Speech Recognition Using Deep Neural Networks: A Systematic Review

Article *in* IEEE Access · February 2019

DOI: 10.1109/ACCESS.2019.2896880

5 authors, including:

Ali Bou Nassif
University of Sharjah
138 PUBLICATIONS **2,361** CITATIONS
SEE PROFILE

Ismail Shahin
University of Sharjah
101 PUBLICATIONS **1,131** CITATIONS
SEE PROFILE

Imtinan Attili
University of Sharjah
7 PUBLICATIONS **344** CITATIONS
SEE PROFILE

Mohammad Azzeh
Princess Sumaya University for Technology
71 PUBLICATIONS **1,418** CITATIONS
SEE PROFILE

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 1, 2019, accepted January 24, 2019, date of publication February 1, 2019, date of current version February 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2896880

Speech Recognition Using Deep Neural Networks: A Systematic Review

ALI BOU NASSIF[1], ISMAIL SHAHIN[1], IMTINAN ATTILI[1], MOHAMMAD AZZEH[2], AND KHALED SHAALAN[3]

[1]Department of Electrical and Computer Engineering, University of Sharjah, Sharjah 27272, United Arab Emirates
[2]Department of Software Engineering, Applied Science Private University, Amman 163, Jordan
[3]Faculty of Engineering and IT, The British University in Dubai, Dubai 345015, United Arab Emirates

Corresponding author: Ali Bou Nassif (anassif@sharjah.ac.ae)

This work was supported by the University of Sharjah through the Competitive Research Project "Emotion Recognition in each of Stressful and Emotional Talking Environments Using Artificial Models" under Grant 1602040348-P. The work of M. Azzeh was supported by the Applied Science Private University, Amman, Jordan.

ABSTRACT Over the past decades, a tremendous amount of research has been done on the use of machine learning for speech processing applications, especially speech recognition. However, in the past few years, research has focused on utilizing deep learning for speech-related applications. This new area of machine learning has yielded far better results when compared to others in a variety of applications including speech, and thus became a very attractive area of research. This paper provides a thorough examination of the different studies that have been conducted since 2006, when deep learning first arose as a new area of machine learning, for speech applications. A thorough statistical analysis is provided in this review which was conducted by extracting specific information from 174 papers published between the years 2006 and 2018. The results provided in this paper shed light on the trends of research in this area as well as bring focus to new research topics.

INDEX TERMS Speech recognition, deep neural network, systematic review.

I. INTRODUCTION

Since the last decade, deep learning has arisen as a new attractive area of machine learning, and ever since has been examined and utilized in a range of different research topics [1]. Deep learning consists of a multiple of machine learning algorithms fed with inputs in the form of multiple layered models. These models are usually neural networks consisting of different levels of non-linear operations. The machine learning algorithms attempt to learn from these deep neural networks by extracting specific features and information [2]. Prior to 2006, searching deep architecture inputs was not a predictable straight forward task; however, the development of deep learning algorithms helped resolve this issue and simplified the process of searching the parameter space of deep architectures [2]. Deep learning models can also operate as a greedy layerwise unsupervised pre-training. This means that it will learn hierarchy from extracted features from each layer at a time. Feature learning is achieved by training each layer with an unsupervised learning algorithm, which takes the features extracted from the previous layer and uses it as an input for the next layer. Thus, feature learning will attempt to learn the transformation of the previously learned features at each new layer. Each iteration feature learning adds one layer of weights to a deep neural network. The resulted layers with learned weights can eventually be loaded to initialize a deep supervised predictor [2], [3]. Using deep architectures has proven to be more efficient in representing non-linear functions in comparison to shallower architectures. Studies have shown that fewer parameters are required to represent a certain non-linear function in a deep architecture in comparison with the large number of parameters needed to represent the same function in a shallower architecture. This shows that deeper architectures are more efficient from a statistical point of view [2], [3].

Deep learning algorithms have been mostly used to further enhance the capabilities of computers so that it understands what humans can do, which includes speech recognition. Speech in particular, being the main method of communication among human beings, received much interest for the

The associate editor coordinating the review of this manuscript and approving it for publication was Malik Jahan Khan.

past five decades right from the introduction of artificial intelligence [4], [5]. Therefore, it is only natural that one of the early applications of deep learning was speech, and up to this day a huge number of research papers have been published in the use of deep learning for speech related applications specifically speech recognition [4], [5], [6], [7]. The conventional speech recognition systems are based on representing speech signals using Gaussian Mixture Models (GMMs) that are based on hidden Markov models (HMMs). This is due to the fact that a speech signal can be considered as a piecewise stationary signal or in other terms a short time stationary signal. In this short time scale, the speech signal can be approximated as a stationary process, thus it can be thought of as a Markov model for many stochastic processes. Each HMM uses a mixture of Gaussian to model a spectral representation of the sound wave. This type of systems is considered simple in design and practical in use. However, they are considered statistically inefficient for modeling non-linear or near non-linear functions [4], [6]. Opposite to HMMs, neural networks permit discriminative training in a much efficient manner. However, it works better for short time signals such as isolated words, when it comes to continuous speech signals it is rarely successful. This is due to its inability to model temporal dependencies for continuous signals. Thus, one solution is using neural networks as a pre-processing e.g. feature transformation, dimensionality reduction for the HMM based recognition [3]. There are many examples that prove that using deep neural networks yield better results than classical models. In 2012, Microsoft released the newest version of their Microsoft Audio Video Indexing Service (MAVIS) which is a speech system based on deep learning. Their final results clearly showed that the word error rate (WER) reduced on four major benchmarks by 30% compared to the state-of-the-art models based on Gaussian mixtures [2].

This systematic literature review (SLR) follows Kitchenham and Charters guidelines [8] and was focusing on identifying the different research papers that have been published from 2006 to 2018 in the area of deep neural networks in speech-related applications. Such applications include: automatic speech recognition, emotional speech recognition, speaker identification and speech enhancement among others. The identified number of papers was originally 230; however, after applying the inclusion/exclusion criteria, only 174 papers were included in the study. The research questions were answered by extracting proper information from these 174 papers and then forming a statistical representation using tables and figures. The results presented are intended to show the trend of research done in this area over the past years as well as bring focus to new interesting research topics.

This paper summarizes the related work in Section 2, whereas information regarding the background such as speech recognition and deep neural networks is presented in Section 3. Section 4 summarizes the methodology used to conduct this review. Section 5 demonstrates the results, where Section 6 concludes the paper.

II. RELATED WORK

Some surveys have been conducted in the area of speech recognition. For instance, Morgan [9] conducted a review in the area of speech recognition assisted with discriminatively trained feed-forward networks. The main focus of the review was to shed the light on papers that employ multiple layers of processing prior to the hidden Markov model based decoding of word sequences. Throughout the paper, some of the methods that incorporate multiple layers of computation for the purpose of either providing large gains for noisy speech in small vocabulary tasks or significant gains for high Signal-to-Noise Ratio (SNR) speech on large vocabulary tasks were described. Moreover, a detailed description was provided about the methods with structures that incorporate a large number of layers (the depth) and multiple streams using Multilayer Perceptrons (MLPs) with a large number of hidden layers. This review paper eventually concluded that even though the deep processing structures are capable of providing improvements in this genre, choice of features and the structure with which they are incorporated, including layer width, can also be significant factors.

Hinton *et al.* [10], presents an overview on the use of deep neural networks that incorporate many number of hidden layers that are trained using some of the new techniques. The overview summarizes the findings of four different research groups that collaborated to reveal the advantage of a feed-forward neural network that has quite a few frames of coefficients as an input and produces subsequent probabilities over HMM states as an output. This technique was studied as an alternative to using the traditional HMMs and GMMs for acoustic modeling in speech recognition. The collected results have shown that deep neural networks that incorporate many hidden layers and are trained by new techniques outperform GMMs - HMMs on a variety of speech recognition benchmarks, by sometimes a large margin.

Deng *et al.* [11] presented an overview summary on the papers that were part of the session at ICASSP- 2013, entitled ''New Types of Deep Neural Network Learning for Speech Recognition and Related Applications,'' which was organized by the authors. In addition to that, the paper presented the history of the development of the deep neural networks for acoustic models for speech recognition. The overview summary focused on the different ways that can be utilized to improve deep learning, which was classified into five different categories: enhanced types of network architecture and activation functions, enhanced optimization methods, enhanced ways of determining the deep neural networks parameters and finally enhanced ways of leveraging a number of languages at the same time. The overview revealed the rapid continues progress in the acoustic models that use deep neural networks which can be seen on several fronts when compared to those based on GMMs. The paper also revealed that these acoustic models can also be applicable and enhance performance in other signal processing applications, and not only speech recognition.

Deng *et al.* [12] conducted a summary on the work done by Microsoft since the year 2009 in the area of speech using deep learning. The paper focused on more recent advances which helped shed some light on the different capabilities as well as limitations of deep learning in the area of speech recognition. This was done by providing samples of the recent experiments carried by Microsoft for advancing speech related applications through the use of deep learning methods. Speech related applications included features extraction, modeling language, acoustic models, understanding speech as well as dialogue estimation. Experimental results have shown clearly that the speech spectrogram features are more advanced to MFCC with deep neural networks compared to the traditional practice using GMMs - HMMs. This paper also shows that improvements should be done on the architecture of deep neural networks in order to improve further the features of acoustic measurements

Li *et al.* [13] presented the basics of the state of the art solutions for automatic spoken language recognition for both, computational and phonological perspectives. Huge progress was achieved in recent years in the area of spoken language recognition which was mostly directed by breakthroughs in relevant signal processing areas such as pattern recognition and cognitive science. Several main aspects relevant to language recognition was discussed such as language characterization, modeling methods, as well as system development techniques. Findings clearly indicate that even though this area has hugely developed in the past years, it is still far from the perfect, especially when it comes to language characterization. In addition, this paper provides an overview on the current research trends and future directions which was carried using the language recognition evaluation (LRE) which is developed by the National Institute of Standards and Technology (NIST).

Li *et al.* [14] provided an overview on modern noise robust techniques for automatic speech recognition developed over the past three decades. More emphasis was given on the techniques that have proven successful over the years and are likely to maintain and further expand in their applicability in the future. The examined techniques were categorized and evaluated using five different criteria, which are: using former knowledge about the acoustic environment distortion, model domain processing versus feature domain processing, using specific environment distortion models, uncertainty processing versus predetermined processing and finally using acoustic models trained by the same model adaptation process used in the testing stage. This study helps the reader differentiate between the different noise-robust techniques, as well as provides a comprehensive insight on the performance complex tradeoffs that should be taken into account when selecting between the available techniques.

This systematic review is different from the above as we are presenting a comprehensive study on the use of deep neural networks in the area of speech recognition. We first provided an overview on machine learning, its categories and definitions, with emphasis on deep learning which is our main concern in the paper. We then provided an overview on speech recognition as well, its different applications, features and types. This provides the reader with a suitable comprehensive theoretical background in order to fully grasp the topic presented which is the use of deep neural networks in the area of speech. Moreover, in order to carry the systematic literature review, 174 papers were used which were published on the span of 12 years between 2006 and 2018. The information extracted from the above papers includes the following:

1) The different types of speech identified.
2) The types of database used to train and test the algorithm.
3) The different languages used to train and test the algorithm.
4) The type of environment (noisy, emotional or neutral) each study used.
5) The different types of features extracted from speech.
6) The type of publication (journal, conference or workshop).
7) The specific name of the conference or journal that published the paper.
8) The distribution of papers over the years.

This extracted information helped identify research patterns over the past decade in the use of deep neural networks in the area of speech. The developed statistics throughout our study helped shed light on research gaps and shortcomings, as well as the past and future direction of studies in this area. This will hopefully help future researchers in identifying new research topics in this area as well as try to find suitable solutions to amend the gaps and shortcomings in the already existing research.

III. BACKGROUND

A. SPEECH SIGNALS

Speech signals can provide us with different kinds of information. Such kinds of information are:

- Speech recognition, which gives information about the content of speech signals.
- Speaker recognition that carries information about the speaker identity.
- Emotion recognition, which delivers information about the speaker's emotional state.
- Health recognition, which offers information on the patient's health status.
- Language recognition, that yields information of the spoken language.
- Accent recognition, which produces information about the speaker accent.
- Age recognition that supplies information about the speaker age.
- Gender recognition, which carries information about the speaker gender.

Automatic speech recognition is the capability of a machine or computer to recognize the content of words and phrases in an uttered language and transform them to

a machine-understandable format. Speech recognition can be used in many applications. Such applications appear in: dictating computers instead of typing, spaceships when the extremities are busy, helping handicapped people, smart homes, and many others.

Automatic speaker recognition can be defined as the process of recognizing the unknown speaker on the basis of the information embedded in his/her speech signal using machine (computer). Speaker recognition is divided into two parts: speaker identification and speaker verification (authentication). The process of determining to which of the registered speaker a given utterance corresponds to is termed as the speaker identification part. This part can be used in public facilities or for the media. These cases comprise, but not limited to, district or other government institutions, calls to radio stations, insurance agencies, or documented conversations [15], [16]. Speaker verification part is described as the procedure of admitting or discarding the claimed speaker identity. The applications of this part comprise the use of voice as a focal factor to authorize the claimed speaker identity. Business relations using a telephone network, dataset access facilities, security control for private information areas, remote access to computers, and intelligent health care systems are some of application areas of this branch [17], [18]. Emotion cue-based speaker recognition becomes one of the research fields for human-machine interaction or affective computing that has recently earned accelerating attentions due to the broad diversity of applications that profit from this up-to-date technology [19]. The main inspiration arises from the demand to mature a human-machine (computer) interface that is more adaptive to a user's identity. The major role of the intelligent human-machine interaction is to enable computers with the capability of affective computing so that computers can recognize the user for distinct applications. The low speaker recognition performance in emotional talking environments is considered as one of the most challenging issues of this field [20]–[23].

Emotion recognition by machine can be defined as the task of recognizing the unknown emotion based on information inserted in speech signals. Emotion recognition field is divided into emotion identification and emotion verification branches. In the first branch, the unknown emotion is identified as the emotion whose model best matches the input speech signal. In the second branch, the goal is to determine whether a given emotion belongs to a specific known emotion or to some other unknown emotions. The applications of emotion recognition clearly appear in [24]–[26] perceiving the speaker emotional status in telephone call center conversations and supplying feedback to an operator for observing purposes, categorizing voice mail messages based on emotions articulated by callers, and recognizing the mistrusted individuals who produced emotional voice (e.g. happiness) in emotional talking environments.

Automatic health recognition is defined as using the patient's voice to provide information on the patient's health status. Automatic health recognition can be used in intelligent health care systems [17], [18]. These systems can be utilized in hospitals which include computerized health categorization and evaluation methods [17]. These systems can also be used in the pathological voice assessment (functional dysphonic voices) [18]. The dysphonic voice can be hoarse or extremely breathy, harsh, or rough. In addition, automatic health recognition systems can be used in the diagnosis of Parkinson's disease. In Massachusetts Institute of Technology (MIT), a team led by Max Little conducted some experiments and tests to analyze and evaluate the voice characteristics of patients who had been detected with Parkinson's disease. They found that they could establish a tool to detect such a disease in the individuals' speech patterns.

Language recognition is the problem of determining which natural language given a speech content is in. One of the immense challenges of language recognition systems is to differentiate between closely correlated languages. Similar languages such as Serbian and Croatian or Indonesian and Malay show significant lexical and structural overlap. Hence, the discrimination between each such two languages become challenging for language recognition systems. The applications of automatic language recognition evidently appear in spoken language translation [27], multilingual speech recognition [28], and spoken document retrieval [29].

The task of accent recognition is the recognition of a speaker's regional accent, within a predetermined language, given the acoustic signal alone. The problem of accent recognition has been considered as more challenging than that of language recognition due to the greater similarity between accents of the same language [30]. Accent recognition has wide range of applications in our daily life. Accent recognition helps in Automatic Speech Recognition (ASR) since speakers with diverse accents pronounce some words differently, constantly varying particular phones. Accent recognition also allows us to conclude the speaker's regional origin and ethnicity and consequently to adjust features used in speaker recognition to regional origin.

Age recognition by voice is the process of estimating the speaker's age (e.g. child, young, adult, senior, etc.) using his/her uttered speech signals. Automatic age recognition can be used in security applications, age-restriction applications, and others [31].

Automatic gender recognition is the process of recognizing whether the speaker is a male or female. Generally, automatic gender recognition results in a great accuracy without much effort because the results of such type of recognition is binary (either male or female) [32], [33]. Automatic gender recognition clearly appears in the applications of call centers of some conservative societies, where automatic dialog systems with the ability of recognizing genders are favored over those without such ability.

B. MACHINE LEARNING

Recently data has become very easily obtained through numerous numbers of open sources. Extracting knowledge from data is considered the real challenge. With the use of

FIGURE 1. Different stages of machine learning.

computers and smart software that can perform numerous computations and calculations in seconds, the process of analyzing data has become easier. Moreover, learning from obtained data is also essential as adapting with new inputs is a highly important process that ensures the continuous development of any smart system. For that reason, a lot of attention has been given on the field of machine learning over the past years.

Machine learning is defined as the field of study that provides computers with the ability to learn from input data without being explicitly programmed to do so. The learning process is done iteratively from analyzed data and new input data. This iterative aspect allows computers to identify hidden insights and repeated patterns and use these findings to adapt when exposed to a new data [34]. The different types of data used in this learning process can vary from observations and examples to instructions and direct experience [34]. The gained knowledge will help in producing reliable and repeated results. Thus, we can describe machine learning as a method that learns from past experiences and uses gained knowledge to do better in the future [34], [35]. Figure 1 illustrates the main stages of machine learning.

Machine learning emphasizes on automatically learning and adapting when exposed to data without the need of human intervention. As mentioned earlier, the past affects the future thus machine learning is viewed as programming by example. In order to solve a certain task, rather than explicitly programming the computer to perform that task, we let it come up with its own program based on provided examples from which the computer learns [35]. Using techniques based on data mining and statistical analysis, machine learning allows computers to emulate human learning behavior, reasoning and decision making. Improving machine learning is highly important since without it there will be no hope of one day reaching artificial intelligence. The reason for that is based on the fact that no system can be described as intelligent if it does not have the ability to learn and adapt [35].

Machine learning has gained significant attention specially in recent years and can be found in several applications such as spam filters, web search, credit scoring, fraud detection, pricing prediction, ad placement, drug design, healthcare, transportation and many other applications [36]. Five main techniques of machine learning exist which are: supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning and finally deep learning. Figure 2 shows the different machine learning types and algorithms that will be mentioned in details later.

In general, supervised learning includes training datasets at which data is presented in pairs, an input and its corresponding correct output. An example for that is the actual price of electronic devices provided the features associated with each device by which the price is directly affected. This process helps to train the algorithm and thus develop a model capable of predicting the price of new input devices that were not included in the training dataset [37]. On the other hand, unsupervised leaning attempts to find common points between the inputs in the dataset. This process can be described as clustering of inputs that are greatly correlated under one general label based on their statistical properties [37]. Semi-supervised learning is a combination of the two previously described types as the algorithm is trained using a dataset that contains both labeled and unlabeled input points. This type is used mainly to enhance the performance of the algorithm through the use of both types of inputs [38]. Reinforcement learning depends on the trial and error process to uncover the set of actions that maximizes a cumulative reward metric, which is used to make the algorithm understand whether it's going in the right direction or not [39]. Lastly is deep learning, this type attempts to model the abstractions found in the dataset using a graph with multiple processing layers. These layers aim to mimic the neural network found within our brains [39]. A detailed description of all five types of machine learning is provided in the subsections below.

1) SUPERVISED LEARNING

This type of machine learning is based on using labeled data to train the learning algorithm. The data is described as labeled since it consists of pairs, an input that can be represented by a vector and its corresponding desired output which can be described as a supervisory signal, [40], [41]. The learning mechanism is described as supervised since the correct output is known and the learning algorithm attempts to iteratively predict this output and is corrected to reduce the variation gap between its predicted and the actual output [41]. Analyzing the training data allows the supervised learning algorithm to produce a function that is called a classifier function if the output was discrete, and a regression function if the output was continuous [40]. The learning algorithm

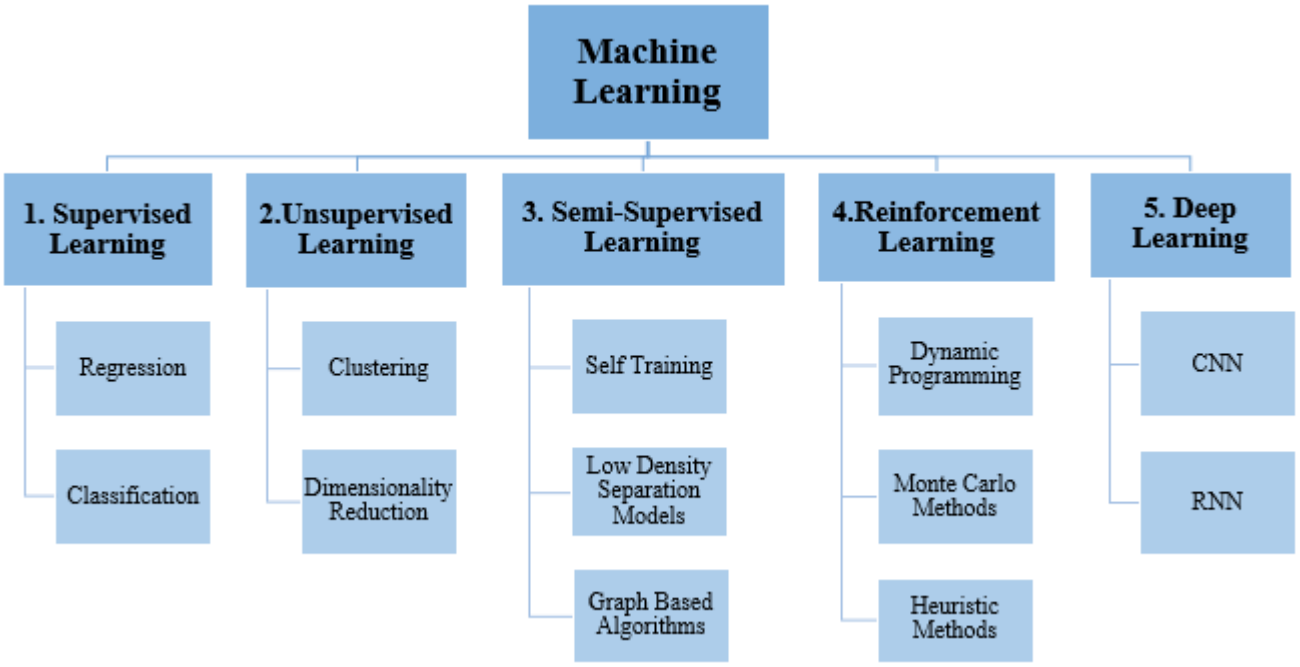

FIGURE 2. Different machine learning types and algorithms.

FIGURE 3. Different stages of the supervised learning.

generalizes detected patterns and features from the training data to a new input data in a reasonable way and thus the produced function predicts the output corresponding to any provided input. Figure 3 illustrates the different stages of the supervised machine learning method.

Regression algorithms (continuous output) and classification algorithms (discrete output) are considered as the main categories of supervised learning. Regression algorithms attempt to uncover the best function that fits points in the training dataset. Regression algorithms include the following main types: linear regression, multiple linear regression and polynomial regression [40]. Classification algorithms, on the other hand, aim to uncover the best fit class for the input data through assigning each input to its correct class. In this case, the output of the predictive function is in the discrete form and its value is one of the different classes available [42].

2) UNSUPERVISED LEARNING

Unlike supervised learning, this method uses an input dataset without any labeled outputs to train the learning algorithm. There is no right or wrong output to each input object and no human intervention to correct or adjust as in supervised learning. Thus, unsupervised learning is more subjective than supervised [43], [44]. The main goal of unsupervised learning is to learn more about the data through identifying the fundamental structure or distribution patterns that is found in the data itself. Learning by itself, the algorithm attempts to represent a particular identified input pattern while reflecting it on the overall structure of input patterns. Thus, the different inputs are clustered into groups based on the features that were extracted from each input object [43]. Figure 4 represents the different stages of unsupervised machine learning method.

Even though the algorithm will not assign names to the resulted clusters, it can still produce and differentiate among them and use some of them to assign new examples into other clusters. This approach is driven by input data and can work well when there is adequate data available for use. An example of that is social information filtering algorithms, similar to those used by Amazon.com to recommend

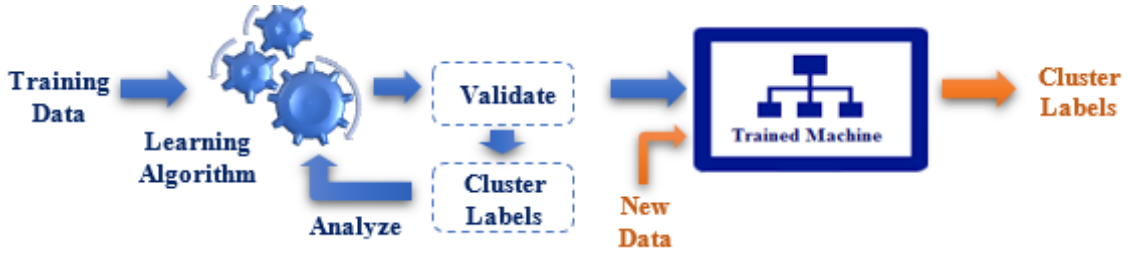

FIGURE 4. The different stages of the unsupervised learning.

books to users. These algorithms are based on finding similar groups of people, then adding new members to these groups [43], [44]. Algorithms included in unsupervised learning can be divided into three main categories, which are: clustering, dimensionality reduction and anomaly detection [44].

3) SEMI-SUPERVISED LEARNING

This method falls between the supervised and unsupervised learning methods where we have a large amount of input data, some of which are labeled and the rest are not. Many real life learning problems fall under this area of machine learning. The reason for that is that semi-supervised requires less human intervention since it utilizes very small amount of labeled data and a large amount of unlabeled data. Utilizing less labeled datasets is more appealing since such datasets are very hard to collect as well as expensive and may require access to domain experts. Unlabeled datasets on the other hand are cheaper and easier to get access to [45].

Both supervised and unsupervised learning techniques can be utilized to train the learning algorithm in semi-supervised learning. Unsupervised learning techniques can be used to unfold hidden structures and patterns in the input dataset. Whereas supervised learning techniques can be utilized to make guess predictions on the unlabeled data, feed the data back to the learning algorithm as training data and use gained knowledge to make predictions on new sets of data. Thus, we can say that unlabeled data is used to modify or reprioritize prediction or hypothesis obtained from labeled data [45]. Figure 5 illustrates the different stages of a semi-supervised machine learning method.

In order to make use of the unlabeled training data, all semi-supervised learning algorithms do at least one of the following assumptions [45]: smoothness assumption, cluster assumption and manifold assumption.

4) REINFORCEMENT

Reinforcement learning is learning by interacting with the problem environment. A reinforcement learning agent learns from its own actions rather than being specifically taught what to do. It selects current actions based on past experiences (exploitation) and new choices (exploration). Thus, it can be described as a trial and error learning process. The success of an action is determined through a signal received by the reinforcement learning agent in the form of a numerical reward value. The agent aims to learn to select actions that maximize the value of the numerical reward [46]. Actions may affect not only the current situation and current reward value, but also affect successive situations and reward values.

Learning agents usually have goals set and it can sense, to some extent, the state of the environment it is in and thus take actions that affect the state and bring it closer to the set goals. Reinforcement learning is different from supervised learning based on the way each method gains knowledge. Supervised learning method learns from examples provided by an external supervisor. Whereas reinforcement learning uses direct interactions with the problem environment to gain knowledge [46], [36].

5) DEEP LEARNING

Since 2006, this class of machine learning has emerged strongly and has been incorporated in hundreds of researches ever since. Areas in which deep learning have been incorporated ranged from information processing to artificial intelligence. Deep Learning can be described as a sub-field of machine learning that is based on algorithms that learn from multiple of levels in order to provide a model that represents complex relations among data. A hierarchy of features is present such that high level features are defined in terms of lower level features and this is why it is referred to as deep architecture. Most of the models incorporated under this class are based on the unsupervised learning representations [47].

Deep learning is basically the intersection point between neural networks, graphical modeling, optimization, artificial intelligence, pattern recognition as well as signal processing. The rationale behind the popularity of deep learning can be summarized in the following: it helped in highly increasing the processing abilities of computer chips, it allowed incorporation of a huge size of training data and it was the reason for the recent advances in machine learning in the area of information and signal processing [47].

a: OVERVIEW ON DEEP LEARNING

Until recently, most of the signal processing techniques were based on the utilization of shallow structured architectures. These architectures typically contained one or two layers at most of non-linear feature transformation. Examples of these shallow architectures include: Gaussian mixture models (GMMs), the support vector machines (SVMs) and

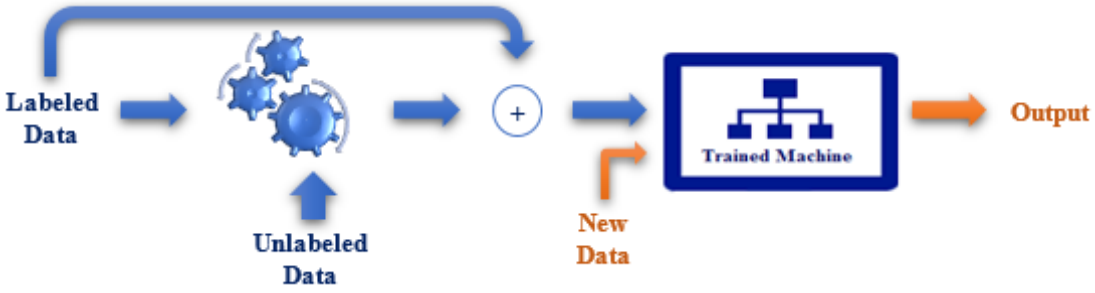

FIGURE 5. Different stages of the semi-supervised learning.

linear or nonlinear dynamical systems [47]. These architectures are best suited for simple or constrained problems as their limited abilities can cause problems in large scale complicated real world problems. Such real world problems may include human speech, language recognition and visual scenes which require a more deep and layered architecture to be able to extract such complex information [48].

The concept of deep learning first originated from artificial network research. A good example of models with a deep architecture is deep neural networks, which are often described as feed-forward neural networks. Back-propagation (BP) was one of the most popular algorithms used for learning the parameters of these networks. However, alone BP did not work well for learning networks that contain more than a small number of hidden layers [49]. The persistent occurrence of local optima in the non-convex objective function of the deep networks are the main source of difficulties in the learning. The difficulty in optimization with the deep models was empirically alleviated when an unsupervised learning algorithm was introduced [50], [51]. Deep belief networks (DBN), which is a class of deep generative models, were introduced. DBN consists of a stack of restricted Boltzmann machines (RBMs). At the core of the DBN is a greedy learning algorithm that optimizes DBN weights at time complexity linear to the size and depth of the networks [52].

Incorporating hidden layers with a huge number of neurons in a DNN has shown to highly improve the modeling abilities of the DNN and therefore create many closely optimal configurations [53]. Even in the case where parameter learning was trapped into a local optimum, the resulting DNN is still capable of performing quite well since the chance of having a poor local optimum gets lower and lower as the number of neurons used is high. However, using deep neural networks would require high computational power during the training process. Since huge computational capabilities were not easily available in the past, it was not until recent years that researchers have started seriously exploring deep neural networks.

b: CLASSES OF DEEP LEARNING

The term ''deep learning'' refers to a wide range of machine learning techniques as well as architectures which are based on the use of many layers of non-linear information processing that are considered hierarchical in nature. Depending on the intention behind using deep learning, whether it synthesis or recognition, generation or classification, a broad classification can be used that defines three different classes of deep learning, which are: deep networks intended for unsupervised (generative) learning, deep networks for supervised learning and hybrid deep networks. As for the first class, it aims to capture high order correlation of the visible data for the purpose of synthesis or pattern analysis given that no information is available about the target class labels. The second class aims to directly provide the discriminative power for the purpose of pattern classification. Labeled data are always present in the direct or indirect form for supervised learning. Finally, the third class aims to perform discriminations which are often assisted with the outcomes of generative or unsupervised deep networks. This is done usually by better optimization of the deep networks in class 2 [53]. There are many different deep learning algorithms, two of these popular algorithms are briefly discussed below.

c: CONVOLUTIONAL NEURAL NETWORKS (CNN)

These networks are considered a type of discriminative deep architecture in which every model contains a convolutional layer and a pooling layer and are stacked on top of each other [54]. Many weights are shared in the convolutional layer, the pooling layer on the other hand sub-samples the output coming from the convolutional layer and decreases the data rate of the below layer. The weight sharing together with properly chosen pooling schemes, results in invariance properties of the CNN. Some have argued that the limited invariance found in CNN is not satisfactory for complicated pattern recognition tasks. However, the CNNs have shown effectiveness when used in computer vision or image recognition tasks [55], [56]. Also, with some appropriate changes in the CNN for image analysis purposes such that it incorporates speech properties, the CNN can be utilized in speech recognition as well [57]. The drive for using the convolution operator in such applications, which is considered a specialized linear operator, is the fact that it uses three important concepts, which are: sparse interactions, parameter sharing, and equivariant representation.

d: RECURRENT NEURAL NETWORKS

Recurrent Neural Networks (RNNs) are considered as a class of deep networks for the use in unsupervised learning in the cases where the depth of the input data sequence can be as large as the length since RNNs allow parameter sharing through the different layers of the network [58], [59]. RNNs are developed by the use of the same set of weights in a recursive manner over a tree like structure, and the tree is traversed in topological order [58], [59]. The RNN is used mainly for the purpose of predicting the future data sequence through the use of previous data samples. The RNN is very prevailing when it comes to modeling sequence data such as speech or text. However, until recently, these networks were not widely used since they are considered hard to train such that it captures the long term dependencies [60]. In recent years, advances in Hessian free optimization [61] have helped in overcoming this obstacle through the use of approximated second order information or stochastic curvature estimates. In some of the recent published work [62], RNNs trained with Hessian free optimization are demonstrated to be well capable of generating sequential text characters.

IV. METHODOLOGY

The survey conducted in this paper is based on the Systematic Literature Review (SLR) presented by Kitchenham and Charters methodology [8]. Their methodology divides work into several phases where each phase includes several stages, which are: the planning phase, conducting phase, and finally reporting phase. We divided the first phase into six different stages. The first stage was identifying the research questions which were based on the objectives set for the review. The second stage was specifying the research strategy used to retrieve related research papers, at this stage we also specified the proper search terms as well as the proper paper selection criteria. The third stage was specifying a proper study selection measures which includes the inclusion/exclusion rules. The fourth stage was designing quality assessment rules which were used to filter the research papers. Stage five was designing the data extraction approach which was used to answer the research questions raised. The last stage was synthesizing the extracted data from the research papers. The following subsections demonstrate the review protocol that was followed in this paper.

A. RESEARCH QUESTIONS

The main goal of this review paper is to identify and examine articles that implement deep neural networks in the area of speech. Based on that, the following research questions were identified:

- RQ1: What are the different types of papers that were included in the study?
- RQ2: What are the different types of speech identified in the research papers?
- RQ3: What are the different types of database used to test and train the algorithm in each paper?
- RQ4: What are the different database languages identified in the research papers?
- RQ5: What type of environment was used to conduct the study?
- RQ6: How features were extracted from speech?
- RQ7: What type of evaluation technique was used in the research papers?
- RQ8: What types of deep neural network models have been used?

B. SEARCH STRATEGY

Below is a detailed explanation on the search strategy that was implemented in this review:

1) SEARCH TERMS

The search terms were identified based on the following [1]:

1) The research questions were used to identify the main search terms.
2) New terms were identified based on published papers and books.
3) Boolean operators were used (ANDs and ORs) in order to limit the search results.

The search terms that were used include the following:

- "deep neural network" AND "speech"
- "deep neural networks" AND "speech"
- DNN AND speech
- "deep neural network" OR "deep neural networks" OR DNN AND speech
- "deep learning" AND Speech

2) SURVEY RESOURCES

The following digital libraries were used to search for the needed research papers:

- Google Scholar
- IEEE Explorer
- Science Direct
- ResearchGate
- Springer

3) SEARCH PHASES

The search terms listed earlier were used to retrieve the research papers from the specified digital libraries. The inclusion/exclusion criteria used is explained in details in the coming section. Based on our used inclusion/exclusion criteria, 174 publications were used in this review.

C. STUDY SELECTION

Originally, we obtained 230 papers which were based on the search conducted using the listed search terms. Further filtration was performed by the authors to ensure only relevant papers were included in this review and the results found were discussed in scheduled regular meetings. The selection and filtration steps that were used are listed below:

1) Step 1: removing all duplicate research papers that were obtained from different digital libraries.

2) Step 2: applying inclusion/exclusion criteria to ensure only relevant papers are included in this review.
3) Step 3: removing review papers from the list of papers. It is important to note that these identified review papers were used to conduct a comparison with our review.
4) Step 4: applying the quality assessment in order to include papers with the highest quality which give the best answers to the raised research questions.

The used inclusion/exclusion criteria in this review paper is defined below: Inclusion criteria:

- Include papers that use deep neural networks in the area of speech.
- Include papers that use deep learning in the area of speech.

Exclusion criteria:

- Exclude papers that use deep neural networks in an area other than speech.
- Exclude papers that are related to speech but do not use deep neural networks.
- Exclude papers with no clear publication information.

D. QUALITY ASSESSMENT RULES

Applying quality assessment rules was the final step used to identify the final list of papers included in this review paper. QARs were applied to evaluate the quality of the research papers in accordance with the set research questions. Ten QARs were identified, each worth 1 mark out of 10. The score of each QAR was selected as follows: when fully answered score = 1, answer above average score = 0.75, average answer score = 0.5, below average answer score = 0.25, completely not answered score = 0. The summation of all ten QARs represents the overall score of each paper. A score of 6 or less means the paper was excluded from this review. The QARs that were used to evaluate the quality of the papers are the following:

1) QAR 1: Is the paper well organized?
2) QAR 2: Are the research objectives identified clearly in the paper?
3) QAR 3: Is there sufficient background information provided in the paper?
4) 4. QAR 4: Is the specific area of speech used clearly defined?
5) QAR 5: Does the paper include practical experimentations?
6) QAR 6: Is the conducted experiment suitable and acceptable?
7) QAR 7: Is the data set used clearly identified?
8) QAR 8: Are the results of the conducted experiments clearly identified and reported?
9) QAR 9: Are the methods used to analyze the results appropriate?
10) QAR 10: Overall, is the paper considered useful?

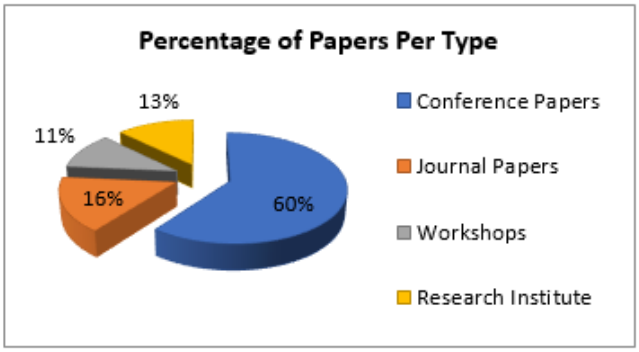

FIGURE 6. Percentage of papers in each type.

TABLE 1. Distribution of papers over the identified 13 conferences.

SLT	A128	1	1.0%
ASRU	A68	1	1.0%
ChinaSIP	A105, A116, A143	3	2.9%
ISCSLP	A125	1	1.0%
DAGA	A146	1	1.0%
ICASSP	A6, A15, A17, A20, A21, A23, A27, A28, A32, A36, A39, A40, A43, A44, A46, A53, A55, A57, A61, A63, A70, A71, A72, A73, A74, A75, A80, A83, A87, A91, A92, A95, A98, A99, A103, A106, A107, A109, A110, A111, A114, A115, A117, A120, A121, A122, A129, A131, A132, A141, A144, A148, A152, A158, A161, A170, A171,	57	54.3%
ICSP	A96, A126	2	1.9%
AAAAI	A154	1	1.0%
ACII	A59	1	1.0%
ICML	A1, A77	2	1.9%
Interspeech	A3, A8, A9, A13, A14, A16, A19, A26, A30, A35, A41, A42, A56, A66, A78, A79, A82, A86, A89, A100, A101, A102, A112, A119, A123, A124, A127, A137, A139, A142, A162, A163, A166	33	31.4%
ISCA	A76	1	1.0%
SIGDIAL	A52	1	1.0%

TABLE 2. Distribution of journal papers.

EURASIP	A136, A138, A145, A149, A153, A155	6	21.4%
IEEE Signal Processing Letters	A90, A93, A135	3	10.7%
IEEE Trans on Audio, Speech, and Language Processing	A7, A24, A29, A84, A85, A88, A94, A104, A108, A113, A130, A156, A167	13	46.4%
Scholarpedia	A4	1	3.6%
Ieee Journal Of Selected Topics In Signal Processing	A160	1	3.6%
Neural Networks	A164	1	3.6%
Speech Communications	A165, A172	2	7.1%
Ieee Transactions on Affective Computing	A174	1	3.6%

E. DATA EXTRACTION STRATEGY

In this stage, the finalized list of papers was used to extract needed information to answer the set of research questions.

TABLE 3. Different areas of speech the papers fall under.

Area	Conference	Journal	Workshops	Research	Total # of Papers	Percentage
speaker identification	A87	A113, A135	A81, A97	A150	6	3.4%
speech emotion recognition	A17, A124	A164, A174	A60	A49	6	3.4%
speech enhancement	A90, A130, A161, A166	A86, A100, A105, A119, A129, A139	A147	A157, A168, A169, A173	15	8.6%
speech recognition	A1, A3, A6, A8, A9, A13, A15, A19, A20, A21, A23, A26, A27, A28, A32, A35, A36, A39, A40, A41, A42, A43, A44, A46, A52, A53, A55, A57, A59, A61, A63, A66, A68, A70, A71, A72, A73, A74, A75, A76, A77, A78, A79, A80, A82, A83, A89, A92, A95, A96, A99, A101, A102, A103, A106, A107, A110, A111, A112, A114, A115, A116, A117, A120, A121, A122, A123, A125, A126, A127, A128, A131, A132, A137, A141, A142, A143, A144, A146, A148, A152, A154, A158, A162, A170, A171	A4, A7, A24, A29, A84, A85, A88, A93, A94, A104, A108, A136, A138, A145, A149, A153, A155, A156, A160, A165, A167, A172	A11, A22, A25, A31, A34, A51, A54, A58, A62, A64, A65, A69, A140, A51	A2, A5, A10, A12, A33, A37, A45, A47, A48, A50, A67, A118, A133, A134, A159	137	78.7%
speech transcription	A14, A30, A56, A109		A18		5	2.9%
Other	A16, A91, A98, A163		A38		5	2.9%

The information extracted from each paper included the following: paper ID, paper title, publication year, publication type, domain, RQ1, RQ2, RQ3, RQ4, RQ5, RQ6 and RQ7. Some difficulties occurred during the extraction process. For instance, in some papers different error percentages were presented to showcase the advantage of their technique in comparison to others without actually explaining how this error percentage was calculated. It is also important to note that not all papers answered all research questions.

F. SYNTHESIS OF EXTRACTED DATA

The extracted information for research question RQ1 – RQ5 and RQ7 were tabulated and presented as quantitative data that was used to develop a statistical comparison between the different findings for each research question. These developed statistics helped uncover certain research patterns as well as research directions that were carried over the past decade. As for RQ6, since extracted data was qualitative, a descriptive comparison was carried which focused on the main similarities and differences that were spotted between the included research papers.

V. RESULTS

A. RESEARCH QUESTION 1

The 174 papers that were included in the study fall into four main different types, which are: conference papers, journal papers, workshop papers and finally research institute publications. Figure 6 provides the distribution of papers between these main types.

The majority of papers used in the study, at a 60%, were identified as conference papers, as it can be clearly seen above. The rest 40% were distributed between the journal papers, workshop papers and research institute papers at a 16%, 11% and 13% respectively. In order to provide even further details about the papers that were used, a detailed statistical data was derived on the different conferences and journals that these papers were published in. Table 1 displays the distribution of papers in the 13 different identified conferences.

As it can be seen, the majority of conference papers, at about a 54%, were published in ICASSP ''IEEE International Conference on Acoustics, Speech and Signal Processing''. This was followed by 31% published in Interspeech. The remaining 15% was divided between the rest of the conferences at a 3% for ChinaSIP '' IEEE China Summit and International Conference on Signal and Information Processing'', 2% for ICML ''International Conference on Machine Learning'' and 2% for ICSP, then finally 1% for each of the remaining 8 conferences. Similarly, Table 2 presents the distribution of journal papers.

As it can be seen, the majority of journal papers, at a 46%, were published in the IEEE Transaction on Audio, Speech

and Language Processing. This was followed by 21% published in EURASIP "The European Association for Signal Processing", 11% in IEEE Signal Processing Letters, 7% in Speech Communications, and finally 4% in all of the following: Scholarpedia, IEEE Journal of Selected Topics in Signal Processing, Neural Networks and the IEEE Transactions of Affective Computing.

B. RESEARCH QUESTION 2

Among the 174 papers, different areas of speech were identified which include: speaker identification, speech emotion recognition, speech enhancement, speech recognition, speech transcription among others. The percentage of papers in each of these speech areas is shown in Table 3.

The majority of papers fall under the speech recognition area at a 79%, followed by about 9% in the speech enhancement area and 3% in the speaker identification, speech emotion recognition and the speech transcription area. Also, 3% of the papers were categorized as other, this category includes the areas of speech that had less than 1% of papers and include: speaker verification, language identification, speech pattern classification and spoken language understanding.

Since a huge percentage of papers fall under the area of speech recognition, further analysis was carried on these papers in order to identify even more details on the different speech recognition areas each paper is focused on. Different subareas were identified including: automatic speech recognition, large vocabulary speech recognition, low resource speech recognition, multilingual speech recognition, noise robust speech recognition, phone recognition, sequence classification, speaker adaptation and speech separation among others. Table 4 provides detailed information about these subcategories.

In speech recognition, the area of large vocabulary speech recognition had the most number of published papers at a 20%, which was followed by Noise robust speech recognition at about 14%. Further, 13% of the papers only mentioned speech recognition without any further details, 10% fall under automatic speech recognition, 6% fall under speaker adaptation, 4% under phone recognition and speech separation, 3% under sequence classification, multilingual speech and low resource speech recognition. Also, 20% of the speech recognition papers were classified as other since they include sub areas that have a less than 1% publication percentage.

C. RESEARCH QUESTION 3

To train and test algorithms, several databases were used in the research papers. Some were private while the majority of the databases, at 83%, were public and available on the web. More details are shown in Table 5. Some of the public databases that were used include: TIMIT dataset, ATIS dataset, Switchboard Hub5 task, Aurora 4, Babel corpus, AMI corpus1 among others.

D. RESEARCH QUESTION 4

Table 6 displays the different languages that were identified in the research papers that used to train and test the algorithms. As it can be clearly seen, the majority of the papers at 85% used an English language based dataset. Only 9% of the papers used multiple languages based datasets,

TABLE 4. The specific speech recognition areas identified in the papers.

Specific Area in Speech Recognition	Papers	Count of Papers	Percentage
Automatic Speech Recognition (ASR)	A3, A4, A5, A19, A25, A65, A75, A138, A140, A151, A154, A155, A165, A167	14	10.2%
Large Vocabulary Speech Recognition	A11, A22, A26, A29, A32, A36, A39, A41, A42, A43, A48, A50, A58, A59, A63, A64, A67, A68, A69, A77, A84, A99, A114, A131, A133, A134, A159	27	19.7%
Low resource speech recognition	A51, A54, A55, A110	4	2.9%
Multilingual speech recognition	A47, A71, A72, A160	4	2.9%
Noise robust speech recognition	A44, A45, A46, A61, A62, A73, A74, A88, A96, A111, A116, A117, A126, A136, A141, A142, A143, A144, A148, A149	20	14.6%
Phone recognition	A10, A15, A24, A40, A145, A162	6	4.4%
Sequence classification	A6, A9, A79, A120	4	2.9%
Speaker Adaptation	A57, A80, A89, A122, A123, A127, A137, A156	8	5.8%
Speech recognition	A1, A13, A20, A23, A31, A33, A35, A70, A83, A92, A94, A101, A103, A104, A106, A112, A146, A171	18	13.1%
speech separation	A95, A85, A108, A153, A170	5	3.6%
Other	A2, A7, A8, A12, A21, A27, A28, A34, A37, A52, A53, A66, A76, A78, A82, A93, A102, A107, A115, A118, A121, A125, A128, A132, A152, A158, A172	27	19.7%

TABLE 5. Type of database used in the research papers.

Type	Count	Percentage
Public	147	84%
Private	27	16%

TABLE 6. Different languages used in the research papers.

Language	Count	Percentage
English	148	85%
Chinese	2	1%
Italian	2	1%
French	2	1%
Japanese	1	1%
low-resource South African languages	1	1%
Multiple languages	15	9%
None Mentioned	3	2%

TABLE 7. Type of environment used in the research papers.

Type	Count	Percentage
Noisy Environment	47	27%
Emotional Speech	3	2%
Neutral	124	71%

TABLE 8. Identified evaluation techniques in the research papers.

Evaluation Technique	# of Papers	Percentage
Label Error Rate (LER)	2	1%
Phone Error Rate (PER)	17	11%
Word Error Rate (WER)	87	56%
Error Rate	17	11%
Accuracy	17	11%
dB SIR Gain	3	2%
Phoneme classification performance	2	1%
Other	29	19%

TABLE 9. Standalone deep neural networks.

Model Name	Number of papers	Percentage (%)
Deep Neural Network (DNN) ***	67	38
Deep Belief Network (DBN)	15	8.6
Convolutional Neural Network (CNN)	26	15
Recurrent Neural Network (RNN)	15	8.6
Deep Maxout Network (DMN)	3	1.72
Deep Convex Network (DCN)	3	1.72
Deep Stacking Network (DSN)	1	0.57
Deep Tensor Network (DTN)	1	0.57
Autoencoder	1	0.57

*** authors only mentioned DNN as a type without further clarifications

TABLE 10. Hybrid deep neural networks.

Model Name	Number of papers	Percentage (%)
DNN – HMM (Hidden Markov Model)	21	12
DBN - HMM	15	8.6
DMN - HMM	2	1.15
DNN – GMM (Gaussian Mixture Model) - HMM	1	0.57
CNN-RNN	1	0.57
CNN-HMM	1	0.57
RNN-HMM	1	0.57

2% of the papers did not mention the used language and the 5% remaining papers used Chinese, Italian, French, Japanese and South African languages at 1% each.

E. RESEARCH QUESTION 5

The environment used to train and test used algorithms varied between noisy, neutral and emotional. 71% of the papers either mentioned using a neutral environment or not mention anything at all and thus assumed neutral. As for noisy environment, 27% of the papers mentioned building a noisy robust system. Only 2% of the papers mentioned using emotional speech to test and train the algorithms. Table 7 provides more details on this.

F. RESEARCH QUESTION 6

TFeatures were extracted from speech using different techniques. It was found that the most popular is the mel-frequency cepstrum coefficients (MFCCs), as 69.5% of the papers (121 papers) used MFCCs to extract features from speech. On the other hand, almost 10% of the papers used the linear discriminate analysis (LDA) transform and almost 5% used the HLDA transform. In addition, it was found that almost 2% of the papers used the short time Fourier transform (STFT). Finally, the rest of the papers used other techniques that include: MLLT, perceptual linear predictive (PLP), log power spectral (LPS), Bark-frequency cepstrum coefficients (BFCC), batch normalization, maximum likelihood linear transform (MLLT) and residual connections.

G. RESEARCH QUESTION 7

Several evaluation techniques were used in the research papers to evaluate the overall performance of the developed system. Table 8 illustrates the different identified techniques, as well as the percentage of papers that used each technique. As it can be seen, 56% of the papers used Word Error Rate (WER) to evaluate the performance of their system, whereas 11% used for each Phone Error Rate (PER), error rate calculations and accuracy calculations. On the other hand 2% of the papers used dB SIR Gain calculations, 1% used Label Error Rate (LER) and Phoneme Classification Performance. On the other hand, 15% of the papers were classified as "other" since used techniques by each of the papers scored less than 1%. Some of the techniques that fall under this category include: Root Mean Square Error (RMSE), Sentence Accuracy, Query Error Rate (QER), unweighted classification accuracy, Gain in dB among others.

TABLE 11. List of extracted papers.

[A1]	Connectionist Temporal Classification: Labelling Unsegmented Sequence Data With Recurrent Neural Networks
[A2]	Deep Neural Network Adaptation For Children's And Adults' Speech Recognition
[A3]	Application Of Pretrained Deep Neural Networks To Large Vocabulary Speech Recognition
[A4]	Deep Belief Networks For Phone Recognition
[A5]	From Speech To Letters - Using A Novel Neural Network Architecture For Grapheme Based ASR
[A6]	Lattice-Based Optimization Of Sequence Classification Criteria For Neural-Network Acoustic Modeling
[A7]	Speech Recognition Using Augmented Conditional Random Fields
[A8]	Binary Coding Of Speech Spectrograms Using A Deep Auto-Encoder
[A9]	Investigation Of Full-Sequence Training Of Deep Belief Networks For Speech Recognition
[A10]	Phone Recognition With The Mean-Covariance Restricted Boltzmann Machine
[A11]	Roles Of Pre-Training And Fine-Tuning In Context-Dependent DBN-HMMs for Real-World Speech Recognition
[A12]	Deep Neural Network For Acoustic-Articulatory Speech Inversion
[A13]	Accelerated Parallelizable Neural Network Learning Algorithm For Speech Recognition
[A14]	Conversational Speech Transcription Using Context-Dependent Deep Neural Networks
[A15]	Deep Belief Networks Using Discriminative Features For Phone Recognition
[A16]	Deep Convex Net: A Scalable Architecture For Speech Pattern Classification
[A17]	Deep Neural Networks For Acoustic Emotion Recognition: Raising The Benchmarks
[A18]	Feature Engineering In Context-Dependent Deep Neural Networks For Conversational Speech Transcription
[A19]	Improved Bottleneck Features Using Pretrained Deep Neural Networks
[A20]	Large Vocabulary Continuous Speech Recognition With Context-Dependent DBN-HMMs
[A21]	Learning A Better Representation Of Speech Soundwaves Using Restricted Boltzmann Machines
[A22]	Making Deep Belief Networks Effective For Large Vocabulary Continuous Speech Recognition
[A23]	Speech Recognition with Segmental Conditional Random Fields: A Summary Of The JHU CLSP 2010 Summer Workshop
[A24]	Acoustic Modeling Using Deep Belief Networks
[A25]	Adaptation Of Context-Dependent Deep Neural Networks For Automatic Speech Recognition
[A26]	Application Of Pretrained Deep Neural Networks To Large Vocabulary Speech Recognition
[A27]	Applying Convolutional Neural Networks Concepts To Hybrid Nn-Hmm Model For Speech Recognition
[A28]	Boosting Attribute And Phone Estimation Accuracies With Deep Neural Networks For Detection-Based Speech Recognition
[A29]	Context-Dependent Pre-Trained Deep Neural Networks For Large-Vocabulary Speech Recognition
[A30]	Conversational Speech Transcription Using Context-Dependent Deep Neural Networks
[A31]	Deep Neural Networks For Acoustic Modeling In Speech Recognition
[A32]	Exploiting Sparseness In Deep Neural Networks For Large Vocabulary Speech Recognition
[A33]	Factorized Deep Neural Networks For Adaptive Speech Recognition
[A34]	Improving Wideband Speech Recognition Using Mixed-Bandwidth Training Data In Cd-DNN-HMM
[A35]	Integrating Deep Neural Networks Into Structured Classification Approach Based On Weighted Finite-State Transducers
[A36]	Investigation On Dimensionality Reduction Of Concatenated Features With Deep Neural Network For LVCSR Systems
[A37]	Scalable Stacking And Learning For Building Deep Architectures
[A38]	Use Of Kernel Deep Convex Networks And End-To-End Learning For Spoken Language Understanding
[A39]	A Cluster-Based Multiple Deep Neural Networks Method For Large Vocabulary Continuous Speech Recognition
[A40]	A Deep Convolutional Neural Network Using Heterogeneous Pooling For Trading Acoustic Invariance With Phonetic Confusion
[A41]	A Scalable Approach To Using DNN-Derived Features In GMM-HMM Based Acoustic Modeling For LVCSR
[A42]	Accurate And Compact Large Vocabulary Speech Recognition On Mobile Devices
[A43]	An Empirical Study Of Learning Rates In Deep Neural Networks For Speech Recognition
[A44]	An Investigation Of Deep Neural Networks For Noise Robust Speech Recognition
[A45]	Audio-Visual Deep Learning For Noise Robust Speech Recognition
[A46]	Automatic Localization Of A Language-Independent Sub-Network On Deep Neural Networks Trained By Multi-Lingual Speech
[A47]	Cross-Language Knowledge Transfer Using Multilingual Deep Neural Network With Shared Hidden Layers
[A48]	Deep Convolutional Neural Networks For LVCSR
[A49]	Deep Learning For Robust Feature Generation In Audiovisual Emotion Recognition In Audiovisual Emotion Recognition
[A50]	Deep Maxout Networks For Low-Resource Speech Recognition
[A51]	Deep Maxout Neural Networks For Speech Recognition
[A52]	Deep Neural Network Approach For The Dialog State Tracking Challenge
[A53]	Deep Neural Network Features And Semi-Supervised Training For Low Resource Speech Recognition
[A54]	Elastic Spectral Distortion For Lowre source Speech Recognition With Deep Neural Networks
[A55]	Error Back Propagation For Sequence Training Of Context-Dependent Deep Networks For Conversational Speech Transcription
[A56]	Exploring Convolutional Neural Network Structures And Optimization Techniques For Speech Recognition
[A57]	Fast Speaker Adaptation Of Hybrid NN/Hmm Model For Speech Recognition Based On Discriminative Learning Of Speaker Code
[A58]	Hybrid Acoustic Models For Distant And Multichannel Large Vocabulary Speech Recognition
[A59]	Hybrid Deep Neural Network - Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition
[A60]	Hybrid Speech Recognition With Deep Bidirectional LSTM
[A61]	Ideal Ratio Mask Estimation Using Deep Neural Networks For Robust Speech Recognition
[A62]	Improvements To Deep Convolutional Neural Networks For LVCSR

TABLE 11. *(Continued.)*

[A63]	Improving Deep Neural Networks For LVCSR Using Rectified Linear Units And Dropout
[A64]	Improving Robustness Of Deep Neural Networks Via Spectral Masking For Automatic Speech Recognition
[A65]	Investigation Of Multilingual Deep Neural Networks For Spoken Term Detection
[A66]	Investigation Of Recurrent-Neural-Network Architectures And Learning Methods For poken Language Understanding
[A67]	Kl-Divergence Regularized Deep Neural Network Adaptation For Improved Large Vocabulary Speech Recognition
[A68]	Large Scale Deep Neural Network Acoustic Modeling With Semi-Supervised Training Data For Youtube Video Transcription
[A69]	Learning Filter Banks Within A Deep Neural Network Framework
[A70]	Low-Rank Matrix Factorization For Deep Neural Network Training With High-Dimensional Output Targets
[A71]	Multilingual Acoustic Models Using Distributed Deep Neural Networks
[A72]	Multilingual Training Of Deep Neural Networks
[A73]	Noise Adaptive Front-End Normalization Based On Vector Taylor Series For Deep Neural Networks In Robust Speech Recognition
[A74]	On Rectified Linear Units For Speech Processing
[A75]	Predicting Speech Recognition Confidence Using Deep Learning With Word Identity And Score Features
[A76]	Rapid And Effective Speaker Adaptation Of Convolutional Neural Network Based Models For Speech Recognition
[A77]	Rectifier Nonlinearities Improve Neural Network Acoustic Models
[A78]	Restructuring Of Deep Neural Network Acoustic Models With Singular Value Decomposition
[A79]	Sequence-Discriminative Training Of Deep Neural Networks
[A80]	Speaker Adaptation Of Context Dependent Deep Neural Networks
[A81]	Speaker Adaptation Of Neural Network Acoustic Models Using I-Vectors
[A82]	Speech Activity Detection On Youtube Using Deep Neural Networks
[A83]	Speech Recognition With Deep Recurrent Neural Networks
[A84]	The Deep Tensor Neural Network With Applications To Large Vocabulary Speech Recognition
[A85]	Towards Scaling Up Classification-Based Speech Separation
[A86]	A Comparative Analytic Study On The Gaussian Mixture And Context Dependent Deep Neural Network Hidden Markov Models
[A87]	A Novel Scheme For Speaker Recognition Using A Phonetically-Aware Deep Neural Network
[A88]	A Spectral Masking Approach To Noise-Robust Speech Recognition Using Deep Neural Networks
[A89]	Adaptation Of Deep Neural Network Acoustic Models Using Factorized I-Vectors
[A90]	An Experimental Study On Speech Enhancement Based On Deep Neural Networks
[A91]	Automatic Language Identification Using Deep Neural Networks
[A92]	Context Dependent State Tying For Speech Recognition Using Deep Neural Network Acoustic Models
[A93]	Convolutional Neural Networks For Distant Speech Recognition
[A94]	Convolutional Neural Networks For Speech Recognition
[A95]	Deep Learning For Monaural Speech Separation
[A96]	Deep Neural Network Based Speech Separation For Robust Speech Recognition
[A97]	Deep Neural Networks For Extracting Baum-Welch Statistics For Speaker Recognition
[A98]	Deep Neural Networks For Small Footprint Text-Dependent Speaker Verification
[A99]	Direct Adaptation Of Hybrid DNN/Hmm Model For Fast Speaker Adaptation In LVCSR Based On Speaker Code
[A100]	Dynamic Noise Aware Training For Speech Enhancement Based On Deep Neural Networks
[A101]	Ensemble Deep Learning For Speech Recognition
[A102]	Experiments On Deep Learning For Speech Denoising
[A103]	Factorized Adaptation For Deep Neural Network
[A104]	Fast Adaptation Of Deep Neural Network Based On Discriminant Codes For Speech Recognition
[A105]	Global Variance Equalization For Improving Deep Neural Network Based Speech Enhancement
[A106]	Improving Deep Neural Network Acoustic Models Using Generalized Maxout Networks
[A107]	Improving DNN Speaker Independence With I-Vector Inputs
[A108]	Improving Robustness Of Deep Neural Network Acoustic Models Via Speech Separation And Joint Adaptive Training
[A109]	I-Vector-Based Speaker Adaptation Of Deep Neural Networks For French Broadcast Audio Transcription
[A110]	Joint Acoustic Modeling Of Triphones And Trigraphemes By Multi-Task Learning Deep Neural Networks For Low-Resource Speech Recognition
[A111]	Joint Noise Adaptive Training For Robust Automatic Speech Recognition
[A112]	Learning Small-Size DNN With Output-Distribution-Based Criteria
[A113]	Learning Spectral Mapping For Speech Dereverberation
[A114]	Mean-Normalized Stochastic Gradient For Large-Scale Deep Learning
[A115]	Multilingual Deep Neural Network Based Acoustic Modeling For Rapid Language Adaptation
[A116]	Noisy Training For Deep Neural Networks
[A117]	Recurrent Deep Neural Networks For Robust Speech Recognition
[A118]	Relation Classification Via Convolutional Deep Neural Network
[A119]	Robust Speech Recognition With Speech Enhanced Deep Neural Networks
[A120]	Sequence Classification Using The High-Level Features Extracted From Deep Neural Networks
[A121]	Single-Channel Mixed Speech Recognition Using Deep Neural Networks
[A122]	Singular Value Decomposition Based Low-Footprint Speaker Adaptation And Personalization For Deep Neural Network
[A123]	Speaker Adaptation Of DNN-Based ASR With I-Vectors: Does It Actually Adapt Models To Speakers?
[A124]	Speech Emotion Recognition Using Deep Neural Network And Extreme Learning Machine
[A125]	Speech Separation Based On Improved Deep Neural Networks With Dual Outputs Of Speech Features For Both Target And Interfering Speakers

TABLE 11. *(Continued.)*

[A126]	Speech Separation Of A Target Speaker Based On Deep Neural Networks
[A127]	Towards Speaker Adaptive Training Of Deep Neural Network Acoustic Models
[A128]	Vocal Tract Length Normalization Approaches To DNN-Based Children' S And Adults' Speech Recognition
[A129]	A Deep Neural Network Approach To Speech Bandwidth Expansion
[A130]	A Regression Approach To Speech Enhancement Based On Deep Neural Networks
[A131]	An Investigation Into Speaker Informed DNN Front-End For LVCSR
[A132]	An Investigation Of Augmenting Speaker Representations To Improve Speaker Normalization For DNN-Based Speech Recognition
[A133]	Building DNN Acoustic Models For Large Vocabulary Speech Recognition
[A134]	Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks
[A135]	Deep Neural Network Approaches To Speaker And Language Recognition
[A136]	Exploiting Spectro-Temporal Locality In Deep Learning Based Acoustic Event Detection
[A137]	FMLLR Based Feature-Space Speaker Adaptation Of DNN Acoustic Models
[A138]	Exploiting Foreign Resources For DNN-Based ASR
[A139]	Integration Of DNN Based Speech Enhancement And ASR
[A140]	Investigating Sparse Deep Neural Networks For Speech Recognition
[A141]	Joint Training Of Front-End And Back-End Deep Neural Networks For Robust Speech Recognition
[A142]	Multi-Resolution Stacking For Speech Separation Based On Boosted DNN
[A143]	Noisy Training For Deep Neural Networks In Speech Recognition
[A144]	On The Importance Of Modeling And Robustness For Deep Neural Network Feature
[A145]	Phone Recognition With Hierarchical Convolutional Deep Maxout Networks
[A146]	Real-Time Dereverberation For Deep Neural Network Speech Recognition
[A147]	Robust ASR Using Neural Network Based Speech Enhancement And Feature Simulation
[A148]	Spatial Diffuseness Features For DNN-Based Speech Recognition In Noisy And Reverberant Environments
[A149]	Speech Recognition In Reverberant And Noisy Environments Employing Multiple Feature Extractors And I-Vector Speaker Adaptation
[A150]	Time Delay Deep Neural Network-Based Universal Background Models For Speaker Recognition
[A151]	Towards Structured Deep Neural Network For Automatic Speech Recognition
[A152]	Exploiting Low-Dimensional Structures To Enhance DNN Based Acoustic Modeling In Speech Recognition
[A153]	Localization Based Stereo Speech Source Separation Using Probabilistic Time-Frequency Masking And Deep Neural Networks
[A154]	Toward A Better Understanding Of Deep Neural Network Based Acoustic Modelling: An Empirical Investigation
[A155]	Wise Teachers Train Better DNN Acoustic Models
[A156]	Small-Footprint Highway Deep Neural Networks For Speech Recognition
[A157]	A Hybrid DSP/Deep Learning Approach To Real-Time Full-Band Speech Enhancement
[A158]	A Network Of Deep Neural Networks For Distant Speech Recognition
[A159]	Accelerating Deep Neural Network Learning For Speech Recognition On A Cluster Of GPUS
[A160]	An End-To-End Deep Learning Approach To Simultaneous Speech Dereverberation And Acoustic Modeling For Robust Speech Recognition
[A161]	Collaborative Deep Learning For Speech Enhancement: A Run-Time Model Selection Method Using Auto encoders
[A162]	Deep Learning-Based Telephony Speech Recognition In The wild
[A163]	Deep Neural Network Embeddings For Text-Independent Speaker Verification
[A164]	Evaluating Deep Learning Architectures For Speech Emotion Recognition
[A165]	Hybrid Convolutional Neural Networks For Articulatory And Acoustic Information Based Speech Recognition
[A166]	Improving Mask Learning Based Speech Enhancement System With Restoration Layers And Residual Connection
[A167]	Multichannel Signal Processing With Deep Neural Networks For Automatic Speech Recognition
[A168]	Multi-Objective Learning And Mask-Based Post-Processing For Deep Neural Network Based Speech Enhancement
[A169]	Multiple-Target Deep Learning For LSTM-RNN Based Speech Enhancement
[A170]	Permutation Invariant Training Of Deep Models For Speaker-Independent Multi-Talker Speech Separation
[A171]	Very Deep Convolutional Networks For End-To-End Speech Recognition
[A172]	Automatic Lexical Stress And Pitch Accent Detection For L2 English Speech Using Multi-Distribution Deep Neural Networks
[A173]	Deep Neural Network Based Monaural Speech Enhancement With Low-Rank Analysis And Speech Present Probability
[A174]	Feature Selection Based Transfer Subspace Learning For Speech Emotion Recognition

H. RESEARCH QUESTION 8

Regarding the types of deep neural network used in speech recognition, out of 174 papers, 132 papers used DNN models as standalone models and 42 papers used hybrid models (two or more models). Tables 9 and 10 display the standalone and hybrid models.

VI. CONCLUSIONS

This paper provided a thorough statistical analysis on the use of deep learning in speech related applications by extracting specific information from 174 papers published between the years 2006 and 2018. The majority of the papers identified (40%) were conference papers and more than 50% of these papers were published in ICASSP. As for the journal papers, it was found that 46% of them were published in the IEEE Transactions on Audio, Speech, and Language Processing. The majority of the papers focused on speech recognition as the application in use. As for the utilized data bases in the included study, they were mostly public and in English and the environment was mostly natural non noisy. As for the

evaluation technique used, it was found that the majority of the papers used the WER (word error rate) to determine the efficiency of their systems. We hope that the results provided in this study would help future researchers identify new and interesting research topics that has not been examined yet, as well as highlight some of the gaps in the existing studies.

It is surprising to see that most of the researchers still use MFCCs as feature extraction for speech signals in deep learning models. MFCCs were heavily used in classical classifiers such as HMM and GMM. It is worth trying when using deep learning models other feature extraction methods such as Linear Predictive Coding (LPC).

As seen in Tables 9 and 10, 75% of DNN models were standalone models where only 25% of the models used hybrid models. Authors are encouraged to use hybrid models as research showed that using HMM or GMM inform of a DNN model gives better results [7].

Another observation is that there is little work on speech recognition using Recurrent Neural Networks (RNN). Authors are highly recommended to conduct research using deep RNN in the future since RNN models, especially Long Short Time Memory (LSTM), are very powerful in speech recognition [53].

Appendix

See Table 11.

REFERENCES

[1] A. H. Meftah, Y. A. Alotaibi, and S.-A. Selouani, "Evaluation of an Arabic speech corpus of emotions: A perceptual and statistical analysis," *IEEE Access*, vol. 6, pp. 72845–72861, 2018.

[2] Y. Xie, L. Le, Y. Zhou, and V. V. Raghavan, "Deep learning for natural language processing," in *Handbook of Statistics*. Amsterdam, The Netherlands: Elsevier, 2018.

[3] J. Padmanabhan and M. J. J. Premkumar, "Machine learning in automatic speech recognition: A survey," *IETE Tech. Rev.*, vol. 32, no. 4, pp. 240–251, 2015.

[4] H. Singh and A. K. Bathla, "A survey on speech recognition," *Int. J. Adv. Res. Comput. Eng. Technol.*, no. 2, no. 6, pp. 2186–2189, 2013.

[5] M. A. Anusuya and S. K. Katti, "Speech recognition by machine: A review," *Int. J. Comput. Sci. Inf. Secur.*, vol. 6, no. 3, pp. 181–205, 2009.

[6] Y. Zhang, "Speech recognition using deep learning algorithms," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2013, pp. 1–5. [Online]. Available: https://scholar.google.com/scholar?as_q=Speech+Recognition+Using+Deep+Learning+Algorithms&as_occt=title&hl=en&as_sdt=0%2C31

[7] I. Shahin, A. B. Nassif, and S. Hamsa, "Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments," *Neural Comput. Appl.*, to be published.

[8] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature reviews in software engineering version 2.3," *Engineering*, vol. 45, no. 4, p. 1051, 2007.

[9] N. Morgan, "Deep and wide: Multiple layers in automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 7–13, Jan. 2012.

[10] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[11] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8599–8603.

[12] L. Deng *et al.*, "Recent advances in deep learning for speech research at Microsoft," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8604–8608.

[13] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: From fundamentals to practice," *Proc. IEEE*, vol. 101, no. 5, pp. 1136–1159, May 2013.

[14] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[15] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 4, May 2002, pp. 4072–4075.

[16] S. Furui, "Speaker-dependent-feature extraction, recognition and processing techniques," *Speech Commun.*, vol. 10, nos. 5–6, pp. 505–520, Dec. 1991.

[17] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 201–216, Mar. 2001.

[18] C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni, and A. Ghio, "Application of Automatic Speaker Recognition techniques to pathological voice assessment (dysphonia)," in *Proc. Eur. Conf. Speech Commun. Technol. (Eurospeech)*, 2005, pp. 149–152.

[19] R. W. Picard, "Affective Computing," MIT Press, Cambridge, MA, USA, Tech. Rep. 321, 1995, pp. 1–16.

[20] I. Shahin, "Employing emotion cues to verify speakers in emotional talking environments," *J. Intell. Syst.*, vol. 25, no. 1, pp. 3–17, 2016.

[21] I. M. A. Shahin, "Employing both gender and emotion cues to enhance speaker identification performance in emotional talking environments," *Int. J. Speech Technol.*, vol. 16, no. 3, pp. 341–351, Sep. 2013.

[22] I. Shahin, "Speaker identification in emotional talking environments based on CSPHMM2s," *Eng. Appl. Artif. Intell.*, vol. 26, no. 7, pp. 1652–1659, Aug. 2013.

[23] I. Shahin, "Identifying speakers using their emotion cues," *Int. J. Speech Technol.*, vol. 14, no. 2, pp. 89–98, 2011.

[24] V. A. Petrushin, "Emotion recognition in speech signal: Experimental study, development, and application," in *Proc. 6th Int. Conf. Spoken Lang. Process. (ICSLP)*, 2000, p. 5.

[25] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human–computer interaction," *Neural Netw.*, vol. 18, no. 4, pp. 389–405, 2005.

[26] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Commun.*, vol. 40, nos. 1–2, pp. 33–60, 2003.

[27] E. Nöth, S. Harbeck, and H. Niemann, "Multilingual speech recognition," in *Computational Models of Speech Pattern Processing* (NATO ASI Series), vol. 169. Springer, 1999, pp. 362–374. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-60087-6_31

[28] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, nos. 1–2, pp. 31–51, 2001.

[29] C. Chelba, T. J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 39–49, May 2008.

[30] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," Ph.D. thesis, Graduate School Arts Sci., Columbia Univ., New York City, NY, USA, 2011, pp. 1–171.

[31] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Nöth, "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar./Apr. 2008, pp. 1605–1608.

[32] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proc. Lang. Resour. Eval. Conf.*, Jan. 2006, pp. 1123–1126.

[33] I. M. A. Shahin, "Gender-dependent emotion recognition based on HMMs and SPHMMs," *Int. J. Speech Technol.*, vol. 16, no. 2, pp. 133–141, 2013.

[34] R. Schapire, *Theoretical Machine Learning* (Lecture). Princeton, NJ, USA: Princeton Univ., 2008, pp. 1–6.

[35] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[36] A. Mukherjee, A. Pal, and P. Misra, "Data analytics in ubiquitous sensor-based health information systems," in *Proc. 6th Int. Conf. Next Generation Mobile Appl. Services Technol. (NGMAST)*, 2012, pp. 193–198.

[37] M. I. Schlesinger and V. Hlavác, "Supervised and unsupervised learning," in *Ten Lectures on Statistical and Structural Pattern Recognition*. Springer, 2002. [Online]. Available: https://www.springer.com/gp/book/9781402006425

[38] J. T. Senders *et al.*, "An introduction and overview of machine learning in neurosurgical care," *Acta Neurochirurgica*, vol. 160, no. 1, pp. 29–38, 2018.
[39] E. Alpaydin, *Introduction to Machine Learning*, 3rd ed. Cambridge, MA, USA: MIT Press, 2015.
[40] *Supervised Learning*. Accessed: Dec. 15, 2017. [Online]. Available: https://en.wikipedia.org/w/index.php?oldid=457924979
[41] P. R. Krugman, M. Obstfeld, and M. J. Melitz, *International Economics: Theory and Policy*. New York, NY, USA: Prentice-Hall, 2012.
[42] L. C. Resende, L. A. F. Manso, W. D. Dutra, and A. M. L. da Silva, "Support vector machine application in composite reliability assessment," in *Proc. 18th Int. Conf. Intell. Syst. Appl. Power Syst. (ISAP)*, Porto, Portugal, 2015. [Online]. Available: https://ieeexplore.ieee.org/document/7325580
[43] R. Rojas, "Unsupervised learning and clustering algorithms," in *Neural Networks*. Springer, 1996, pp. 99–121. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-61068-4_5
[44] M. Caza-Szoka, D. Massicotte, and F. Nougarou, "Naive Bayesian learning for small training samples: Application on chronic low back pain diagnostic with sEMG sensors," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2015, pp. 470–475.
[45] X. Zhu, "Semi-Supervised Learning Literature Survey Contents," Dept. Comput. Sci., Univ. Wisconsin–Madison, Madison, WI, USA, Tech. Rep. 1530, 2008, p. 10.
[46] P. R. Montague, "Reinforcement Learning: An Introduction, by Sutton, R.S. and Barto, A.G.," *Trends Cogn. Sci.*, vol. 3, no. 9, p. 360, 1999.
[47] Y. Cho and L. K. Saul, "Kernel methods for deep learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 22, 2009, pp. 342–350.
[48] P. Dayan, "Unsupervised learning," in *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA, USA: MIT Press, 2009, pp. 1–7.
[49] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[50] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.
[51] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[52] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[53] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, May 2015.
[54] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4277–4280.
[55] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Proc. NIPS*, 2012, pp. 1–9.
[56] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, *Deep, Big, Simple Neural Nets for Handwritten Digit Recognition*, vol. 22, no. 12. Cambridge, MA, USA: MIT Press, Dec. 2010, pp. 3207–3220. [Online]. Available: https://www.mitpressjournals.org/doi/abs/10.1162/NECO_a_00052
[57] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2015.
[58] L. Deng, "Design and learning of output representations for speech recognition," in *Proc. NIPS Workshop Learn. Output Representations*, Dec. 2013.
[59] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Trans. Signal Inf. Process.*, vol. 3, no. e2, pp. 1–29, 2014.
[60] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, pp. 157–166, 1994.
[61] J. Martens, "Deep learning via Hessian-free optimization," in *Proc. 27th Int. Conf. Mach. Learn.*, vol. 951, 2010, pp. 735–742.
[62] J. Martens and I. Sutskever, "Learning recurrent neural networks with Hessian-free optimization," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 1033–1040.
[63] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.
[64] R. Serizel and D. Giuliani, "Deep neural network adaptation for children's and adults' speech recognition," in *Proc. 1st Italian Comput. Linguistics Conf.*, 2014, pp. 1–6.
[65] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," in *Proc. Interspeech*, 2012, pp. 2–5.
[66] A.-R. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," *Scholarpedia*, vol. 4, no. 5, pp. 1–9, 2009.
[67] F. Eyben, M. Wöllmer, B. Schuller, and A. Graves, "From speech to letters—Using a novel neural network architecture for grapheme based ASR," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Nov./Dec. 2010, pp. 376–380.
[68] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proc. ICASSP*, 2009, pp. 3761–3764.
[69] E. Fosler-Lussier, Y. He, P. Jyothi and R. Prabhavalkar, "Conditional random fields in speech, audio, and language processing," *Proc. IEEE*, vol. 101, no. 5, pp. 1054–1075, 2013.
[70] L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," in *Proc. Interspeech*, Sep. 2010, pp. 1692–1695.
[71] A.-R. Mohamed, D. Yu, and L. Deng, "Investigation of full-sequence training of deep belief networks for speech recognition," in *Proc. Interspeech*, Sep. 2010, pp. 2846–2849.
[72] G. E. Dahl, M. Ranzato, A.-R. Mohamed, and G. Hinton, "Phone recognition with the mean-covariance restricted Boltzmann machine," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 469–477.
[73] D. Yu, L. Deng, and G. E. Dahl, "Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," in *Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn.*, 2010, pp. 1–8.
[74] B. Uria, S. Renals, and K. Richmond, "A deep neural network for acoustic-articulatory speech inversion," in *Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn.*, 2011, pp. 1–9.
[75] D. Yu and L. Deng, "Accelerated parallelizable neural network learning algorithm for speech recognition," in *Proc. Interspeech*, Aug. 2011, pp. 2281–2284.
[76] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Aug. 2011, pp. 437–440.
[77] A.-R. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2011, pp. 5060–5063.
[78] L. Deng and D. Yu, "Deep convex network: A scalable architecture for speech pattern classification," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Aug. 2011, pp. 2285–2288.
[79] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5688–5691.
[80] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2011, pp. 24–29.
[81] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Aug. 2011, pp. 237–240.
[82] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Large vocabulary continuous speech recognition with context-dependent DBN-HMMS," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2011, pp. 4688–4691.
[83] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted Boltzmann machines," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 1, May 2011, pp. 5884–5887.
[84] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A.-R. Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2011, pp. 30–35.

[85] G. Zweig *et al.*, "Speech recognition with segmental conditional random fields: Final report from the 2010 JHU summer workshop," Baseline, Tech. Rep., 2010, pp. 7–10.

[86] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[87] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of Context-Dependent Deep Neural Networks for Automatic Speech Recognition," in *Proc. SLT*, 2012, pp. 366–369.

[88] D. Yu, S. M. Siniscalchi, L. Deng, and C.-H. Lee, "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4169–4172.

[89] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[90] D. Yu, F. Seide, and G. Li, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 1–2.

[91] D. Yu, F. Seide, G. Li, and L. Deng, "Exploiting sparseness in deep neural networks for large vocabulary speech recognition," in *Proc. ICASSP*, 2012, pp. 4409–4412.

[92] D. Yu, X. Chen, and L. Deng, "Factorized deep neural networks for adaptive speech recognition," in *Proc. Int. Workshop Stat. Mach. Learn. Speech Process.*, 2012, pp. 1–5.

[93] J. Li, D. Yu, J.-T. Huang, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM," in *Proc. IEEE Workshop Spoken Lang. Technol. (SLT)*, Dec. 2012, pp. 131–136.

[94] Y. Kubo, T. Hori, and A. Nakamura, "Integrating deep neural networks into structural classification approach based on weighted finite-state transducers," in *Proc. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2012, pp. 2594–2597.

[95] Y. Bao, H. Jiang, C. Liu, Y. Hu, and L. Dai, "Investigation on dimensionality reduction of concatenated features with deep neural network for LVCSR systems," in *Proc. Int. Conf. Signal Process. (ICSP)*, vol. 1, no. 1, Oct. 2012, pp. 562–566.

[96] L. Deng, D. Yu, and J. Platt, "Scalable stacking and learning for building deep architectures," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 2133–2136.

[97] L. Deng, G. Tur, X. He, and D. Hakkani-Tür, "Use of kernel deep convex networks and end-to-end learning for spoken language understanding," in *Proc. IEEE Workshop Spoken Lang. Technol. (SLT)*, Dec. 2012, pp. 210–215.

[98] P. Zhou, C. Liu, Q. Liu, L. Dai, and H. Jiang, "A cluster-based multiple deep neural networks method for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6650–6654.

[99] L. Deng, O. Abdel-Hamid, and D. Yu, "A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2013, pp. 6669–6673.

[100] Z. Yan, Q. Huo, and J. Xu, "A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2013, pp. 104–108.

[101] X. Lei, A. Senior, A. Gruenstein, and J. Sorensen, "Accurate and compact large vocabulary speech recognition on mobile devices," in *Proc. Interspeech*, Aug. 2013, pp. 662–665.

[102] A. Senior, G. Heigold, M. Ranzato, and K. Yang, "An empirical study of learning rates in deep neural networks for speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2013, pp. 6724–6728.

[103] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2013, pp. 7398–7402.

[104] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2013, pp. 7596–7599.

[105] S. Matsuda, X. Lu, and H. Kashioka, "Automatic localization of a language-independent sub-network on deep neural networks trained by multi-lingual speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7359–7362.

[106] J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2013, pp. 7304–7308.

[107] T. N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. ICASSP*, May 2013, pp. 8614–8618.

[108] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3687–3691.

[109] Y. Miao, F. Metze, and S. Rawat, "Deep maxout networks for low-resource speech recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 398–403.

[110] M. Cai, Y. Shi, and J. Liu, "Deep maxout neural networks for speech recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2013, pp. 291–296.

[111] M. Henderson, B. Thomson, and S. Young, "Deep neural network approach for the dialog state tracking challenge," in *Proc. SIGDIAL Conf.*, 2013, pp. 467–471.

[112] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2013, pp. 6704–6708.

[113] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 309–314.

[114] H. Su, G. Li, D. Yu, and F. Seide, "Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2013, pp. 6664–6668.

[115] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Aug. 2013, pp. 3366–3370.

[116] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2013, pp. 7942–7946.

[117] P. Swietojanski, A. Ghoshal, and S. Renals, "Revisiting hybrid and GMM-HMM system combination techniques," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6744–6748.

[118] L. Li *et al.*, "Hybrid deep neural network–hidden Markov model (DNN-HMM) based speech emotion recognition," in *Proc. Hum. Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 312–317.

[119] A. Graves, N. Jaitly, and A.-R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 273–278.

[120] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2013, pp. 7092–7096.

[121] T. N. Sainath *et al.*, "Improvements to deep convolutional neural networks for LVCSR," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2013, pp. 315–320.

[122] G. E. Dahl, T. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2013, pp. 8609–8613.

[123] B. Li and K. C. Sim, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2013, pp. 279–284.

[124] K. M. Knill, M. J. F. Gales, S. P. Rath, P. C. Woodland, C. Zhang, and S.-X. Zhang, "Investigation of multilingual deep neural networks for spoken term detection," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2013, pp. 138–143.

[125] G. Mesnil, X. He, L. Deng, Y. Bengio, and F. Flight, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding," in *Proc. Interspeech*, vol. 2, 2013, pp. 3771–3775.

[126] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7893–7897.

[127] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2013, pp. 368–373.
[128] T. N. Sainath, B. Kingsbury, A.-R. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2013, pp. 297–302.
[129] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2013, pp. 6655–6659.
[130] G. Heigold *et al.*, "Multilingual acoustic models using distributed deep neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2013, pp. 8619–8623.
[131] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2013, pp. 7319–7323.
[132] B. Li and K. C. Sim, "Noise adaptive front-end normalization based on Vector Taylor Series for Deep Neural Networks in robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2013, pp. 7408–7412.
[133] M. D. Zeiler *et al.*, "On rectified linear units for speech processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3517–3521.
[134] P.-S. Huang, K. Kumar, C. Liu, Y. Gong, and L. Deng, "Predicting speech recognition confidence using deep learning with word identity and score features," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 1, May 2013, pp. 7413–7417.
[135] O. Abdel-Hamid and H. Jiang, "Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition," in *Proc. Interspeech*, Jan. 2016, pp. 1248–1252.
[136] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. 30th Int. Conf. Mach. Learn.*, vol. 28, 2013, p. 6.
[137] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. Interspeech*, Aug. 2013, pp. 2365–2369.
[138] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, vol. 1, 2013, pp. 2345–2349.
[139] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7947–7951.
[140] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 55–59.
[141] N. Ryant, M. Y. Liberman, J. Yuan, N. Ryant, M. Y. Liberman, and J. Yuan, "Speech activity detection on YouTube using deep neural networks," in *Proc. Interspeech*, 2013, pp. 728–731.
[142] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 6, May 2013, pp. 6645–6649.
[143] D. Yu, L. Deng, and F. Seide, "The deep tensor neural network with applications to large vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 2, pp. 388–396, Feb. 2013.
[144] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.
[145] Y. Huang, D. Yu, C. Liu, and Y. Gong, "A comparative analytic study on the Gaussian mixture and context dependent deep neural network hidden Markov models," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Sep. 2014, pp. 1895–1899.
[146] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1695–1699.
[147] B. Li and K. C. Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 8, pp. 1296–1305, Aug. 2014.
[148] P. Karanasou, Y. Wang, M. J. F. Gales, and P. C. Woodland, "Adaptation of deep neural network acoustic models using factorised I-vectors," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Sep. 2014, pp. 2180–2184.
[149] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.
[150] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. J. Moreno, "Automatic language identification using deep neural networks," in *Proc. ICASSP*, May 2014, pp. 5337–5341.
[151] M. Bacchiani and D. Rybach, "Context dependent state tying for speech recognition using deep neural network acoustic models," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2014, pp. 230–234.
[152] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1120–1124, Sep. 2014.
[153] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. 39th IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, May 2014, pp. 1562–1566.
[154] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, "Deep neural network based speech separation for robust speech recognition," in *Proc. 12th Int. Conf. Signal Process.*, Oct. 2014, pp. 532–536.
[155] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and M. J. Alam, "Deep Neural Networks for extracting Baum-Welch statistics for speaker recognition," in *Proc. Odyssey-Speak. Lang. Recognit. Workshop*, Jun. 2014, pp. 293–298.
[156] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2014, pp. 4052–4056.
[157] S. Xue, O. Abdel-Hamid, H. Jiang, and L. Dai, "Direct adaptation of hybrid DNN/HMM model for fast speaker adaptation in LVCSR based on speaker code," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2014, pp. 6339–6343.
[158] Y. Xu, J. Du, L. Dai, and C. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2014, pp. 2670–2674.
[159] L. Deng and J. C. Platt, "Ensemble deep learning for speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH()*, Sep. 2014, pp. 1915–1919.
[160] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Sep. 2014, pp. 2685–2689.
[161] J. Li, J.-T. Huang, and Y. Gong, "Factorized adaptation for deep neural network," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2014, pp. 5537–5541.
[162] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1713–1725, Dec. 2014.
[163] Y. Xu, J. Du, L. Dai, and C.-H. Lee, "Global variance equalization for improving deep neural network based speech enhancement," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP)*, Jul. 2014, pp. 71–75.
[164] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2014, pp. 215–219.
[165] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with I-vector inputs," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 225–229.
[166] A. Narayanan and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 92–101, Jan. 2015.
[167] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2014, pp. 6334–6338.
[168] D. Chen, B. Mak, C.-C. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2014, pp. 5629–5633.

[169] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2014, pp. 2504–2508.

[170] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Sep. 2014, pp. 1910–1914.

[171] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4661–4665.

[172] S. Wiesler, A. Richard, R. Schlüter, and H. Ney, "Mean-normalized stochastic gradient for large-scale deep learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2014, vol. 1, no. 2, pp. 180–184.

[173] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Proc. ICASSP*, May 2014, pp. 7639–7643.

[174] X. Meng, C. Liu, Z. Zhang, and D. Wang, "Noisy training for deep neural networks," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, Xi'an, China, 2014, pp. 16–20.

[175] C. Weng, D. Yu, S. Watanabe, and B.-H. F. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, May 2014, pp. 5532–5536.

[176] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in *Proc. COLING*, 2011, pp. 2335–2344.

[177] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Sep. 2014, pp. 616–620.

[178] L. Deng and J. Chen, "Sequence classification using the high-level features extracted from deep neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2014, pp. 6844–6848.

[179] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, "Single-channel mixed speech recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2014, pp. 5669–5673.

[180] J. Xue, J. Li, D. Yu, M. L. Seltzer, and Y. Gong, "Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2014, pp. 6359–6363.

[181] M. Rouvier and B. Favre, "Speaker adaptation of DNN-based ASR with I-vectors: Does it actually adapt models to speakers?" *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Sep. 2014, pp. 3007–3011.

[182] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. 15th Annu. Conf.*, Sep. 2014, pp. 223–227.

[183] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *Proc. 9th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Sep. 2014, pp. 250–254.

[184] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. 12th Int. Conf. Signal Process. (ICSP)*, Oct. 2014, pp. 473–477.

[185] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," in *Proc. Interspeech*, vol. 20, Sep. 2014, pp. 2189–2193.

[186] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition," in *Proc. IEEE Workshop Spoken Lang. Technol. (SLT)*, Dec. 2014, pp. 135–140.

[187] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. ICASSP*, Apr. 2015, pp. 4395–4399.

[188] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[189] Y. Liu, P. Karanasou, and T. Hain, "An investigation into speaker informed DNN front-end for LVCSR," in *Proc. ICASSP*, vol. 1, Apr. 2015, pp. 4300–4304.

[190] H. Huang and K. C. Sim, "An investigation of augmenting speaker representations to improve speaker normalisation for DNN-based speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4610–4613.

[191] A. L. Maas *et al.*, "Building DNN acoustic models for large vocabulary speech recognition," *Comput. Speech Lang.*, vol. 41, pp. 195–213, Jan. 2017.

[192] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Aug. 2015, pp. 4580–4584.

[193] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1671–1675, Oct. 2015.

[194] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 26, pp. 1–12, 2015.

[195] S. H. K. Parthasarathi, B. Hoffmeister, S. Matsoukas, A. Mandal, N. Strom, and S. Garimella, "fMLLR based feature-space speaker adaptation of DNN acoustic models," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, vol. 1, 2015, pp. 3630–3634.

[196] P. Motlicek, D. Imseng, B. Potard, P. N. Garner, and I. Himawan, "Exploiting foreign resources for DNN-based ASR," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 17, pp. 1–10, 2015.

[197] R. F. Astudillo, J. Correia, and I. Trancoso, "Integration of DNN based Speech Enhancement and ASR," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2015, pp. 3576–3580.

[198] G. Pironkov, S. Dupont, and T. Dutoit, "Investigating sparse deep neural networks for speech recognition," in *Proc. ASRU*, Dec. 2015, pp. 124–129.

[199] Y.-H. Tu, J. Du, L.-R. Dai, and C.-H. Lee, "Speech Separation based on signal-noise-dependent deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 61–65.

[200] X. L. Zhang and D. Wang, "Multi-resolution stacking for speech separation based on boosted DNN," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, vol. 1, 2015, pp. 1745–1749.

[201] S. Yin *et al.*, "Noisy training for deep neural networks in speech recognition," *EURASIP J. Audio Speech Music Process.*, vol. 2015, no. 1, pp. 1–14, 2015.

[202] S.-Y. Chang and S. Wegmann, "On the importance of modeling and robustness for deep neural network feature," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4530–4534.

[203] L. Tóth, "Convolutional deep maxout networks for phone recognition," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2014, pp. 1078–1082.

[204] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann, "Real-time dereverberation for deep neural network speech recognition coherence based spectral enhancement," in *Proc. DAGA*, Erlangen, Germany, 2015, pp. 1–4.

[205] S. Sivasankaran *et al.*, "Robust ASR using neural network based speech enhancement and feature simulation," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2015, pp. 482–489.

[206] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann, "Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments," in *Proc. ICASSP*, Apr. 2015, pp. 4380–4384.

[207] M. J. Alam, V. Gupta, P. Kenny, and P. Dumouchel, "Speech recognition in reverberant and noisy environments employing multiple feature extractors and i-vector speaker adaptation," *J. Adv. Signal Process.*, vol. 2015, no. 1, p. 50, 2015.

[208] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2016, pp. 92–97.

[209] Y.-H. Liao, H.-Y. Lee, and L.-S. Lee, "Towards structured deep neural network for automatic speech recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2015, pp. 3–6.

[210] P. Dighe, G. Luyet, A. Asaei, and H. Bourlard, "Exploiting low-dimensional structures to enhance DNN based acoustic modeling in speech recognition," in *Proc. ICASSP*, Mar. 2016, pp. 5690–5694.

[211] Y. Yu, W. Wang, and P. Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks," *EURASIP J. Audio, Speech, Music Process.*, vol. 2016, no. 7, pp. 1–18, 2016.

[212] X. Wang, L. Wang, J. Chen, and L. Wu, "Toward a better understanding of deep neural network based acoustic modelling: An empirical investigation," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 2173–2179.

[213] R. Price, K. ichi Iso, and K. Shinoda, "Wise teachers train better DNN acoustic models," *EURASIP J. Audio, Speech, Music Process.*, vol. 2016, no. 10, pp. 1–19, 2016.
[214] L. Lu and S. Renals, "Small-footprint highway deep neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1502–1511, Jul. 2017.
[215] J.-M. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process. (MMSP)*, Aug. 2017, pp. 1–5.
[216] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "A network of deep neural networks for Distant Speech Recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4880–4884.
[217] G. Cong, B. Kingsbury, S. Gosh, G. Saon, and F. Zhou, "Accelerating deep neural network learning for speech recognition on a cluster of GPUs," in *Proc. Mach. Learn. HPC Environ.*, 2017, Art. no. 3.
[218] B. Wu *et al.*, "An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 99, pp. 1289–1300, Sep. 2017.
[219] M. Kim, "Collaborative deep learning for speech enhancement: A run-time model selection method using autoencoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 76–80.
[220] K. J. Han, S. Hahm, B. H. Kim, J. Kim, and I. Lane, "Deep learning-based telephony speech recognition in the wild," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. Belmont, CA, USA: Capio Inc., 2017, pp. 1323–1327, Aug. 2017.
[221] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. INTERSPEECH*, 2017, pp. 999–1003.
[222] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Netw.*, vol. 92, pp. 60–68, Aug. 2017.
[223] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman, and M. Tiede, "Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition," *Speech Commun.*, vol. 89, pp. 103–112, May 2017.
[224] Z. Chen, Y. Huang, J. Li, and Y. Gong, "Improving mask learning based speech enhancement system with restorationlayers and residual connection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Aug. 2017, pp. 3632–3636.
[225] T. N. Sainath *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 5, pp. 965–979, May 2017.
[226] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee. (2017). "Multi-objective learning and mask-based post-processing for deep neural network based speech Enhancement." [Online]. Available: https://arxiv.org/abs/1703.07172
[227] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. Hands-Free Speech Commun. Microphone Arrays (HSCMA)*, Mar. 2017, pp. 136–140.
[228] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.(ICASSP)*, New Orleans, LA, USA, 2017, pp. 1–5.
[229] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2017, pp. 4845–4849.
[230] K. Li, S. Mao, X. Li, Z. Wu, and H. Meng, "Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks," *Speech Commun.*, vol. 96, pp. 28–36, Feb. 2018.
[231] W. Shi *et al.*, "Deep neural network based monaural speech enhancement with low-rank analysis and speech present probability," *IEICE Trans. Fundamentals Electron., Commun. Comput. Sci.*, vol. E101A, no. 3, pp. 585–589, 2018.
[232] P. Song and W. Zheng, "Feature selection based transfer subspace learning for speech emotion recognition," *IEEE Trans. Affect. Comput.*, to be published. doi: 10.1109/TAFFC.2018.2800046.

ALI BOU NASSIF received the master's degree in computer science and the Ph.D. degree in electrical and computer engineering from Western University, Canada, in 2009 and 2012, respectively. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering and also the Assistant Dean of graduate studies with the University of Sharjah, United Arab Emirates. He is also an Adjunct Research Professor with Western University, Canada. His research interests include the applications of statistical and artificial intelligence models in different areas such as software engineering, electrical engineering, e-learning, security, signal processing, and social media. He is a Registered Professional Engineer in ON, Canada, and a member of the IEEE Computer Society.

ISMAIL SHAHIN received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Southern Illinois University, Carbondale, USA, in 1992, 1994, and 1998, respectively. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Sharjah, United Arab Emirates. He has more than 55 journal and conference publications. He has remarkable contribution in organizing many conferences, symposiums, and workshops. His research interests include speech recognition, speaker recognition under neutral, stressful, and emotional talking conditions, emotion and talking condition recognition, gender recognition using voice, and accent recognition.

IMTINAN ATTILI received the B.Sc. degree in electrical and communications engineering from United Arab Emirates University, in 2009, and the M.Sc. degree in project management from The British University in Dubai, in 2013, where she is currently pursuing the M.Sc. degree in electrical and electronics engineering. She has been a Lab Engineer with the University of Sharjah, since 2010. Her research interest include mixed analog digital IC design, low voltage mixed mode CMOS circuits, transconductance, and operational amplifier circuit design.

MOHAMMAD AZZEH received the M.Sc. degree in software engineering from the University of the West of England, Bristol, U.K., and the Ph.D. degree in computing from the University of Bradford, Bradford, U.K., in 2010. He is currently an Associate Professor with the Faculty of Information Technology, Department of Software Engineering, Applied Science Private University. He has published more than 40 research articles in reputable journals and conferences, such as *IET Software*, *Software: Evolution and Process*, *Empirical Software Engineering*, *Applied Soft Computing*, and the *Journal of Systems and Software*. His research interests include software cost estimation, empirical software engineering, data science, mining software repositories, and machine learning for software engineering problems. He was a Conference Chair of CSIT2016 and CSIT2018, and he is a Co-Chair of many IT-related workshops.

KHALED SHAALAN received the B.Sc. degree in computer science (artificial intelligence and software engineering), the M.Sc. degree in informatics and the M.Sc. degree in IT management, and the Ph.D. degree in computer science. He is currently the Head of Programmes. He is also a Full Professor of computer science with The British University in Dubai, United Arab Emirates (UAE). He is also an Honorary Fellow with the School of Informatics, University of Edinburgh, U.K. He has a long experience in teaching in computer science for both core and advanced undergraduate and postgraduate levels. He has taught more than 30 different courses at the undergraduate and postgraduate levels. Over the last two decades, he has been contributing to a wide range of research topics in Arabic Natural Language Processing, including machine translation, parsing, spelling and grammatical checking, named entity recognition, and diacritization. Moreover, he has also worked on topics in knowledge management, knowledge-based systems, knowledge engineering methodology, including expert systems building tools, expert systems development, and knowledge verification. Nevertheless, he worked on health informatics topics, including context-aware knowledge modeling for decision support in E-Health and game-based learning. Furthermore, he worked in educational topics, including intelligent tutoring, item banking, distance learning, and mobile learning. He has been the Principal Investigator or Co-Investigator on research grants from USA, U.K., and UAE funding bodies. He has published more than 164 referred publications and the impact of his research using GoogleScholar's H-index metric is 28. He has several research publications in his name in highly reputed journals, such as *Computational Linguistics*, the *Journal of Natural Language Engineering*, the *Journal of the American Society for Information Science and Technology*, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *Expert Systems with Applications*, *Software-Practice and Experience*, the *Journal of Information Science*, *Computer Assisted Language Learning*, and the *European Journal of Scientific Research*. His research work is cited extensively worldwide (see his Google Scholar citation indices). He has guided several master's and Ph.D. students in Arabic natural language processing, healthcare, intelligent tutoring systems, and knowledge management. He encourages and supports his students in publishing at highly ranked journals and conference proceedings. He has been actively and extensively supporting the local and international academic community. He has participated in seminars and invited talks locally and internationally, invited to international group meetings, invited to review papers from leading conferences and premier journals in his field, and invited for reviewing promotion applications to the ranks of an Associate and a Full Professor for applicants from both British and Arab Universities. He is the Founder and a Co-Chair of the International Conference on Arabic Computational Linguistic. He is an Associate Editor of *ACM Transactions of Asian and Low Resource Language Information Processing* Editorial Board and the Association for Computing Machinery.