiac arrhythmias. *ARPN Journal of Engineering and Applied Sciences*, *3*(3), 1–6.

Anvarjon, T., Mustaqeem, & Kwon, S. (2020). Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features. *Sensors*, *20*(18), 5212. https://doi.org/10.3390/s20185212

Bargarai, F., Abdulazeez, A., Tiryaki, V., & Zeebaree, D. (2020). *Management of Wireless Communication Systems Using Artificial Intelligence-Based Software Defined Radio*.

Bingol, M. C., & Aydogmus, O. (2020). Performing predefined tasks using the human–robot interaction on speech recognition for an industrial robot. *Engineering Applications of Artificial Intelligence*, *95*, 103903. https://doi.org/10.1016/j.engappai.2020.103903

Bird, J. J., Wanner, E., Ekárt, A., & Faria, D. R. (2020). Optimisation of phonetic aware speech recognition through multi-objective evolutionary algorithms. *Expert Systems with Applications*, *153*, 113402. https://doi.org/10.1016/j.eswa.2020.113402

Bocklet, T., Maier, A., Bauer, J. G., Burkhardt, F., & Noth, E. (2008). Age and gender recognition for telephone applications based on gmm supervectors and support vector machines. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1605–1608.

Chaudhary, D., & Vasuja, E. R. (2019). *A Review on Various Algorithms used in Machine Learning*.

Deepak, S., & Ameer, P. M. (2019). Brain tumor classification using deep CNN features via transfer learning. *Computers in Biology and Medicine*, *111*, 103345. https://doi.org/10.1016/j.compbiomed.2019.103345

Fantaye, T. G., Yu, J., & Hailu, T. T. (2020). Advanced Convolutional Neural Network-Based Hybrid Acoustic Models for Low-Resource Speech Recognition. *Computers*, *9*(2), 36. https://doi.org/10.3390/computers9020036

Furui, S. (1991). Speaker-dependent-feature extraction, recognition and processing techniques. *Speech Communication*, *10*(5–6), 505–520.

G., S., R., V., & K.P., S. (2018). Diabetes detection using deep learning algorithms. *ICT Express*, *4*(4), 243–246. https://doi.org/10.1016/j.icte.2018.10.005

Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A review on speech recognition technique. *International Journal of Computer Applications*, *10*(3), 16–24.

Ghosal, D., & Kolekar, M. H. (2018). Music Genre Recognition Using Deep Neural Networks and Transfer Learning. *Interspeech 2018*, 2087–2091. https://doi.org/10.21437/Interspeech.2018-2045

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649.

Hasan, D. A., & Abdulazeez, A. M. (2020). A Modified Convolutional Neural Networks Model for Medical Image Segmentation. *Learning*, *20*, 22.

Huang, K.-Y., Wu, C.-H., Hong, Q.-B., Su, M.-H., & Chen, Y.-H. (2019). Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5866–5870. https://doi.org/10.1109/ICASSP.2019.8682283

Jahwar, A. F., & Abdulazeez, A. M. (2020). META-HEURISTIC ALGORITHMS FOR K-MEANS CLUSTERING: A REVIEW. *PalArch's Journal of Archaeology of Egypt/Egyptology*, *17*(7), 12002–12020.

Jansson, P. (2018). *Single-word speech recognition with Convolutional Neural Networks on raw waveforms*. 31.

Jing, W., Jiang, T., Zhang, X., & Zhu, L. (2012). *The optimisation of speech recognition based on convolutional neural network*. 10.

Kavi B. Obaid, Zeebaree, S. R. M., & Ahmed, O. M. (2020). *Deep Learning Models Based on Image Classification: A Review*. https://doi.org/10.5281/ZENODO.4108433

Klevans, R. L., & Rodman, R. D. (1997). *Voice recognition*. Artech House, Inc.

Korvel, G., Treigys, P., Tamulevicus, G., Bernataviciene, J., & Kostek, B. (2018). Analysis of 2D Feature Spaces for Deep Learning-Based Speech Recognition. *Journal of the Audio Engineering Society*, *66*(12), 1072–1081. https://doi.org/10.17743/jaes.2018.0066

Laine, U. (2017). *Analytic Filter Bank for Speech Analysis, Feature Extraction and Perceptual Studies* (p. 453). https://doi.org/10.21437/Interspeech.2017-1232

Lalitha, S., Tripathi, S., & Gupta, D. (2019). Enhanced speech emotion detection using deep neural networks. *International Journal of Speech Technology*, *22*(3), 497–510. https://doi.org/10.1007/s10772-018-09572-8

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

LeCun, Y., Bengio, Y., & Laboratories, T. B. (1995). *Convolutional Networks for Images, Speech, and Time-Series*. 15.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

LeCun, Y., Huang, F. J., & Bottou, L. (2004a). Learning methods for generic object recognition with invariance to pose and lighting. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, *2*, II–104.

LeCun, Y., Huang, F. J., & Bottou, L. (2004b). Learning methods for generic object recognition with invariance to pose and lighting. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, *2*, II–104.

Liu, C.-R., Qu, D., & Yang, X.-K. (2019). *Long Short Term Memory Networks Weighted Prediction Error for Far-Field Speech Recognition*. 4.

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, *1*(4), 140–147. https://doi.org/10.38094/jastt1457

Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network. *IEEE Access*, *7*, 125868–125881. https://doi.org/10.1109/ACCESS.2019.2938007

Métais, E., Meziane, F., Vadera, S., Sugumaran, V., & Saraee, M. (Eds.). (2019). *Natural Language Processing and Information Systems: 24th International Conference on Applications of Natural Language to Information Systems, NLDB 2019, Salford, UK, June 26–28, 2019, Proceedings* (Vol. 11608). Springer International Publishing. https://doi.org/10.1007/978-3-030-23281-8

Morgan, N. (2011). Deep and wide: Multiple layers in automatic speech recognition. *Ieee Transactions on Audio, Speech, and Language Processing*, *20*(1), 7–13.

Muhammad, M. A., Zeebaree, D. Q., Abdulazeez, A. M., Saeed, J. N., & Zebari, A. (2020). *A Review on Region of Interest Segmentation Based on Clustering Techniques for Breast Cancer Ultrasound Images*. *01*(03), 14.

Nagajyothi, D., & Siddaiah, P. (2018). Speech Recognition Using Convolutional Neural Networks. *International Journal of Engineering & Technology*, *7*(4.6), 133. https://doi.org/10.14419/ijet.v7i4.6.20449

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access*, *7*, 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

Nweke, H. F., Teh, Y. W., Al-garadi, M. A., & Alo, U. R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications*, *105*, 233–261. https://doi.org/10.1016/j.eswa.2018.03.056

Omar, N., Abdulazeez, A. M., Sengur, A., & Al-Ali, S. G. S. (2020). Fused faster RCNNs for efficient detection of the license plates. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(2), 974–982.

Ordóñez, F. J., & Roggen, D. (2016). Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, *16*(1), 115.

Palaz, D., & Collobert, R. (2015). *Analysis of cnn-based speech recognition system using raw speech as input*. Idiap.

Palaz, D., Magimai.-Doss, M., & Collobert, R. (2015). Convolutional Neural Networks-based continuous speech recognition using raw speech signal. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4295–4299. https://doi.org/10.1109/ICASSP.2015.7178781

Passricha, V., & Aggarwal, R. K. (2019). A Hybrid of Deep CNN and Bidirectional LSTM for Automatic Speech Recognition. *Journal of Intelligent Systems*, *29*(1), 1261–1274. https://doi.org/10.1515/jisys-2018-0372

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. https://doi.org/10.1109/TASLP.2016.2602884

Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *4*, IV-4072-IV–4075.

Sainath, T. N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A., Dahl, G., & Ramabhadran, B. (2015). Deep convolutional neural networks for large-scale speech tasks. *Neural Networks*, *64*, 39–48.

Sajjad, M., Khan, S., Muhammad, K., Wu, W., Ullah, A., & Baik, S. W. (2019). Multi-grade brain tumor classification using deep CNN with extensive data augmentation. *Journal of Computational Science*, *30*, 174–182. https://doi.org/10.1016/j.jocs.2018.12.003

Sethi, J., & Mittal, M. (2019). Ambient air quality estimation using supervised learning techniques. *EAI Endorsed Transactions on Scalable Information Systems*, *6*(22).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.

Solanki, A., & Pandey, S. (2019). Music instrument recognition using deep convolutional neural networks. *International Journal of Information Technology*. https://doi.org/10.1007/s41870-019-00285-y

Song, Z. (2020). English speech recognition based on deep learning with multiple features. *Computing*, *102*(3), 663–682. https://doi.org/10.1007/s00607-019-00753-0

Sornam, M., Muthusubash, K., & Vanitha, V. (2017). A Survey on Image Classification and Activity Recognition using Deep Convolutional Neural Network Architecture. *2017 Ninth International Conference on Advanced Computing (ICoAC)*, 121–126. https://doi.org/10.1109/ICoAC.2017.8441512

Swietojanski, P., Ghoshal, A., & Renals, S. (2014). Convolutional Neural Networks for Distant Speech Recognition. *IEEE Signal Processing Letters*, *21*(9), 1120–1124. https://doi.org/10.1109/LSP.2014.2325781

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

V.Chapaneri, S. (2012). Spoken Digits Recognition using Weighted MFCC and Improved Features for Dynamic Time Warping. *International Journal of Computer Applications*, *40*(3), 6–12. https://doi.org/10.5120/5022-7167

Vogt, T., & André, E. (2006). Improving Automatic Emotion Recognition from Speech via Gender Differentiaion. *LREC*, 1123–1126.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56–70. https://doi.org/10.38094/jastt1224

Zeebaree, D. Q., Haron, H., & Abdulazeez, A. M. (2018). Gene Selection and Classification of Microarray Data Using Convolutional Neural Network. *2018 International Conference on Advanced Science and Engineering (ICOASE)*, 145–150. https://doi.org/10.1109/ICOASE.2018.8548836

Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. (2019a). Machine learning and Region Growing for Breast Cancer Segmentation. *2019 International Conference on Advanced Science and Engineering (ICOASE)*, 88–93. https://doi.org/10.1109/ICOASE.2019.8723832

Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. (2019b). Trainable Model Based on New Uniform LBP Feature to Identify the Risk of the Breast Cancer. *2019 International Conference on Advanced Science and Engineering (ICOASE)*, 106–111. https://doi.org/10.1109/ICOASE.2019.8723827

Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zeebaree, S. R. (2017). Combination of K-means clustering with Genetic Algorithm: A review. *International Journal of Applied Engineering Research*, *12*(24), 14238–14245.

Zeebaree, S. R., Haji, L. M., Rashid, I., Zebari, R. R., Ahmed, O. M., Jacksi, K., & Shukur, H. M. (2020). *Multicomputer Multicore System Influence on Maximum Multi-Processes Execution Time*.

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, *26*(7), 3142–3155.

Zhang, M., Zeng, Y., Han, Z., & Gong, Y. (2018). Automatic Modulation Recognition Using Deep Learning Architectures. *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 1–5. https://doi.org/10.1109/SPAWC.2018.8446021

Zhang, S., Zhang, S., Huang, T., & Gao, W. (2018). Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Transactions on Multimedia*, *20*(6), 1576–1590. https://doi.org/10.1109/TMM.2017.2766843

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035

Zoughi, T., Homayounpour, M. M., & Deypir, M. (2020). Adaptive windows multiple deep residual networks for speech recognition. *Expert Systems with Applications*, *139*, 112840. https://doi.org/10.1016/j.eswa.2019.112840

Cite this article:

Kazheen Ismael Taher & Adnan Mohsin Abdulazeez (2021). Deep Learning Convolutional Neural Network for Speech Recognition: A Review. ***International Journal of Science and Business***, *5*(3), 1-14. doi: https://doi.org/10.5281/zenodo.4475361

Retrieved from http://ijsab.com/wp-content/uploads/681.pdf

Published by

IJSAB
International