Ain Shams Engineering Journal 16 (2025) 103405

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

Review Article

# Deep neural networks for speech enhancement and speech recognition: A systematic review

Sureshkumar Natarajan [a,*], Syed Abdul Rahman Al-Haddad [a,*], Faisul Arif Ahmad [a], Raja Kamil [b], Mohd Khair Hassan [b], Syaril Azrad [c], June Francis Macleans [d], Sadiq H. Abdulhussain [e], Basheera M. Mahmmod [e], Nurbek Saparkhojayev [f], Aigul Dauitbayeva [g]

[a] Department of Computer and Communication Systems Engineering, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[b] Department of Electrical and Electronic Engineering, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[c] Department of Aerospace Engineering, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[d] Independent Researcher, Malaysia
[e] Department of Computer Engineering, University of Baghdad, Iraq
[f] Rudny Industrial University, Kazakhstan
[g] Department of Computer Science, Korkyt Ata Kyzylorda State University, Kazakhstan

ARTICLE INFO

*Keywords:*
Acoustic modeling
Denoising
Reverberation
Beamforming
Speech enhancement
Speech recognition
Machine learning
Deep neural network
Systematic review

ABSTRACT

The field of speech signal processing has undergone significant transformation through extensive research. There is growing interest in Speech Enhancement (SE) and Automatic Speech Recognition (ASR), with SE serving as a crucial preliminary step to enhance ASR performance. This paper addresses key challenges, particularly the need to maintain speech quality and improve intelligibility in ASR systems. Recently, deep learning techniques have emerged as powerful tools for tackling these challenges. This systematic review examines speech enhancement and recognition techniques, emphasizing denoising, acoustic modeling, and beamforming. Various deep learning architectures, such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, and Hybrid Neural Networks, are reviewed to highlight their roles in enhancement and recognition. The review specifically details their usage, the features utilized in each study, the databases employed, performance, and limitations, all presented in a structured tabular format. This approach provides valuable insights into the strengths and weaknesses of each method, guiding future advancements in the field. In particular, it emphasizes that LSTM-RNN models excel in temporal signal processing, while hybrid models demonstrate superior performance in optimizing task outcomes. The paper conducts a comprehensive statistical analysis of 187 research papers that exclusively utilize deep neural networks to address the challenges of speech enhancement and recognition, presenting the latest advances in the field. The review examines publications from 2012 to 2024, shedding light on research trends and patterns, while the proposed solutions aim to bridge gaps for researchers in this evolving domain.

## 1. Introduction

Speech information transfer in human communication is of great importance and has become increasingly associated with advanced technology in recent decades. Many studies have focused on speech enhancement and speech recognition due to their importance in daily applications [146,189]. Various issues that arise when speech data is captured at a distance from its source such as in the case of far-field communication. Environmental daily scenarios lead to speech degradation. Stationary and non-stationary background noise, convoluted reverberations, and unpredictable spectral characteristics, all of which greatly affect the characteristics of the speech signal [1]. To address

* Corresponding authors.
*E-mail addresses:* suresh45619@gmail.com (S. Natarajan), sar@upm.edu.my (S.A. Rahman Al-Haddad), faisul@upm.edu.my (F.A. Ahmad), kamil@upm.edu.my (R. Kamil), khair@upm.edu.my (M.K. Hassan), syaril@upm.edu.my (S. Azrad), macleans30june@gmail.com (J.F. Macleans), sadiqhabeeb@coeng.uobaghdad.edu.iq (S.H. Abdulhussain), basheera.m@coeng.uobaghdad.edu.iq (B.M. Mahmmod), nursp81@gmail.com (N. Saparkhojayev), AICOS@mail.ru (A. Dauitbayeva).

https://doi.org/10.1016/j.asej.2025.103405
Received 22 September 2024; Received in revised form 31 January 2025; Accepted 4 April 2025
Available online 30 April 2025

these difficult situations, researchers have been exploring different solutions for years, including, transform based processing, machine learning, deep learning, neural networks, and beamforming for distant speech enhancement and recognition [2–4,191–195,201,212]. Since speech data is inherently linked to language, as the primary means of communication in human society, all speech information needs to be encoded in an understandable language [5]. Speech information generally exists as an acoustic form of energy that is manipulated according to the desired form of information encoded by the receptor based on the desired process such as speech enhancement using the beamforming method and speech recognition based on deep neural network (DNN) [2,6,7,171].

Speech enhancement is a crucial and a primary stage in speech processing and can be applied to single-channel or multi-channel speech signals under different environmental conditions. This process involves separation of actual speech information from noisy audio signals [5,8,9]. Technological advances such as beamforming and DNN have been introduced to enhance speech quality and intelligibility [10]. This process involves extracting unwanted noise data and amplifying speech information for speech synthesis with maintaining high quality and minimal processing to obtain better synthesis results [11]. Although speech enhancement can be easily achieved in a near-field communications environment, additional techniques are needed to improve signal quality for far-field speech enhancement techniques and configurations. In general, the acoustic or speech signal is captured by external or internal microphones. Microphone array technology is particularly useful for remote channel speech enhancement, providing superior results in a reverberant environment compared to single-channel or 2-microphone setups [12]. Studies have explored different multi-microphone setups, such as 4-microphone and 6-microphone configurations, depending on the application, which provide improved performance in terms of quality, information, and sequential speech synthesis [13]. In speech enhancement techniques, multichannel filters outperform single-channel filters, and spatial information is essential for identifying target speakers through localization. Multichannel speech enhancement uses a larger array of input devices to amplify the target speaker's voice, a process known as beamforming, while blind source separation (BSS) can be achieved using small microphone arrays. This process involves considering specific parameters, such as direction of arrival and array geometry to choose the optimal filter coefficients [14]. Besides, spatial covariance matrices must be configured to enhance speech signal acoustic transfer functions and signals [15]. When considering all these factors, it is essential to ensure that the actual speech information is not corrupted or lost. Various methods have been introduced to address this concern. It is worth noting that the use of DNN in speech intelligibility perception model provides improved speech information [124], using short-term objective intelligibility measure with loss function, in addition to mean square error that improve the data and achieve better quality of speech enhancement [16].

Based on the previous discussion of speech enhancement and beamforming, it is important to note that beamforming includes different methods specifically designed for different speech sources and environmental conditions. The general types of these methods include adaptive beamforming [10,17] and MVDR beamforming [18]. In addition, there are other models for beamforming, such as NMF-informed [19], eigenvector [20,21], and permutation invariant beamforming, all of which excel in enhancing noisy speech signals in multichannel environments [15,22]. It is worth noting that recent developments have seen the application of DNN to beamforming, which has led to a significant improvement in performance for speech enhancement [14,188]. In the context of multichannel speech enhancement and speech recognition, beamforming is utilized for feature extraction to reduce processing time and aid in DNN training, thereby streamlining the overall speech processing time for the system [1,23].

Once speech data has been enhanced, the following step involves identifying and segregating informative speech data from the enhanced speech. Speech recognition, a cutting-edge model, is capable of accurately identifying a speaker's actual information by utilizing specific feature extraction components to improve the recognition of data [24,25]. With the ongoing research on speech enhancement in recent years, several techniques have been developed for speech recognition [26,27]. In the initial stages, machine learning was introduced for speech recognition, significantly impacting machine-to-human communication. During this period, the hidden Markov model was also incorporated into speech recognition. However, machine learning faced certain limitations with respect to meeting the requirements of acoustic speech data. Consequently, extensive research in the field of speech recognition led to the introduction of DNN, which provided superior results, especially in diverse environmental conditions such as various types of noises and reverberant signals [28]. Various deep learning or DNN models were utilized to segregate and extract all required elements [29–31]. The fundamental DNN introduced for speech recognition is the feed-forward neural network (FFNN), which is suitable for word recognition, focusing on desired vocabulary or predefined words. This system was designed to be less processing intensive, targeting specific words as per the defined training datasets. However, this approach has proven insufficient for robust, acoustic, or unknown speech recognition. Consequently, the basic DNN was further integrated with various DNN models such as the recurrent neural network (RNN), convolutional neural network (CNN), and long and short-term memory (LSTM) to align with the defined objectives for speech recognition [17,32,33]. These models were developed by creating training sets and conducting feature extraction from enhanced speech data to obtain better recognized speech information. Speech data in all these models were represented by time and frequency domain activities, and then later trained to achieve the required recognized speech information from noisy speech data [34]. By implementing the above models based on the research, speech separation was improved using advanced RNN [1,148,164].

Speech enhancement and speech recognition are critical technologies in modern human–computer interaction, enabling applications ranging from virtual assistants to accessibility tools. Recent advancements in deep neural networks (DNNs) have significantly enhanced the performance of these systems, particularly in complex and noisy environments. However, despite the strides made in improving accuracy, robustness, and noise resilience, there remain substantial challenges, including limited generalization to diverse acoustic environments, the computational complexity of DNN-based models, and the integration of enhancement and recognition tasks into a unified framework.

This systematic review synthesizes existing research on DNNs in the domains of speech enhancement and recognition, with a particular focus on insights gained from a statistical analysis of 187 papers. The review aims to provide a comprehensive evaluation of current methods and identify trends, challenges, and future directions based on this extensive body of work.

Research Objectives and Aims: The main objective of this review is to analyze and synthesize the current state of DNN-based approaches to speech enhancement and speech recognition. Specifically, this review aims to:

1. Evaluate the effectiveness of various DNN architectures (e.g., CNNs, RNNs, LSTM, Transformer, and hybrid models) in improving speech enhancement and recognition performance across different noise conditions.
2. Identify challenges and limitations in current DNN-based approaches including generalization to unseen noise types, real-time processing constraints, and computational complexity.
3. Analyze the most commonly used features in DNN-based systems and how these features contribute to performance improvements.
4. Examine widely used datasets and evaluation techniques, assessing their role in advancing DNN-based systems and their impact on performance benchmarks.

By addressing these objectives, this review aims to offer valuable insights for advancing future research in DNN-based speech enhancement and recognition.

## 2. Related works

In recent studies on speech enhancement and speech recognition, various surveys have been conducted. For instance, Hinton et al. [28] conducted an in-depth investigation that involved four different research groups using DNN acoustic modeling. This study marks a pivotal time for the field of speech following the advent of automatic speech recognition, showcasing the great progress made by DNN. The comprehensive overview encompasses a detailed analysis of GMM-HMM and deep belief network models, shedding light on various speaker-independent phonetic models with PER accuracies, thus facilitating meticulous model analysis. Central to this study is the TIMIT database, which plays a crucial role in driving progress. A noteworthy aspect is the comparative study between DBN-DNN and GMM for large vocabulary in speech recognition, including a thorough examination of DNN-HMM [149] and GMM-HMM based models. It is crucial to note the focus on DNN techniques, which have outperformed traditional GMM-HMM methods; however, they face challenges in terms of complete optimizations and efficient time dominance, particularly in adverse real-time conditions.

Li Deng et al. [35] conducted an in-depth analysis of all types of DNN utilized in speech recognition applications. Their exploration led to proposing measures to enhance deep learning methods, encompassing improved optimization, enhanced neural activation functions, advanced network architecture, methods for determining numerous hyper-parameters of DNN, and ways to leverage multiple languages or Gaussian mixture models (GMM). Highlighting the latest techniques and their impact on upcoming signal processing and speech recognition applications, this study provides valuable insights. However, it's crucial to note the absence of a discussion centering on the issues faced in overall acoustic models. A key point from their study emphasizes the effectiveness of implementing a DNN in the training segment of the model, thus enhancing performance and optimizing the trainable datasets [36].

Ajay Shrestha and Ausif Mahmood [37] have provided a comprehensive overview of deep learning, focusing on classifications, implementation, and frameworks. Their analysis offers a deep dive into the trends in deep learning frameworks, encompassing a wide range of network classifications, from feed-forward neural networks to modular networks, and exploring the common applications developed for various datasets. The overview emphasizes the popularity of the TensorFlow framework as the most widely used deep learning algorithm. Additionally, it sheds light on the distinctions between algorithms and outlines the existing shortcomings that call for improvement. The overview concludes with a thought-provoking reflection on the ambiguities of deep learning, leaving readers with a stimulating perspective to contemplate.

Nassif et al. emphasized the remarkable outcomes achieved using DL approaches in different speech-related applications, highlighting their superiority over other methods [26]. Their work conducted a comprehensive statistical analysis of a wide range of research carried out since the beginning of DL as a new domain in machine learning for speech applications in 2006. From 2006 to 2018, a thorough analysis was done on a total of 174 research papers. These papers were mostly presented at various conferences and journals, with a considerable number of them being featured at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). The works were scrutinized with great attention to detail, aiming to extract useful insights and information pertaining to the subject matter. The comprehensive review offers invaluable insights into the existing research trends and identifies new areas of interest in the field of speech processing and speech recognition. It highlights the transformative impact of deep learning on this domain and its potential to revolutionize the future of speech processing applications. This paper offers a detailed examination of the theoretical research analysis in machine learning, particularly deep learning for speech applications, it does not delve into the recent resources and advancements that can be utilized to practically implement these techniques for developing sophisticated speech recognition systems. This omission may limit the paper's comprehensiveness in addressing the current state-of-the-art and prospects in the field, which is crucial for the development of advanced speech recognition systems.

Ibrahim et al. presented a comprehensive review of Automatic Speech Recognition (ASR) Systems, emphasizing their architecture, recent advancements utilizing neural networks, and proposing a novel direction for future research [38]. The study emphasizes the need for continued advancements in ASR systems, particularly in the context of integrating speech processing and computer vision for enhanced user-machine interaction. This research paper comprehensively explores the two extremes of Automatic Speech Recognition (ASR) systems, addressing both fundamental audio aspects and recent advancements. However, it lacks a detailed discussion on the techniques and methodologies that have been instrumental in developing high-accuracy systems within this domain, which could have provided valuable insights for future enhancements.

Lee et al. discussed the challenges and advancements in voice recognition technology using deep learning techniques [39]. The commonly used voice recognition algorithm that uses DNN or a combination of DNN and HMM has shown better performance than conventional acoustic models. However, the recognition rate was significantly lower in noisy environments. To improve speech recognition performance, researchers actively studied noise reduction techniques such as spectral subtraction, spectral masking, and statistical methods. The paper also highlighted the challenges in lip-reading using deep learning, such as difficulties in accurately measuring changes in lip movement, extracting features of lips due to external factors, and phonemes with the same mouth shape. The review also highlighted the possibility of developing speech recognition using sensors in conjunction with deep learning techniques. The sensor-based voice recognition deep learning model can be developed and applied to systems or wearable devices, enabling personalized speech recognition and future artificial intelligence services that can collaborate with humans.

Alharbi et al. (2021) conducted a comprehensive investigation of ASR, focusing on the main themes that have emerged in recent years and the significant challenges faced in practical scenarios [40]. The study involved an analysis of literature from five different research databases, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The authors identified 82 relevant conferences and articles published between 2015 and 2020 that matched the study's scope.

Dhanjal et al. in their study [41] referred to papers from the years 2015 to 2021 and focused on current research issues facing the implementation of speech recognition. The authors analyzed existing research works, evaluation metrics, datasets, and tools used in the latest study, which are the factors that were considered in their work. Additionally, the author discussed the contribution of deep learning methods, which have played a crucial role in enhancing the performance of ASR. The findings indicate that both Mozilla and Facebook have made significant contributions to the field of speech recognition using neural networks. The main aim of the study is to provide direction for new researchers by highlighting the advancements made in the field of speech recognition using neural networks.

Kheddar et al. discussed the significant challenges faced by the field of ASR due to recent advancements in DL [42]. These challenges included the need for extensive training datasets, high computational resources, and the requirement for adaptive systems to perform well in dynamic environments. Traditional DL techniques often assumed that training and testing data came from the same domain, which was not always the case. To address these issues, innovative DL approaches like

deep transfer learning (DTL), federated learning (FL), and deep reinforcement learning (DRL) emerged. With DTL, high-performance models could be achieved using small yet related datasets, while FL allowed training on confidential data without dataset possession. DRL optimized decision-making in dynamic environments, reducing computation costs. This study provided a comprehensive review that focused on DTL, FL, DRL-based ASR frameworks, and their application to Transformers, an advanced DL technique known for capturing extensive dependencies in the input ASR sequence. The review began by providing a thorough background on DTL, FL, DRL, and Transformers. It then adopted a well-structured taxonomy to outline the state-of-the-art (SOTA) approaches in the field. A critical analysis was conducted to identify the strengths and weaknesses of each framework, followed by a comparative study to highlight the existing challenges and open up avenues for future research opportunities. This research aimed to provide valuable insights for researchers and professionals in the field of ASR, aiding them in understanding the current challenges and potential solutions.

Douglas O Shaughnessy presented a comprehensive analysis of speech enhancement techniques, their practical applications, and the challenges faced in enhancing the quality of distorted speech signals [43]. The main objective of the research paper was to provide a detailed understanding of classical SE techniques such as the spectral subtraction method, wiener filtering method, and a few other methods and their significance in various practical contexts, with a particular emphasis on the importance of monaural enhancement. The study included an in-depth review of SE techniques and their applications across a variety of contexts. Moreover, the paper examined the significant components of DNN architectures commonly used in SE applications, such as CNNs, RNNs, Autoencoders, Attention mechanisms, Multichannel DNN-SE, and Codebook SE. These components played a critical role in processing speech signals, managing temporal contexts, focusing on specific data segments, and enhancing speech quality using various techniques, such as filtering, encoding/decoding, and beamforming. This discussion provided valuable insights into the computational complexity, information processing capabilities, and potential improvements in SE methods using DNN architectures. The paper also discussed the difficulties involved in determining and evaluating acoustic measures that approximate PESQ and provided suggestions for improving these methods.

Some of speech enhancement and speech recognition works based on deep learning have been explained in detail in sections 2.1 and 2.2.

### 2.1. Speech enhancement

Deep learning has led to significant improvements in the performance of the speech enhancement task, where deep neural networks are trained to recover clean speech signals from noisy mixtures. Many works have been proposed over the past decades to address this problem and have shown good improvement [196].

Xu et al. explored the use of DNN in speech enhancement, introducing a regression-based framework with multiple layers of deep networks that simulated 100 h of speech data, proving to be vital for achieving state-of-the-art performance due to its advanced capabilities [44]. However, in reverberant speech, the temporal correlation of consecutive frames becomes weak at low RT60. To address this, utilizing more acoustic context information and refining DNN learning can improve the continuity of enhanced speech [45]. Furthermore, DNN has effectively impacted speech enhancement by de-reverberation in a single-channel system [25]. A reverberation-time-aware DNN-based model, taking into account frame shift, speech framing, and acoustic context size, proved to be effective in de-reverberation. During the training stage, the module is trained on a multi-condition dataset containing pairs of reverberant and anechoic speech represented by log power spectra (LPS), using a linear activation function in the output layer [25]. In 2017, Ming Tu and Xianxian Zhang picked up from these findings and further utilized DNN for speech enhancement by introducing skip connections into a feed-forward network-based architecture, which were established between the input and output networks. The ideal ratio mask was estimated using a DNN [9]. Building on this foundation, in 2018, Karjol et al. [46] introduced a multiple DNN for speech enhancement, where each DNN contributes to obtaining enhanced speech. This system utilizes a gating network that provides weights for combining the outputs of the individual DNNs. Furthermore, it was found that the DNN was applied to estimate the clean speech spectrum, calculate the weighted average, and use the average value for joint training of multiple DNNs, resulting in a 0.07 % improvement in signal-to-noise ratio compared to a single DNN-based system. However, challenges arise when implementing multiple DNN acoustic models together due to increased mismatch during training, which leads to performance degradation at lower SNRs. Additionally, it's noteworthy that the TIMIT dataset, which offers very controlled conditions, can impact performance analysis under various conditions [46].

Thanh T. Vu, Benjamin Bigot, and Eng Siong Chng [47] examined the CHIME-3 Challenge, focusing on robust speech recognition and speech enhancement using beamforming and non-negative matrix analysis. The objective was to isolate and identify speech from multi-channel recordings in noisy environments. The challenge involved 12 speakers and six microphone arrays positioned on a tablet in a noisy setting. Their approach involved channel selection based on cross-correlation with minimum variance distortion less response (MVDR) beamforming, and non-negative matrix factorization for speech enhancement. Signal processing was carried out using DNN and RNN- language model (RNN-LM). The system's front and back ends were managed independently by combining cross-correlation and DNN acoustic models designed on fMLLR features and an RNN language model, respectively. The results demonstrated a Word Error Rate (WER) of 11.94 % on real noisy recordings, representing a 65 % improvement from the baseline challenge. However, the method's computational complexity and power usage could be further optimized to enhance speech intelligibility and mitigate noise.

Anurag Kumar and Dinei Florencio [48] presented a system for speech enhancement under multi-noise conditions using DNN. The model undergoes noise-aware training to effectively identify and segregate multiple noises at the input side. To minimize the error rate of clean speech, they introduced the weighted square error (WSE). Then, a study was conducted by Ming Tu and Xianxian [9] on speech enhancement utilizing DNN, demonstrating an innovative approach to handling a standard feed-forward network structure for learning the non-linear mapping function from input to output. Their proposed method incorporated a skip connection with the identity weight matrix between the input–output network, along with experiments involving the stacking of multiple network blocks. This led to the improvement of the results and developed a method for learning a log-compressed ideal ratio mask. The PESQ values were higher in sDNN2 compared to sDNN1, with similar STOI values. Where, at 5db, the STOI values were 0.854 and 0.853 for sDNN1 and sDNN2, respectively. While the models performed better than the baseline DNN, they did not exhibit significant improvement relative to each other.

Shuai Nie et al. [49] developed an advanced speech enhancement system for single-channel applications. This system used deep learning to determine the presence of speech and updated noise statistics for the Weiner filter-based speech enhancement framework, resulting in improved speech quality. The system's performance was thoroughly evaluated using important parameters such as Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Artifact Ratio (SAR), as well as average gains of SDR and Perceptual Evaluation of Speech Quality (PESQ), demonstrating superior performance compared to existing methods. The authors demonstrated that their proposed deep noise tracking network (DNTN) outperformed a pure deep learning-based approach and had better generalization capabilities. Notably, the system's effectiveness was tested under various noise conditions, showing strong performance in unfamiliar noise environments. One of

the limitations of the proposed approach was that it relied heavily on the choice of the smoothing factor αx, which could be challenging to set in practice, especially in scenarios with varying noise conditions.

Jen-Cheng Hou et al. [50] explored an innovative approach to speech enhancement by utilizing CNN for both audio and visuals and integrating the resultant outputs into a unified network. The model was trained in an end-to-end format, employing an audio-visual encoder-decoder architecture. This heterogeneous system leveraged visuals for enhancement, evolving into a multi-task learning algorithm-based speech enhancement framework. Subjective listening tests indicated ratings of 1.95 for the audio CNN system and 2.95 for the audio-visual CNN system. Additionally, further improvements were achieved through the application of a fully connected network and U-NET. Augmenting the training data with real-world conditions has the potential to enhance system performance [50]. One limitation of the proposed AVDCNN model is its dependence on visual input quality, which may not be consistent or reliable in real-world scenarios, potentially affecting the overall performance of the speech enhancement algorithm. Furthermore, the model's effectiveness in practical applications may be affected by factors such as lighting conditions, camera angles, and background noise.

Addressing the limitations of single-channel speech enhancement methodologies, an inherent challenge arises from the distortion introduced during the filtering process. This challenge can be overcome through a multi-channel approach that incorporates spatial information. In 2020, Nicolas Furnon et al. proposed a solution by introducing a fully connected array of microphones and synchronized sensors, combined with a Convolutional Recurrent Neural Network (CRNN). A comparative study between Recurrent Neural Network (RNN) and CRNN illustrated the dominance of CRNN in effectively suppressing noise and significantly enhancing speech quality. However, it's important to note that the Wiener filter has a drawback of introducing unwanted music noise, and the iterative nature of the process is time-consuming, making it complex to implement for further application processes [51].

The deep auto-encoder (DAE) system is known to have issues with speech distortion, while mask-based speech enhancement algorithms have lower distortion but struggle to suppress noise effectively. In 2020, Yochai Yemini et al. introduced a combination of DAE with a mask-based algorithm, using DNN to overcome these issues. The system obtained speech presence probability and DNN output in the noisy log-spectrum domain. By using a Comp-Net approach that combines mask net and spec net, better noise suppression was achieved, leading to significant speech enhancement [52]. One potential limitation of this study is the lack of a comprehensive comparison with other state-of-the-art speech enhancement methods, which makes it difficult to fully assess the novelty and effectiveness of the proposed approach.

An unresolved issue was the system's struggle to perform well in very noisy environments and restore speech under such conditions. Yoshiaki Bando et al. [53] proposed an adaptive neural network, using two networks: one for denoising the VAE encoder to estimate a latent vector from a noisy signal, and the other for training and learning from the spectrogram using a feedback mechanism. This effectively increased the system's enhancement ability in terms of STOI, SDR, and PESQ. However, the proposed model required significant computational resources to train the denoising VAE and generative decoder network, which might not have been feasible for real-time implementation.

In 2020, a study conducted by Wenhao Yuan proposed an innovative method for speech enhancement that leveraged DNNs to extract local features of noisy speech in a causal manner [54]. The authors employed long short-term memory (LSTM) and CNN architectures to capture correlations in the time and frequency domains, respectively, drawing inspiration from improved minima-controlled recursive averaging (IMCRA) and the time–frequency correlation method. Additionally, a time–frequency smoothing neural network (TFSNN) was presented, incorporating elements of statistic-based speech enhancement methods to augment local feature calculations through time and frequency correlations. By emulating IMCRA's time–frequency correlation modeling through CNN for frequency domain convolutional smoothing and LSTM for time domain recursive averaging, TFSNN was purpose-built for speech enhancement. Objective tests demonstrated the superiority of TFSNN over other causal speech enhancement systems in terms of enhancing speech quality and intelligibility, suggesting that TFSNN could serve as an effective tool for speech enhancement.

A novel study introduced multi-objective-based multi-channel speech enhancement technique, named bidirectional Long Short-Time Memory (BiLSTM), designed to process signals affected by noise and reverberation [55]. BiLSTM generated two outputs, Log Power Spectra (LPS) and Ideal Ratio Mask (IRM), with the aim of predicting both LPS and IRM features of clean speech. These multi-channel features are then combined to create single-channel features, which are subsequently processed through a deep neural network (DNN) to obtain the speech-enhanced signal. The authors have demonstrated that the proposed method outperforms single-channel and single-objective methods in terms of WER, PESQ, and STOI evaluation parameters. The study further revealed that the value of STOI was observed to be slightly higher at a distance of 0.6 m between speaker and microphone, while it drops for 1.7 m for two RT60 values of 0.45 s and 0.57 s. Similarly, the PESQ value drops from 2.30 to 2.29 when the distance between speaker and microphone values increases from 0.6 m to 1.7 m at RT60 = 0.45 s. It is worth noting that the PESQ value of 2.28 was recorded at both distances of 0.6 m and 1.7 m when measured at RT60 = 0.57 s, which raises a question about its reliability.

Li et al. [56] identified issues with the traditional generalized side-lobe canceller (GSC) in improving automatic speech recognition (ASR) when dealing with speech recorded from a distance. To address this, authors introduced a novel dual-channel deep neural network (DNN)-based GSC structure called nnGSC. This approach allowed nnGSC to automatically track the sound source direction and showed a significant 23.7 % improvement in character error rate (CER) compared to microphone observations. Additionally, it achieved a 13.5 % improvement compared to oracle direction-based super-directive beamformers, a 12.2 % improvement compared to oracle direction-based traditional GSC, and a 5.9 % improvement compared to oracle mask-based minimum variance distortionless response (MVDR) beamformers.

Zhang et al. addressed the issues faced by far-field multi-talker automatic speech recognition systems, which had gained more attention in recent years. The performance of such systems often degraded due to the impact of background noise, reverberation, and other interfering speakers involved in the conversation process simultaneously [57]. An end-to-end multichannel-based system was proposed, utilizing a speech enhancement block as a front end and the ASR block as the back end. The front-end system comprised dereverberation, denoising, and speech separation. Specifically, a WPE dereverberation filter was used to deal with the problem of reverberation, and neural-based beamforming was employed as a denoising module with two purposes: to enhance the target speaker signal and to isolate the target speaker speech signal from the undesired speaker signal. The main idea was to address the training issue faced by existing systems by filling the gap of end-to-end training in more realistic environments such as cocktail party scenarios, where the target speaker, two or more interfering speakers, background noise, and reverberation effects would have a huge impact on the system. For implementing the backend system, an end-to-end ASR architecture was used. The experiment was conducted and evaluated using several multi-speaker benchmark datasets, and the results were compared over single-channel baseline systems. The proposed end-to-end model achieved a 35 % relative word error rate over the baseline used in their study.

Tesch et al. investigated the influence of different types of information, such as spatial, spectral, and temporal, on a DNN-based filter for speech enhancement or speech extraction problems [58]. The authors aimed to understand the nature of a non-linear spatial filter (NSF) and its interdependencies with temporal and spectral information. Various network architectures, including narrowband version of joint non-linear

filter (T-JNF), wideband version of joint non-linear filter (F-JNF), T-NSF, F-NSF, and FT-NSF, were used to analyze these relationships. The FT-JNF architecture flipped the first LSTM layer to the frequency axis to learn cross-band information, which was used for multichannel speech enhancement. The proposed SpatialNet shared a similar purpose with FT-JNF but replaced the LSTM networks with a more powerful Conformer narrow-band block and a convolutional-linear cross-band block. The F-JNF and T-JNF base network architectures jointly processed spatial information with either temporal or spectral information. The FT-JNF architecture was introduced to utilize all three sources of information by manipulating the data arrangement. The non-linear spatial filtering architectures, T-NSF, F-NSF, and FT-NSF were designed to study the properties of a non-linear spatial filter separately from tempo-spectral processing. Lastly, a single-channel post-filter was introduced to jointly process temporal and spectral information. The study trained these networks based on a complex ideal ratio mask to facilitate phase enhancement.

In 2023, Kuang et al. proposed a neural-based beamformer technique for multi-channel speech enhancement [59]. The approach was designed to work in three steps. The first step involved using a set of masks to estimate the filter coefficients of the MVDR beamformer. In the second step, the output of the beamformer, driven by a deep neural network-based post-filtering technique, was used to remove residual noise. Finally, a distortion compensation filter was applied to enhance the speech quality and intelligibility of the signals. The proposed hybrid approach, known as the TriU-net model, demonstrated superior performance on the CHIME-3 dataset, achieving evaluation metric scores of 2.854 and 92.57 % for wb-PESQ score and ESTOI, respectively. However, there was a significant drawback to this method: the model required retraining whenever the array geometry configuration changed. Moreover, the training process involved batch processing, which limited its use in online platforms.

Cherukuru et al. addressed the issue of environmental noise affecting the performance of multi-channel speech activity devices used in real-time applications [60]. To improve speech quality, a new MCSE system that employed deep learning and preprocessing algorithms was proposed. The system used wavelet transform, CNN, and BiLSTM models to analyze data patterns and eliminate noise from the speech signals. The proposed MCSE system surpassed previous studies by achieving a WRR of 70.55 % at −10 dB SNR and 75.44 % at 15 dB SNR, while the existing system scored only 5.82 % and 88.8 %, respectively. The ANOVA analysis confirmed the significant difference in performance between the two systems. Additionally, this research showed that word recognition accuracy remained acceptable even at low SNR levels.

Our exploration of speech enhancement involved an in-depth analysis of cutting-edge advancements in DNN and pioneering research in this field. The findings were summarized into three comprehensive tables, offering a brief overview of speech enhancement based on denoising techniques, acoustic modeling, and beamforming techniques. Table 1 features an overview of speech enhancement through denoising, including details of techniques, datasets, features, usage, performance, and limitations observed in recent studies. A significant breakthrough identified is the shift towards training acoustic models using noisy speech data, resulting in a noteworthy enhancement in system performance. Table 2 contains in-depth information on acoustic modeling-based speech enhancement techniques, showcasing diverse datasets, usage, and features, along with performance discussions and limitations. Additionally, Table 3 provides an exhaustive overview of speech enhancement using beamforming techniques, encompassing a range of approaches, datasets, features, usage, performance insights, and limitations found in different studies. These three tables collectively offer a comprehensive perspective, providing a glimpse into the rapidly evolving landscape of speech enhancement.

### 2.2. Speech recognition

Deep learning is a new attractive area of machine learning over the last decades and has been utilized in a range of different research topics. Deep learning is used to enhance computers capabilities to understand what humans can do, involving speech recognition. Since speech is the main method of human communication, it received a big interest for the past decades right from the introduction of artificial intelligence. So, speech is the early applications of deep learning and up to this day a huge number of research papers have been published in the use of deep learning for different speech applications specifically speech recognition [26,152,154,181,184,185].

Hinton et al. presented a highly effective and comprehensive paper detailing ASR performance advancements. The paper highlighted several key factors contributing to these advancements, including: i) the utilization of deep neural network (DNN), recurrent neural network (RNN), and long short-term memory (LSTM) models, enabling the learning and optimization of large variations for the ASR method; ii) the incorporation of larger vocabularies in training, which enhanced the recognition process; and iii) the introduction of powerful GPUs, increasing the feasibility of training extensive models on big data. Nevertheless, they encounter difficulties in achieving full optimizations and efficient time dominance, especially in unfavorable real-time conditions [28].

In the training phase of DNN models that address noise, researchers have developed a variety of techniques known as a-noise-aware networks. One approach involves using the vector Taylor series (VTS) as a noise estimator to generate inputs that adapt to the noise. However, the proposed system did not result in a significant decrease in the Word Error Rate (WER). Despite these advancements, it remains challenging to synchronize three critical steps in the process. Real-time applications require a robust system to function effectively [62]. In 2013, a DNN was created to adapt all parameters using a small amount of retraining. Nevertheless, a lack of data led to overfitting and a loss of valuable training information. To address this, a senone distribution approach was proposed to maintain proximity to the original distribution. This involved introducing the Kullback-Leibler divergence between adapted and unadapted posteriors into the optimization criterion [63]. Another method involves injecting different noises into the input speech data during training to help the network learn the noise patterns and enhance its ability to perform well in noisy environments, thereby improving the robustness of ASR systems. The system's approach to noise injection is complex, and its performance is lowest when car noise is considered. It performs best in a noisy restaurant environment [64]. In 2014 (Juni Duj et al.), an integrated end-to-end ideal system of jointly modeling front-end and back-end solutions was developed to improve speech robustness in reverberant conditions due to the difficulty in recovering clean speech from corrupted signals using only the front-end system. The system showed an error rate reduction of approximately 50 % from the baseline. DNN-based speech enhancement performed better than clean-condition training, as demonstrated in the Aurora-4 task [65]. In 2015 (Tiano Gaos et al.), a joint training approach was evaluated to tackle ASR in a noisy environment, and the evaluation of ASR performance was conducted on the AURORA 4.0 corpus. The DNN base was evaluated, showing better performance in three different acoustic features [66].

DNN systems have a significant impact on the field of speech signals, particularly in enhancement and recognition. However, the detrimental effects of reverberation and background noise are more pronounced in distant communication environments [153], leading to signal degradation. Therefore, there is ample room for improvement in speech signal processing for distant communication scenarios. Giri et al. [67] presented a system that employs two approaches to enhance the robustness of DNN acoustic models to reverberation. The first approach utilizes a multi-task learning method to train the DNN, leveraging parallel training data where the signal representation is shared to simultaneously learn and perform both senone classification and feature

**Table 1**
An overview of speech enhancement based on denoising process.

| Research Author | Technique | Dataset | Features | Usage | Performance | Limitation |
|---|---|---|---|---|---|---|
| Cherukuru Pavani et al. [60] | DWT is used to remove noise. CNN is used to improve the quality of speech signals. | AURORA, LibriSpeech, and NOIZEUS | Spectral | The proposed MCSE system employs DWT preprocessing and deep learning using CNN. The system filtered environmental noise in low to high SNR conditions, leading to improved accuracy in word recognition rate. This technique has practical applications in developing speech enhancement systems. | The proposed DWT-CNN-MCSE outperforms BAV-MCSE system in WRR and shows superior performance in stationary and non-stationary environments for different SNR levels (−10 dB to 10 dB). It performs slightly better with non-stationary noise. | The proposed DWT-CNN-MCSE system's reliability is in question due to inconsistent results at higher SNR levels, while it performs better at lower SNR levels. It performs inferiorly compared to BAV-MCSE at SNR of 15 dB. |
| Yoshiaki Bando *et al.* [53] | De-noising variational auto-encoder (VAE) based speech enhancement. | CHiME4, LITIS ROUEN Audio scene, TIMIT + ROUEN | – | To enhance the speech combination of generative and discriminative approaches. Unsupervised noise model is combined with VAE that estimates latent vectors of clean speech. Besides, multitask learning for denoising is applied. | The obtained STOI value is 0.89 and for noisy mixture is 0.83. The proposed system performs better than other algorithms. | It is very likely that VAE eliminate useful information, and tends to be sensitive towards input errors leading to poor performance in unknown conditions |
| Yan Zhao et al. [136] | Deep learning-based speech enhancement. | IEEE Corpus, DEMAND | FFT | A noise removal model is used where IRM is applied to the magnitude spectrum. A time domain signal reconstruction (TDR) module with a new objective function is applied for the enhancement process. | The obtained STOI results show that the method outperformed other methods while it does not achieve significant performance on the PESQ parameter. | One system limitation is the susceptibility to highly damaging noise, which can significantly impact its performance in real-world scenarios. |
| Nasir Saleem et al. [118] | DNN and less aggressive Wiener filtering-based speech enhancement algorithm. | Noizeus and Aurora | STFT | DNN is used to calculate the magnitude spectrum of noise signals and a Less aggressive wiener filter is placed as an extra layer to enhance the magnitude spectrum. | Speech quality is improved compared to baselines: spectral subtraction, wiener filtration, log MMSE and performs better in noisy conditions. | Only limited conditions were considered for noise suppression where Weiner filter introduces unwanted musical noise as a major limitation. |
| Babafemi O. Odelowo et al. [137] | Time domain-based speech enhancement. | IEEE Corpus and NOIZEUS database | LPS, Log Magnitude Spectral | To enhance speech by using noise prediction model with time domain subtraction. | The system shown better performance than conventional systems in seen noise but could not give satisfactory results in unseen noise. | Performance is good for high SNR values but deteriorates at low SNR levels. There is no significant improvement in unseen noise due to post-filtering based on masks. |
| Ming Tu et al. [9] | DNN based speech enhancement | TIMIT | Mel-frequency spectra | To learn from non-linear mapping of DNN, it adds skip connection with identity weight matrix between input and output. It learns log compressed ideal ratio mask when input noisy and out clean are compresses. | The performance was estimated for the noise condition of shopping malls and indoor cafes. DNN1 performs better in STOI, and DNN2 outperforms other systems in PESQ parameter. | The improvement over the baseline was not better at low signal-to-noise ratios due to the consistent skip connections at each stage in the DNN2 model. |
| Sivaramakrishna Yecchuri et al. [202] | The TANSCUNet that integrates attention modules such as adaptive time–frequency attention and adaptive hierarchical attention. | Common Voice corpus and Noizeus | Spatial covariance matrices | The TANSCUNet model effectively reduces noise while preserving important speech elements by integrating attention modules such as adaptive time–frequency attention and adaptive hierarchical attention. | The model has shown significant improvements. In known noise conditions, it achieves an average SDR of 10.52 dB, STOI of 88.23 %, and a PESQ score of 3.36. In unknown noise conditions, it delivers an average SDR of 10.12 dB, STOI of 87.14 %, and a PESQ score of 3.24. | The model's exceptional performance is offset by its substantial computational demands and training inefficiencies, making it less suitable for real-time applications. |
| Nasir Saleem et al. [203] | A novel Time-Domain Convolutional Neural Network (TDCNN) was introduced featuring a Transformer bottleneck | WSJ0 SI-84 | local and global features | Primarily designed for speech enhancement, this technique significantly improves speech intelligibility and quality in noisy environments. Its | The TDCNN + MTAT model demonstrates exceptional performance, surpassing existing SE models. It achieves remarkable improvements | Although the model exhibits exceptional performance, its computational complexity and extensive training time due to its deep |

(*continued on next page*)

**Table 1** (*continued*)

| Research Author | Technique | Dataset | Features | Usage | Performance | Limitation |
|---|---|---|---|---|---|---|
| | empowered by Time-Attention (TAT). | | | applications extend to fields such as automatic speech recognition (ASR) systems. | in STOI by up to 24.91 % and delivers significant enhancements in PESQ scores, particularly in multi-talker babble noise and cafeteria noise tests. | architecture pose notable limitations. Future endeavors aim to mitigate these limitations by reducing the number of trainable parameters for real-time applications. |
| Nika Sljubura et al. [204] | Deep Convolutional Recurrent Network (DCRN) | DNS, ESC-50, UrbanSound8k, and Librispeech | Mel-Spectrogram | An innovative system designed to integrate seamlessly into industrial helmets, reducing background noise while preserving critical sounds like alarms and sirens | The cutting-edge system delivers a lightning-fast 150 ms inference time and consumes just 48 mJ of energy per inference on a Cortex-M7 microcontroller. It also boosts classification accuracy by 5 %, achieving an impressive 95 % accuracy in detecting emergency signals. | One drawback is the slightly lower PESQ (speech quality) score. Additionally, the system has a significant memory footprint and needs model optimization to operate on devices with limited resources like microcontrollers. |
| Nasir Saleem et al. [206] | Dual-path High-order Transformer-style Fully Attentional Network (DPHT-ANet) | WSJ0-SI84 and VCTK, DEMAND | Real-Imaginary (RI) spectrogram | Monaural speech enhancement in noisy environments | Achieves ESTOI improvement of 38.28 %, PESQ 1.21, and SDR 10.83 dB on WSJ0-SI84 dataset. | High computational complexity and potential limitations in real-world acoustic environments due to transformer architecture. |
| Manal Abdullah Alohali et al. [207] | Spiking Neural Networks (SNNs) | WSJ0-SI84 and VCTK, DEMAND | Spectrogram | The model is used for speech enhancement, focusing on improving speech quality and intelligibility in noisy environments | The model achieved 33.69 % improvement in ESTOI, 1.05 PESQ, and 11.36 dB SDR over noisy mixtures. | While the model improves temporal feature processing, it may underutilize historical information across different time steps. |
| Jawad Ali et al. [208] | Capsule Networks (CapsNet) with Convolutional Recurrent Neural Networks (CRNN) | LibriSpeech, VoiceBank, DEMAND | spatio-temporal features | The model is applied for monaural speech enhancement | The model improves speech intelligibility by up to 24.2 % in STOI and 33.72 % in PESQ. | The model has a higher computational cost due to the use of capsule networks. |
| Zehua Zhang et al. [210] | Supervised Attention Multi-Scale Temporal Convolutional Network (SA-MSTCN). It uses complex compressed spectrum (CCS) and complex ratio masking (CRM) with a supervised attention mechanism to enhance monaural speech | DNS | CCS and CRM features used | Speech enhancement, particularly for applications like hearing aids, robust speech recognition, video conferencing, and environments with varying signal-to-noise ratios (SNR). | Achieves state-of-the-art speech quality and intelligibility with stable denoising across various noise levels. Performance is robust in both noisy and reverberant environments. | Computational cost is slightly higher than baseline models due to the additional complexity of attention mechanisms. It also shows some limitations in processing very high-SNR clean speech. |

enhancement tasks. The second approach involves a framework that characterizes the reverberant environment extracted from the input signal to train a room-aware DNN. The utilization of multi-task learning yielded superior results in terms of signal denoising and effectively addressed the task of de-reverberation. Furthermore, its application was observed to be highly advantageous in real-time conditions.

Battenberg et al. highlighted the significance of implementing three deep learning methods namely CTC, RNN-Transducer, and attention-based models for the development of an end-to-end automatic speech recognition system [68]. The study compared these three techniques by analyzing the role of their respective encoders in enhancing the performance of the models. Although each of these three techniques was implemented for end-to-end automatic speech recognition, it was worth noting that they differed in their training processes. Specifically, large language models were required for decoding, a factor that was not thoroughly examined in the study.

To address the challenge of multi speaker speech recognition using a transformer model [69], two key issues were identified. Firstly, the transformer model was replaced with RNN-based encoder-decoder models for speech recognition. Secondly, the complexity associated with the transformer model, such as high computational requirements and memory consumption, was addressed by modifying the self-attention component to restrict its scope to a segment instead of the entire sequence, thereby reducing computational complexity. Additionally, an external dereverberation preprocessing technique called Weighted Prediction Error (WPE) was incorporated to improve the model's ability to handle reverberated signals. The experiments conducted on the spatialized wsj1-2mix corpus demonstrated that the Transformer-based models achieved a 40.9 % and 25.6 % relative Word Error Rate (WER) reduction, resulting in 12.1 % and 6.4 % WER in anechoic conditions for single-channel and multi-channel tasks, respectively. In the reverberant scenario, the models achieved a 41.5 % and 13.8 % relative WER reduction, resulting in 16.5 % and 15.2 % WER. Furthermore, the integration of the external dereverberation method significantly reduced the performance gap between reverberant and anechoic conditions, and future improvements are expected through a more integrated approach within the model.

A novel hybrid approach for speech recognition using a transformer-based acoustic model was introduced by Wang et al. [70]. The study explored various modeling techniques, including positional embedding methods and an iterated loss function, to understand the impact of self-attention in replacing recurrence within the Transformer model. Experimental results indicated that the proposed transformer-based acoustic model outperformed the BLSTM model on the LibriSpeech dataset. However, the study did not extensively investigate the specific contribution of self-attention to the Transformer model's superiority, raising questions about the key factors driving its performance advantage.

**Table 2**
An overview of speech enhancement based on acoustic modelling.

| Research Author | Technique | Dataset | Features | Usage | Performance | Limitation |
|---|---|---|---|---|---|---|
| Hoang Ngoc Chau et al. [174] | Deep learning-based end-to-end architectures use GNN, specifically employing temporal graph convolutional network approach. | INTERSPEECH 2021 ConferencingSpeech Challenge dataset, Librispeech, AISHELL-1, AISHELL-3, AudioSet, MUSAN. | – | Used for multi-channel speech enhancement. The technique is validated on the INTERSPEECH 2021 Conferencing Speech Challenge dataset and has practical applications in video conferencing and speech signal enhancement. | The proposed is effective and superior to state-of-the-art methods, as demonstrated through comparisons on the Conferencing Speech 2021 Challenge dataset. | It does not provide sufficient details on the potential joint leverage of spatial, temporal, and spectral representations by GNNs in the latent space. This raises question about the sufficiency of GNNs for multi-channel speech modeling. |
| Anil Garg [182] | LSTM | NOIZEUS | Fractional Delta AMS | The method addressed instability issues during training. To overcome these limitations, it uses modified wiener filter and Long Short-Term Memory (LSTM) networks to enhance intelligibility and quality, while also considering the inherent variability of speech characteristics. | The model showed better performance than existing models, achieving higher SDR, PESQ, CORR, ESTOI, STOI, and SNR values, with lower RMSE values. It is highly effective in reducing airport noise, outperforming other models. | The proposed work has great potential, but it does not fully address all noise sources. It does not provide a comprehensive estimate of the spectrum magnitude and phase, which leads to inaccuracy and reduces its effectiveness. |
| Mhd Modar Halimeh et al. [176] | Complex-valued spatial autoencoder (COSPA) | TIMIT, MUSAN | – | Complex-valued Spatial Autoencoder (COSPA) data-driven approach was employed for multichannel signal enhancement. | COSPA method outperformed four out of five benchmarks, except for the Oracle knowledge Generalized MVDR beamformer (OGMVDR). One advantage of the COSPA method is that it does not require prior knowledge of the direction of arrival (DOA) of the source. | COSPA effectiveness may be limited varying noise conditions because it lacks integration of lateral information such as DOA. And because it is a fully data-driven method, it may not perform as model-based techniques in certain scenarios, unlike knowledge-based Oracle methods. |
| Ashutosh Pandey et al. [177] | Dual-path attentive recurrent network (ARN) augmented with self-attention. | WSJCAM0, REVERB Challenge, DNS challenge corpus | – | Triple-Path Attentive Recurrent Network (TPARN) model is utilized for multichannel speech enhancement. The model processes signals independently through ARN with self-attention and incorporates spatial context aggregation to enhance speech quality and suppress noise. | This technique surpasses existing state-of-the-art approaches by utilizing spatial information to achieve superior results under challenging noisy and reverberant conditions. The TPARN model improves speech quality while retaining the speaker's natural voice. | The study has certain limitations, as it does not provide information on how the technique was used for removing room reverberation using time-domain approaches. Additionally, it does not report results for varying reverberation levels. |
| Chang-Le Liu et al. [175] | Fully convolutional network (FCN) with Sinc and dilated convolutional layers (SDFCN). | CHiME3, WSJ0. | LPS | A fully convolutional network technique was used to process raw data using waveform mapping concepts based on neural network models to enhance multichannel speech signals, improving quality and intelligibility. | The study can remove audio noise and achieve better subjective listening scores, speech recognition performance, and objective metrics such as STOI and PESQ. rSDFCN system outperformed the DDAE system in multichannel settings. | This work has good evaluation metric scores for: SDFCN and rSDFCN, over existing methods. But it lacks specific details about the involved parameters and complexities. |
| Jen-Cheng Hou *et al.* [50] | Audio Visual based speech enhancement | Mandarin dataset – Taiwan MHINT | Mel Filter Bank | CNN is used to extract visual information, and the obtained speaker's lip movements are used to find the corrupted speech audio input. A deep denoising encoder is used to enhance the audio signal. | Without early fusion, the obtained STOI is equal to 0.66 in AVDCNN while with early fusion is 0.51 and PESQ is 2.41 and 1.52 respectively. | Main problem with these features is that they are not interpreted accurately and not suitable for confidential information publications. Also more data is needed to improve performance. |
| Soumitro Chakrabarty *et al.* [96] | T-F mask- based speech enhancement. | TIMIT, Librispeech | STFT | CNN along with T-F mask is used to differentiate between spatial and desired directional characteristic and remove the uncorrelated noise from it for enhancement. | Proposed system effectively outperforms baseline in both room setups, it eliminates additive noise very efficiently. | Performance gap is noticed compared to oracle IRM due to choice of features with masking for regression. |
| Samba Raju Chiluveru et al. [157] | DNN and Phase estimation-based speech enhancement. | TIMIT, Aurora2 | Not mentioned | Regression model with DNN mapping is used for speech enhancement in low SNR conditions. | Clear distinctive reconstruction is generated from the noisy signal. | Complexity of the system is high, cepstral analysis is computationally expensive. |
| Yan-Hui Tu *et al.* [93] | Improved minima controlled recursive averaging (IMCRA) and deep learning-based speech enhancement. | CHIME4,WSJ0 | LPS | DNN regression model is used to estimate phase and amplitude information of clean signals for speech reconstruction. | The student model with DNN in causal mode is compared to the bidirectional gated recurrent units (BGRUs) that gives consistent improvement with WER reduction of 7.94 %. | Proposed system is based on single channel which faces limitation of limited received information. |
| Nursadul Mamun et al. [211] | Deep Complex Convolution Transformer Network (DCCTN) | TIMIT, AURORA | Real and Imaginary components | Speech enhancement, specifically aimed at improving speech intelligibility for cochlear implant (CI) recipients in noisy environments. | The model achieves significant improvements in speech intelligibility (up to + 40 %) and quality (+31 %) compared to baseline models like DCCRN, especially in noisy conditions such as babble and car noise. | While effective, the model requires high computational resources due to the complex-valued processing and transformer layers, which might limit real-time applications. |

**Table 3**
An overview of speech enhancement based on beamforming.

| Research Author | Technique | Dataset | Features | Usage | Performance | Limitation |
|---|---|---|---|---|---|---|
| Changsheng Quan et al. [173] | The proposed SpatialNet utilizes a novel convolutional linear cross band block based on CNN, which efficiently learns complex spatial information. | SMS-WSJ, WHAMR, WSJ0-2mix, LibriCSS, Reverb Challenge, and CHiME 3/4 Challege. | Spatial features. | The study presents the development of SpatialNet, which effectively leverages spatial information in multichannel speech enhancement tasks. The proposed model demonstrates state-of-the-art performance and holds potential for real-world applications requiring speech separation, denoising, and dereverberation. | The results of the study highlight the effectiveness of SpatialNet in achieving state-of-the-art performance on various speech enhancement tasks, effectively leveraging spatial information. | The limitation of this study is that the proposed SpatialNet is designed for offline (non-causal) processing. |
| Hansol Kim et al. [172] | DNN approach was used for estimating the MVDR beamformer coefficients for enhancing signals in speech and noise spatial covariance matrices estimation. | TIMIT, NOISEX-92, and Aurora 2. | – | This research used DNN to estimate MVDR beamformer factors, improving performance for multi-channel speech enhancement. The techniques and findings can be applied to enhance speech quality scores using deep learning-based beamforming in multi-channel speech enhancement. | The proposed factorized MVDR beamformer effectively mimics the characteristics of the MVDR beamformer with true relative transfer function and noise SCM, outperforming the traditional MVDR beamformer with deep learning-based pre-and post-processing in terms of the PESQ scores. | The study has limitations, including potential signal preprocessing disrupting phase relationships, using time averages that may not be optimal for beamformer performance, and the potential distortion of phase relationships by complex spectral mapping using DNNs, leading to reduced performance. |
| Wenzhe Liu et al. [187] | Neural Beamspace-Domain Filter implemented using CNN-based encoding–decoding architecture. | DNS-Challenge dataset, TIMIT, NOISEX92, CHIME-3. | Spectro-temporal-spatial discriminative features. | This article presents a new approach to designing beamforming coefficients using a causal neural filter that utilizes spectro-temporal-spatial information in the beamspace domain. The method involves designing multiple beams, learning a deep-learning-based filter, and incorporating a residual estimation module to suppress interference components effectively. | A new neural beamspace-domain filter has been developed for multi-channel speech enhancement, outperforming many existing methods by effectively addressing directional interference and diffused babble noise. It offers better speech quality and intelligibility compared to previous approaches, setting a new benchmark for the field. | Lacks information on the computational cost. |
| Kaizhi Qian *et al.* [14] | Multichannel DNN based speech enhancement. | TIMIT,Free SFX, VCTK | Not mentioned | To implement ad hoc microphone network called Deep-Beam used for forming natural sounding speech and filter coefficient which are produced by using monaural speech enhancement. | Deep-Beam outperforms conventional methods like MVDR, GRAB, CLOSEST, IVA on both simulation and real time conditions. | The model's performance is hindered by inadequate compression and the complexity of determining the suitable depth for the application based on depth and intensity. |
| Xueliang Zhang et al. [76] | Single-Multi microphone-based speech enhancement | CHIME 3 | AMS, MFCC, RASTA-PLP | To map spectral features using T-F mask, DNN is trained from single microphone, while multi-microphone estimate is used to calculate noise covariance matrix and steering vector with MVDR beam-former. | Proposed system gives 5.05 % of WER value in real time conditions while system without iteration provides 5.42 % WER that outperforms other baseline models. | Due to re-estimation of mask, steering vector value may not be calculated accurately. |
| Tsubasa Ochiai *et al.* [94] | Joint training based BLSTM | CHIME4, AMI | LMFB | Neural beam-former with filter estimation and mask estimation approaches are used for speech enhancement. | This model dominates other system with character error rate of 18.2 % in real development set and 26.8 % in real evaluated set. | The model's performance falls short of the state-of-the-art due to ineffective features and the lack of robustness and efficiency in the neural beam-former. |
| Thanh T. Vu et al. [47] | Sparse NMF-based speech enhancement | CHIME 3 | MFCC, fMLLR | The system utilizes Signal Dependent MVDR beamforming as a component in the front-end module to improve speech quality in noisy multichannel recordings by eliminating signals below a specific threshold value. | A significant 65 % relative improvement has been observed from the baseline in WER% rates, with the pedestrian environment yielding the best results. | The combination of multiple techniques increases the risk of unforeseen interactions and limitations. |

Lee et al. demonstrated in their research that the Transformer model was more vulnerable to input sparsity than the CNN [71]. The study investigated the performance drop and identified the Transformer's structural properties as the primary cause. The authors proposed a distinct regularization technique that enhanced the Transformer's resilience to input sparsity. This method regulated attention weights without requiring additional module training or excessive computation by leveraging silence label information in forced alignment. This approach offered a practical solution to improve the Transformer model's performance under sparse input conditions.

Li et al. addressed an issue identified in Transformer-based end-to-end models for Automatic Speech Recognition (ASR) tasks [72]. The authors observed a lack of diversity in the intermediate features from different input streams in these models, especially concerning speaker characteristics. To tackle this issue, the author proposed a multi-level acoustic feature extraction framework that combined shallow and deep streams. This framework was designed to capture both traditional features for classification and speaker-invariant deep features to enhance feature diversity. Additionally, the authors employed a feature correlation-based fusion (FCF) strategy to combine intermediate features across frequency and time domains. This approach involved correlating and combining the features before feeding them into the Transformer encoder–decoder module, ultimately resulting in improved performance of the model in ASR tasks.

An end-to-end multichannel speech recognition system was developed by Zhang et al. [73]. This system was developed and implemented with a focus on integrating a deverberation module into the frontend [73]. The system used a multi-source weighted prediction error (WPE) to suppress reverberant components using a mask-based concept. The main goal of the study was to create an algorithm capable of recognizing distant speech. The work proposed extending the weighted power minimization distortionless response (WPD) convolutional beamformer to handle separation and dereverberation tasks simultaneously, enhancing the model's stability during back-propagation. The proposed system was optimized based on the ASR criterion, showing good dereverberation and separation skills. Two innovative frontend architectures were explored and evaluated, demonstrating promising performance on the spatialized wsj1-2mix corpus compared to the previous MIMO-Speech model. Experimental results on the REVERB dataset confirmed the effectiveness of the proposed WPD-based architecture. However, the study's limitation is its restricted evaluation of beamformer variants ($K \in \{1, 3, 5, 7, 10\}$) and fixed channel count ($C = 6$) with limited mask types, possibly preventing optimal settings for the proposed techniques.

Nagano et al. addressed the challenges faced by robots in real environments by proposing the U-TasNet-Beam method [74]. The aim was to improve speech extraction performance and speech recognition accuracy in difficult environments characterized by reverberation, noise, and multiple speakers. The proposed method outperformed conventional methods in sound source localization and speech recognition accuracy, highlighting its potential for robot applications. The paper provided a detailed explanation of the neural beamformer method, emphasizing the use of the SCM loss function to enhance spatial information learning, which ultimately improved speech extraction performance. The method's advantages included the simultaneous handling of multiple challenges such as reverberation, noise, and multi-speaker speech, leading to improved speech recognition accuracy and sound source localization. The experimental results conducted using an actual machine demonstrated a 27.8 % improvement in speech recognition rate compared to unprocessed signals. However, the performance of the proposed method decreased with high volumes of interfering speakers, and it proved to be sensitive to high reverberation levels, indicating a need for further improvement in such scenarios.

Dimah Al-Fraihat et al. conducted a comprehensive analysis of the advancements in speech recognition spanning from 2019 to 2022, concentrating on the increasing significance of hybrid deep learning models. The research effectively captures the latest trends and the transition towards more intricate architectures. However, it primarily centers on widely utilized datasets and languages, predominantly English, neglecting the importance of addressing low-resource languages. This omission is crucial for expanding the applicability of these models. Furthermore, although the paper acknowledges the potential of transformer models, it provides limited comparative analysis, missing an opportunity to delve deeper into their advantages. A commendable aspect of the paper is its forward-looking perspective on underutilized techniques, such as alternative feature extraction methods and the growing utilization of transformers, offering a clear direction for future research. Despite its strengths, the paper could augment its impact by acknowledging the limitations of current evaluation metrics, such as Word Error Rate (WER), and advocating for more diverse approaches. In conclusion, while this review effectively summarizes recent developments, future studies should focus on addressing these critical gaps to comprehensively tackle the challenges in modern speech recognition [209].

Yasin Gormez introduced a hybrid deep learning model for Turkish Automatic Speech Recognition (ASR), effectively addressing the challenges posed by the Turkish language's structure. The study's use of multiple deep learning architectures and the integration of the Zemberek-based language model significantly improved word error rate (WER) and character error rate (CER), leading to improved recognition accuracy. However, notable concerns include the model's computational complexity and limitations in handling certain words due to the nature of the Turkish language [205].

Our investigation into the field of speech recognition involved conducting a thorough analysis of the latest developments in ASR systems using deep neural networks. Our research was summarized in three detailed tables, providing a concise overview of speech recognition techniques, including denoising, acoustic modeling, and beamforming. Table 4 offered an overview of speech recognition using denoising techniques, highlighting the emergence of new training methods using diverse speech datasets, features, usage, performance, and limitations. In Table 5, a detailed breakdown of acoustic modeling-based speech recognition techniques was presented, showcasing a wide array of datasets, features, usage, and performance along with limitations. Additionally, Table 6 comprehensively covered various beamforming-based speech recognition techniques, datasets, features, usage, and provided insights into their performance and limitations across different studies. Collectively, these three tables provided a comprehensive perspective, shedding light on the rapidly evolving landscape of speech recognition research.

The systematic review presented here stands out from other reviews by offering a comprehensive analysis of studies on speech enhancement and speech recognition using deep neural networks. This study offers a captivating exploration of speech enhancement with various techniques [125,128,129] and seamlessly transitions into the realm of speech recognition integrated with deep neural networks [126,127,131,132], providing valuable insights into the core categories of speech enhancement and recognition including far-field scenarios, serving as an enlightening guide for readers from various backgrounds. By shedding light on various datasets, features, types of deep neural networks, and fields of usage, performance, limitation, this study equips readers with a comprehensive theoretical understanding to address any uncertainties related to speech synthesis, speech enhancement, and speech recognition.

Furthermore, the limitations of several deep learning techniques, such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, Auto-Encoders, Transformer models, and the Convolutional Gated Recurrent Unit Deep Neural Network (CGDNN), have been comprehensively outlined in Table 1 through Table 6. These approaches have been extensively utilized in tasks related to speech enhancement and recognition, with considerable investigation into their

**Table 4**
An overview of speech recognition based on denoising.

| Research Author | Technique | Dataset | Features | Usage | Performance | Limitation |
|---|---|---|---|---|---|---|
| Wolfgang Mack *et al.* [138] | MMSE Optimization technique | Libri Speech and TIMIT | STFT | DNN estimates ratio mask of reverberant magnitude spectrum to desired magnitude spectrum which is used for training DNN to obtain minimum-mean-squared-error MMSE. | The system evaluated with same parameters in different rooms yielded good results but when compared with oracle result, there was a gap noticed in the performance. | There is a significant gap in all metrics between the results obtained by the proposed algorithm and the oracle mask, indicating that the proposed algorithm did not achieve optimal performance. |
| Mirco Ravanelli *et al.* [161] | DNN based distant speech recognition | TIMIT and WSJ | MFCC | To overcome lack of matching and communication in the current system this system is designed as a pipelined DNN for enhancement and recognition. | DNN network obtained PER is 28.7 % on TIMIT and 7.6 % PER on WSJ dataset while WSJ with noise is 12.3 % PER. | At various level of architecture, it is found that the performance is not significant under noisy conditions. |
| Takuya Yoshioka et al. [134] | CNN-DNN-HMM based far field recognition. | REVERB | Log mel filter bank | Integrating advanced methods like dereverberation, feature extraction, and normalization, along with a unique CNN-DNN approach to produce HMM state posteriors, this innovative system enables robust speech recognition by modeling short-term patterns and adapting well to new testing conditions. | The performance of the Trigram and RNN models was compared, and a notable observation was made that CNN-DNN-HMM systems outperformed on simulated data compared to real-time data in all four testing scenarios. | The system's high complexity makes it difficult to synchronize mismatched data, leading to operational challenges. Furthermore, it fails to show improvement in speech recognition performance in seen test conditions. |
| Ritwik Giri *et al.* [67] | DNN based multi-task learning | REVERB, WSJCAM0 corpus | Log mel filter bank | The proposed method aims to classify Senones and extract secondary features to enhance speech recognition by minimizing the error between clean and noise-contaminated speech features to develop a robust system. | The proposed system shows a marginal improvement in performance compared to other systems, achieving a word error rate (WER) of 7.77 % on simulated data. | The proposed method did not perform significantly better on real-world data compared to spectral subtraction, dereverberation, and GMM-SGMM-DNN ROVER method. |
| Mengyuan Zhao *et al.* [141] | Convolutional DAE (CDAE) based speech recognition | Aurora 4 and Chinese dataset | Fbank Features | To extract the repeating patterns of music in the spectral and temporal domains, bringing significant improvement in ASR due to the convolution layer. | The WER value achieved on the multilingual CDAE Aurora4 and 863 datasets with embedded piano are 8.76 % and 23.33 % respectively. | One potential limitation/ challenge is the varying acoustic conditions between datasets, which may contribute to a performance gap. |

performance capabilities. Although the constraints of these models are well established, it is crucial to also investigate their future potential in these specific applications. This includes exploring how each of these methodologies can advance, especially in terms of improvements that could overcome their existing challenges and enable broader applications in complex, real-world speech processing scenarios.

DNNs have shown considerable potential in speech enhancement and recognition; however, they struggle with slow convergence rates during training, particularly in real-time speech enhancement systems using large, noisy datasets. Advancements in optimization algorithms, regularization techniques like batch normalization, and adaptive learning rates could help mitigate this issue. Additionally, incorporating transfer learning and utilizing unsupervised pre-training could enhance the adaptability of DNNs to new speech data with limited labels.

CNNs are effective at extracting spatial features from speech spectrograms in enhancement and recognition tasks, yet they face challenges such as high computational costs and limited resilience to variations like background noise and reverberation. The future of CNNs may involve integrating energy-efficient architectures and employing neural architecture search (NAS) to minimize computational demands while maintaining performance. Further, combining CNNs with self-supervised learning and generative models could fortify their robustness in noisy environments.

RNNs and LSTMs are advantageous for sequential speech recognition due to their capacity to model temporal dependencies. Nonetheless, they encounter problems like vanishing gradients and prolonged training times that impede real-time application performance. Implementing gradient clipping and sophisticated training methods, such as teacher forcing, may alleviate gradient concerns. Furthermore, integrating attention mechanisms or Transformer-based models could enhance their capabilities for managing long-range dependencies in both recognition and enhancement tasks [156].

Auto-Encoders are useful for denoising and unsupervised learning in speech enhancement [142,155,166,167]; however, their generalization can be compromised in diverse noise environments. Future enhancements could arise from combining variational autoencoders (VAEs) and generative adversarial networks (GANs), which may improve feature extraction and create cleaner representations of noisy inputs, significantly boosting performance in challenging conditions.

Transformers have excelled in speech recognition tasks but face challenges such as high computational complexity and inefficiencies in processing lengthy sequences. Their future potential lies in techniques like model pruning, sparse attention mechanisms, and knowledge distillation to enhance computational efficiency. Additionally, expanding Transformers to handle multimodal data, including both audio and text, could enrich their performance in tasks necessitating the comprehension of linguistic and acoustic features. Vision Transformers (ViTs) may also be investigated for their applicability to speech enhancement, especially in scenarios with multi-channel or spatially diverse microphone setups.

Although the CGDNN displays promising results through its hybrid architecture in speech enhancement and recognition, it encounters challenges related to slower convergence rates and computational demands, especially in real-time applications. These obstacles could be addressed by adopting dynamic model pruning, quantization strategies, and investigating multi-scale architectures to reduce complexity while sustaining performance. The integration of reinforcement learning (RL) for adaptive optimization could further enhance its scalability and efficacy across varying real-time speech processing contexts, ensuring a better balance between accuracy and computational load.

Despite the inherent limitations of the discussed models, their future in speech enhancement and recognition rests on the development of more efficient, adaptable, and scalable architectures. By tackling issues such as slow convergence, high computational costs, and lack of

**Table 5**
An overview of speech recognition based on acoustic modelling.

| Research Author | Technique | Dataset | Features | Usage | Performance | Limitation |
|---|---|---|---|---|---|---|
| Zhenqing Li et al. [180] | LSTM | TIMIT, LibriSpeech, and VoiceBank | Spectral features | A hourglass-shaped LSTM architecture was introduced to address the long-term dependencies in LSTM models for ASR applications. It is used to improve quality and intelligibility. An IRM mask was employed to eliminate additive noise, ultimately improving speech properties for the target speaker. | This model achieved a WER of 19.13 %, outperforming baseline methods in trainable parameters with only 18.89 M compared to 46.18 M parameters. It also improved STOI and PESQ values compared to baseline methods. | The implementation of deep learning algorithms on portable communication devices is often difficult due to high computational costs, in addition to the obstacle of the limited computational power. |
| Jane Oruh et al. [169] | LSTM-RNN | Publicly available dataset from the librosa library, consisting of isolated spoken English digits. | MFCC | Automatic Speech Recognition (ASR) using an enhanced deep learning LSTM recurrent neural network (RNN) model. | The proposed LSTM-RNN model, achieves 99.36 % accuracy on the spoken English digit dataset. It effectively processes continuous input streams, surpassing the limitations of traditional LSTM models. | A potential limitation of the modified LSTM architecture is the increase in computational resources required. |
| Renjian Feng et al. [170] | Low-rank factorization and batch normalization. | ST-CMDS, AIDATATANG, AISHELL-1, AISHELL-2, and LibriSpeech. | – | This study' findings have practical applications in speech recognition, specifically in enhancing acoustic modeling accuracy and efficiency using the Norm-PmGRU architecture. | The Norm-PmGRU model outperforms other baseline models in various ASR tasks by achieving lower CER and WER, while also using fewer parameters than mGRUIP-Ctx, TDNN-OPGRU, TDNN-LSTMP, and other RNN baseline acoustic models. | Despite its impressive performance, the Norm-PmGRU model has some limitations. It may face scalability issues as context length increases. Careful optimization of context information is also necessary. Additionally, trade-off between different model enhancements is required to achieve optimal performance. |
| Zhong-Qiu Wang et al. [179] | Complex spectral mapping approach using DNN. | CHiME-4 | RI components are concatenated and used as input features | Two (DNNs) are used to estimate complex-spectra mapping, and compute the MVDR beamformer output, which contains crucial spatial information. This spatial information, in combination with the mixture, trains the DNN to enhance signal quality and improve ASR performance. | The method outperformed the previous best results in the WER score without using a model ensemble. Performance was tested using one, two, and six-channel microphones, showing WER scores of 6.82 %, 3.19 %, and 1.99 %, respectively. The results outperformed TI-MVDR when two microphones were used. | Limitations include the lack of evaluation in real-time scenarios, the absence of discussion on performance in dynamic noise and reverberation environments, and the unaddressed computational complexity of the proposed algorithm. |
| Ilyes Rebai *et al.* [133] | DNN based speech recognition | French Database | fMLLR | Data Augmentation and Ensemble Method are applied for integration of posterior probabilities generated by DNNs together to improve recognition accuracy. | Proposed system performs better with simulation data but it does not give expected results in real time. | The matching of the three acoustic models simultaneously is not practically robust. Complex due to ensemble learning along with data augmentation. |
| Bo Wu *et al.* [4] | Two techniques are applied DNN-based regression to enhance reverberant and noisy speech, DNN-based multi-condition training. | REVERB, WSJCAM0, TIMIT | LPS | A Reverberation Time Aware (RTA) based architecture and simultaneous exploitation of the data acquired with a multi-channel microphone are applied to outperform techniques like beamforming. | CD-DNN-HMM outperforms the other system based on GMM and successfully attains top Standardized Root Mean Square Residual (SRMR) scores. | In real time data the 8-channel system under performed than single channel with introduction of new data. |
| Rezki Trianto et al. [105] | LSTM based speech recognition | CHIME 4 | MFCC | In the multi-channel speech system GEV beam-former is used to enhance the utterances and, in the back-end LSTM is used with TDNN for better sequential features of audio stream for recognition. | The average WER% on Fast LSTM is 8.12 % and with Fast bi-directional gave 5.19 % in real time. | Time consumption by bidirectional LSTM is high with more computational power that is ineffective under real time conditions. |

(*continued on next page*)

**Table 5** (*continued*)

| Research Author | Technique | Dataset | Features | Usage | Performance | Limitation |
|---|---|---|---|---|---|---|
| Wei Wang *et al.* [81] | Total variability algorithm with deep neural network. | NIST LRE 2007 | MFCC | To perform language recognition using FCN model and extracted I-vector features. | Accuracy is evaluated based on layers number, activation function, number of node and dropout ratio and obtained significant EER values are 3.54 % while baseline has 5.36 % as resultant output. | Fast recognition and more robustness in the real time environment is yet a problem faced with I-vector features. |
| Hinda DRIDI et al. [117] | Convolutional Gated Recurrent Unit, Deep Neural Network (CGDNN). | TIMIT | fMLLR, MFCC, Fbank Features, LDC, STC | It is an integrated speech recognition model that has CNN LSTM and Feed forward networks applied on TIMIT database for phenome recognition. | The obtained PER% on testing data with DNN is 21.18 %, with CNN is 18.83 % and with DLSTM is 18.97 %. | There are two limitations: slower convergence rates due to deep architectures, and difficulties in reducing dimensionality. |
| Yasin Gormez [205] | CNN + GRU + LSTM + Transformer. | Turkish Microphone Speech Corpus (METU-1.0) and Turkish Speech Corpus (TSC). | STFT | Turkish automatic speech recognition, applicable in fields such as voice assistants, speech-to-text services, and smart devices | Achieved WER of 9.85 and CER of 5.35 for the METU-1.0 dataset, and WER of 8.4 and CER of 2.7 for the TSC dataset with the best-performing hybrid model. | High computational demands for real-time processing; language model unable to correct certain valid but incorrect word predictions due to Turkish agglutinative structure. |

**Table 6**
An overview of speech recognition based on beamforming.

| Research Author | Technique | Dataset | Features | Usage | Performance | Limitation |
|---|---|---|---|---|---|---|
| Xiong Xiao *et al.* [23]. | Deep beamforming | AMI meeting transcription | MFCC | It uses neural network for predicting the parameters with complex values in forming frequency domain beam-former. | The comparison between GMM, DNN and LSTM acoustic models are done on basis of WER% and predicting beamforming works well on the unseen noises for far field ASR. | The limitation lies in its computationally intensive training process, involving separate and joint training of beamforming and acoustic model networks. |
| Bo Li *et al.* [10] | Multi-channel speech enhancement | Artificially created Own database | LMFB | The Neural Network Adaptive Beamforming (NAB) technique consists of filter prediction, a filter and sum beamformer with multitask learning (MTL), and DNN LSTM layers. It aims to reduce computational complexity and improve speech recognition. | The WER% obtained on cross entropy unfactored system is 21.7 % and 20.4 % on factored system while NAB gives 20.5 % value. | The experiment's-controlled conditions may not accurately reflect real-world scenarios, affecting the generalizability of the results. |
| Zhong Meng *et al.* [17] | Deep LSTM based multi-channel speech recognition | CHIME 3 | LMFB | The units of top layer of LSTM acoustic model are used as the auxiliary input of beam-former to predict current filter coefficients. | The baseline in real time gives 32.88 % of WER and Beamformer with feedback LSTM model gives 24.91 which is better than all other previous systems. | Due to parallel running Acoustic LSTM model, it is difficult to match these systems for synchronized output. |
| Xiong Xiao *et al.* [22] | LSTM MVDR based speech recognition | CHIME 4 | fMLLR | It improves LSTM mask predictor by minimizing the ASR cost function and manually tuning is avoided during IBM for better and reliable results. | After pooling of the masks obtained result on evaluation set mean in real condition is 9.8 while median after three iterations is 6.3. | Beam-formed spectrum still contains ample amount of noise. |
| Takaaki Hori *et al.* [1] | Multi-channel-based speech recognition | CHIME 3 | MFCC | Long short-term memory (LSTM) RNNs along with WDAS, MVDR and GEV beam-former in enhancement phase together form recognition stage. | The combination of multiple hypotheses of systems implied in the research gives better WER reduction from state of art by 5.05 % in real data, while 44.5 % reduction from the official challenge is reported. | Using multiple audio models in the SE stage makes it difficult to integrate the results in the application stage. |

robustness through hybrid models, advanced training methods, and energy-efficient designs, we can alleviate these challenges. Such innovations promise to significantly enhance the practical application of deep learning models in intricate and dynamic speech processing environments, paving the way for broader usage in the field.

Our systematic literature review meticulously examined a total of 302 papers, ultimately extracting 187 published papers to capture the essence of the field. All the information or data extracted from the 187 dedicated papers are as follows:

1. Types of speech studied.
2. Types of datasets used for training and testing of speech information.
3. Types of features extracted from speech data, noisy data or training sets.
4. Types of publications (journals, research papers and conferences).
5. Names of conferences, journals and research articles published in papers.
6. Systematic research conducted with specified year.
7. Types of deep neural network used.
8. Types of hybrid models defined over the years.

The information provided above offers fascinating insights into the

advancements in speech enhancement and recognition through deep neural networks over the past few decades. This comprehensive analysis serves as a valuable tool for identifying areas that require further development to keep pace with the evolving system designs and methods. By shedding light on these areas, it paves the way for researchers to explore new topics in speech enhancement, speech recognition, and devise solutions for the existing research limitations.

## 3. Speech and Machine learning

This section provides an overview of speech signals, machine learning techniques, and deep learning approaches, with a focus on various deep neural network models and their applications in speech processing.

### 3.1. Speech signals

Speech signals are the form of acoustic energy used for communication in human society. Speech signals carry informative data that are encoded in some languages and decided by other human beings with the same language understanding. The actual information was interpreted in the form of language data. This information is as follows:

1. Speech recognition, which helps to understand the content of speech.
2. Speaker recognition helps to understand whose speech is delivered and carries useful information about speaker identity.
3. Emotion recognition, which helps to understand the emotional state of speaker.
4. Accent recognition, which helps to understand the accent information of speaker.
5. Language recognition, which helps to understand language of the speaker.
6. Health recognition, which helps to understand the health of the speaker.
7. Age recognition, which helps to identify the age of speaker.
8. Gender recognition, which his helps to identify the gender of speaker.

Automatic speech recognition (ASR) is a process that recognizes spoken words and converts them into text. ASR refers to the capability of a machine or a system to identify words and phrases in a specific language and convert them into a machine-understandable format [75]. Speech recognition is crucial for extracting information from any speech signal, especially in noisy or reverberant environments [143,144]. It has various applications, including speech-to-text converter applications, assistance for physically challenged individuals, and smart home applications [76,77].

Automatic speaker recognition involves identifying a specific speaker within challenging environments, such as those with noise, reverberation, or other acoustic distortions [140]. This process relies on the speaker's unique characteristics, such as voice tone, pitch, rhythm, and the acoustic conditions during speech recording [78]. A significant area of research in speaker recognition is determining the speaker's location. Identifying the speaker's location contributes to enhancing the voices of specific speakers in the speech enhancement process, which is valuable for applications related to dedicated voice processing [67].

On the other hand, emotion recognition is a type of speech processing that aims to recognize the emotions expressed in spoken language. It is an important piece of information for artificial intelligence (AI) development, which interacts with human society [79]. Emotion is the kind of information that is used to recognize the mental condition of humans to understand the nature of the same. The emotion information consists of various dedicated words and the pitch of the voice. This needs to be implemented with a DNN for better recognition and extraction of emotion information from speech data [80].

Automatic language recognition from speech data is used to recognize vocabulary, words, sentences or phrases. Recognition of language from speech information is achieved by training the system with dedicated languages to be recognized for speech syntheses. Language recognition is used for applications related to automatic translator systems and the conversion of one language of speech to another language of speech data [81].

Health recognition, age recognition, and gender recognition are speech information that is compared with the pitch of the speaker to segregate the dedicated information. It is considered that as the age of humans develops, vocal tracks also develop. The vocal track is used to generate voice or speech signals from humans to communicate. Pitch information is considered for the discrimination of speakers according to their age and gender that helps to obtain health information of the person. Therefore, pitch recognition or segregation is also an important process in the development of artificial intelligence (AI) for human society development.

In general speech signal is very important and plays a crucial role in numerous signal processing applications and recently, speech signal has bought gigantic evolution based on machine learning model. It has a close relationship with human–machine interaction, natural language processing, computer linguistics, and psycholinguistics based on the different applied field as shown in Fig. 1 [197].

### 3.2. Machine learning

Over the years, as automation and digitization have advanced, businesses have been increasingly focused on gathering data. The proliferation of open sources has led to an abundance of data being available for efficient knowledge extraction and real-time improvement. Thanks to high-speed computers and advanced software, performing calculations and computations has become incredibly easy and efficient. The significance of smartly extracting information from data inputs and adapting to changes through learning has been increasingly recognized in recent years. The general term "artificial intelligence" refers to any machine and/or software that provides a level of intelligence. A subfield of artificial intelligence is machine learning which deals with specific algorithms. These algorithms are able to learn patterns from data without any human assistance. Later, with the availability of large amounts of data and computing resources, the development of machine learning increased. One important subfield of machine learning is deep learning (DL) which uses artificial neural networks that mimic human learning and are essentially structured algorithmic structures [198]. Machine learning encompasses various methods. The major categorization of ML with their main steps is depicted in Fig. 2 (a) [198]. Where, Machine Learning Techniques have general steps that applied to reach the desires points and transforms raw data into valuable insights. The main steps that show how does ML work shown in Fig. 2 (b).

The field of machine learning (ML) is dedicated to enabling computers to learn from input data, making it an iterative process that involves identifying and analyzing insights from patterns. These insights are then applied to new input data [82], allowing the acquired knowledge to be used for different examples and direct experiences. Fig. 2 illustrates some important stages of ML. In supervised learning, data is fed with expected output as labels and trained accordingly, as shown in Fig. 3 (a). For example, in the context of a cosmetic product with a specific set of ingredients affecting its price, this process enables the algorithm to train itself and develop a prediction model for creating a new cosmetic product with a different price. While unsupervised learning involves the clustering of input data based on highly correlated patterns determined by statistical properties, as depicted in Fig. 3 (b). Notably, unsupervised learning does not include input labels [82]. Deep learning, on the other hand, focuses on studying neurons to mimic the brain structure, enabling the extraction of complex analogies and derivatives from variable input data [83].

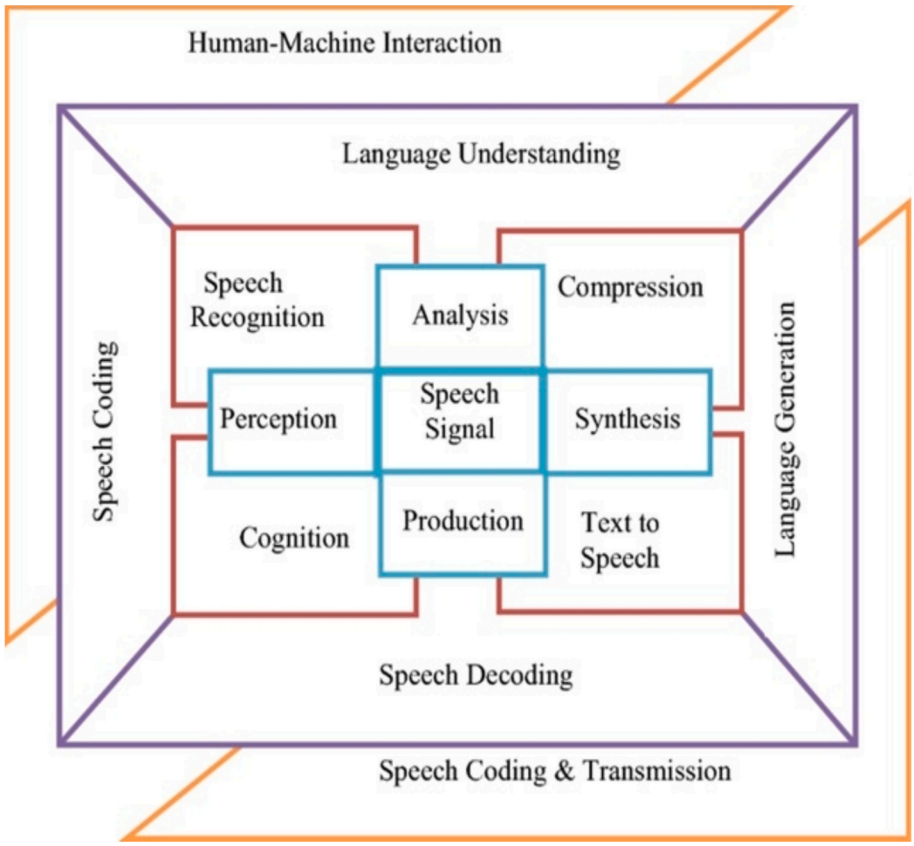



**Fig. 1.** A detailed classification of speech processing fields ( adopted from [197]).

*3.2.1. Supervised learning*

This form of learning operates in pairs, utilizing input represented by sets of vectors corresponding to their expected outputs [84]. The algorithm is supervised, dynamically learning by comparing the predicted output with the actual output [82]. This data analysis produces a classifier that accurately categorizes the data. Supervised learning encompasses regression and classification, providing essential classifications. However, the limited availability of labeled input sets poses a challenge in this process.

*3.2.2. Unsupervised learning*

In unsupervised learning, the algorithm operates without pre-trained output or awareness of the correct output. There is no human intervention in the training, and it does not require labeled vector input [85]. Unsupervised learning is characterized by its subjective nature compared to supervised learning algorithms. It focuses on the fundamental patterns within the data and derives valuable insights from the input structure, identifying and clustering data based on these patterns. Even if the algorithm doesn't fully understand the entire dataset, it discerns differences between patterns and uses them to classify data into specific clusters. This approach is especially effective when dealing with limited amounts of data. An analogy of unsupervised learning is Spotify, where people with similar music tastes are grouped into categories and recommended music accordingly. Unsupervised learning encompasses three main categories: clustering, dimensionality reduction, and anomaly detection [82]. Dimensionality reduction is employed on large datasets containing unnecessary data that hinders performance, modifying input vector dimensions to extract insightful features. Anomaly detection involves analyzing data to identify deviations from the normal behavior of input data points, focusing on the intricate properties of data that lead to unusual behavior in processes.

*3.2.3. Deep learning*

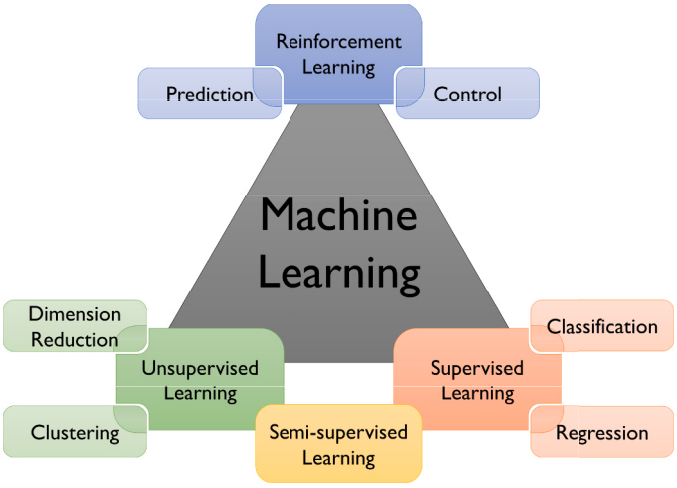
Deep learning has been at the forefront of technological advancements, gaining widespread popularity among researchers worldwide since 2006. Inspired by the human brain, deep learning algorithms are structured in layers, passing high-level to low-level features, forming what is known as deep architecture. These algorithms use weights and biases to calculate nodes and incorporate concepts such as intersection of different neural networks, optimization, pattern recognition, and signal processing. The impact of deep learning has been profound, enabling the training of large datasets, increasing processing capacity, and fundamentally transforming the field of machine learning. It has truly revolutionized the way we approach data analysis and pattern recognition [86].

Recently, DL was adapted in image processing fields, then it was used in almost all signal processing fields such as music, speech, and environmental signal processing [1]. Research based on DL for speech signal processing has shown momentous advacment in the performance over traditional models like Gaussian Mixture Model and Hidden Markov Models [200]. There are different types of deep learning used for various signal processing as shown in Fig. 4 [200]. Many of them have been used in speech processing such as Deep Neural Network, Auto-Encoder, Generative Adversarial Network, Deep Belief Network, Deep Reinforcement LearningRestricted Boltzmann Machine, Convolutional Neural Network, and Recurrent Neural Network.

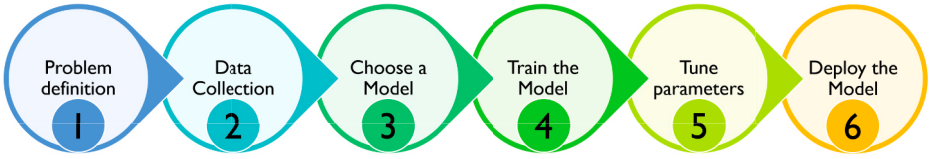
In the following section, we will delve into an overview of deep neural networks and their types, providing insights into their applications and capabilities.

*3.3. Deep neural network (DNN)*

Human communication is a complex process that depends on the transmission of acoustic energy waves, generated by the human vocal

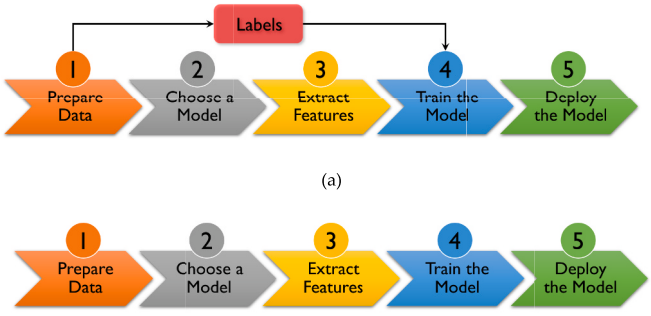



(a)




(b)

**Fig. 2.** (a) Main Types of Machine Learning Techniques (b) Main steps of machine learning work.




(a)

(b)

**Fig. 3.** (a) The stages of supervised learning (b) The stages of unsupervised learning [199].

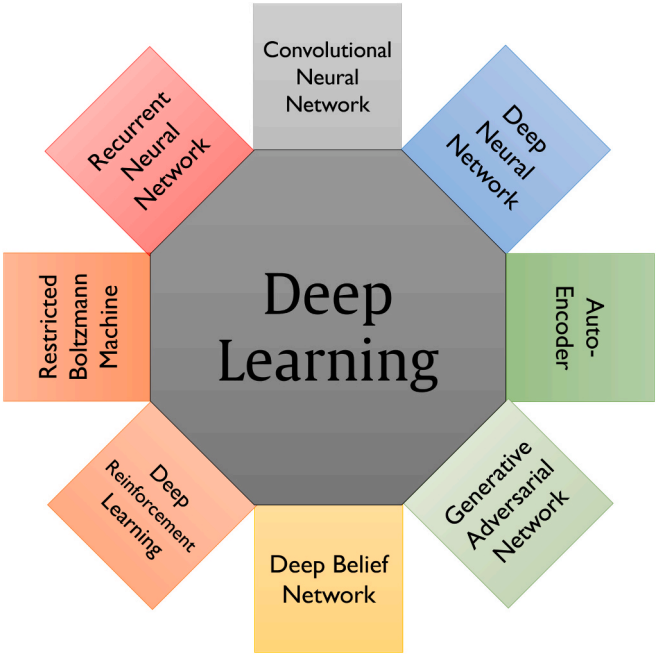



**Fig. 4.** Different general types of deep learning.

tract, to convey meaning and intent between individuals and machines. The field of speech research has been focused on improving speech data synthesis and analysis, aiming to develop applications that seamlessly integrate with human society. By addressing the challenges and issues faced by society, researchers strive to create innovative solutions that enhance quality of life and facilitate meaningful interactions between humans and machines. Speech data or signals contain a vast amount of information. To extract the actual information, speech recognition is developed, and for clear amplified data extraction, speech enhancement is introduced [139,147,158]. It is crucial to segregate all this information for dedicated speaker analysis. This implementation involves a significant amount of data and processing layers. To manage this complex process, deep neural networks (DNNs) were introduced in the field of speech synthesis, akin to the human neural system. Fig. 5 illustrates the types of DNN architectures used for speech synthesis over the past few decades, and more detailed information can be found in sections 3.3.1, 3.3.2, 3.3.3, and 3.3.4. DNNs have the capability to handle large data segments for complex data processing, with an advanced learning ability to provide optimal outcomes for applications developed by the system. In today's world, this technology is also known as "artificial

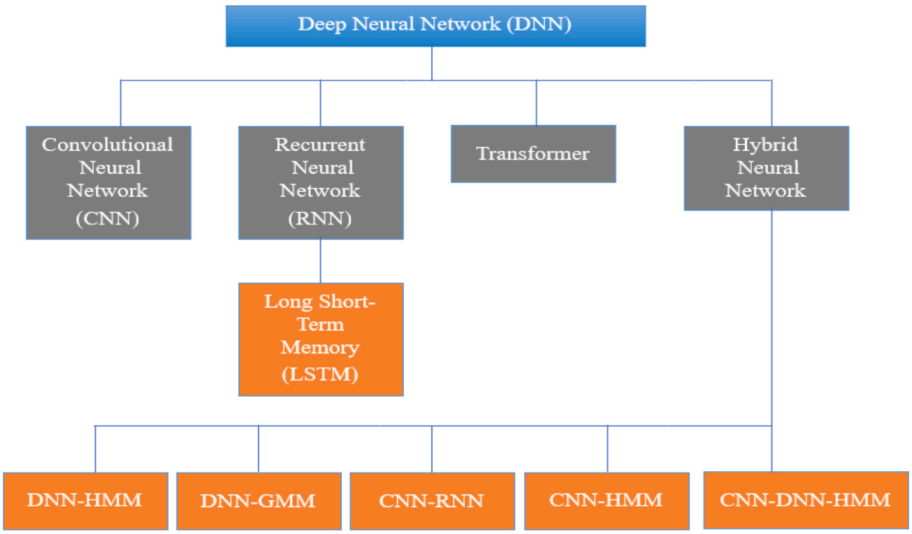



**Fig. 5.** Types of deep neural network.

intelligence (AI)," representing the future of advanced technological development in human society. In addition to this, it also facilitates learning and testing to achieve the best outcomes in the implemented system. This is achieved through multi-layer backgrounds to deliver accurate information for system development. Numerous applications have been introduced in recent years, with a notable achievement being the development of customer-level products with advancements such as "Alex," an AI-based smart device controlled by human speech. Many other technologies, including smart car control management systems, AI for desktops like Microsoft CORTANA for Windows 10, Apple Siri, Amazon Alexa, Google Home, and more, have been developed using DNNs. Observing this trend, the growth of DNN in the field of speech synthesis is expected to become even more advanced and futuristic [87,88].

This systematic review presenting a comprehensive study on the utilization of DNNs in the area of speech enhancement and speech recognition since there are a significant advanced in these fields recently [165]. Specifically, DNNs have demonstrated effectiveness in noise suppression for speech enhancement, a critical step in enhancing the quality of degraded or noisy speech [89–91,150,151,186]. When referring to the input for a system or device, we are addressing all the noisy elements, reverberant signals, and unwanted speech data. The introduction of beamforming processes has greatly improved the speech enhancement process, particularly when combined with DNN technology, resulting in more accurate speech enhancement [178]. The utilization of DNNs in multichannel speech has led to better results, owing to the extensive use of training and testing sets to achieve the desired outcomes. Researchers have explored hybrid combinations, such as LSTM-RNN-based DNN for speech enhancement [92], indicating ongoing improvement in DNN-based speech enhancement due to advanced research and the development of superior technology [49,93].

Speech recognition plays a crucial role in identifying spoken words and phrases, enabling us to extract meaningful information from noisy and multi-speaker audio data. By analyzing the acoustic features of audio data, speech recognition algorithms can accurately identify a dedicated speaker's voice from a combination of noisy and multi-speaker signals. This is made possible through the utilization of deep neural networks (DNNs), which can handle multiple tasks simultaneously, such as feature extraction and noise reduction, thereby improving the system's overall performance [61,94,116,118–120,122,161–163]. Over the past few decades, researchers have explored various approaches to train these systems, including noise information comparison and speech information comparison. The results consistently indicate that speech training sets yield better outcomes than noise training sets [37].

The above review is related to the significant advancements in speech recognition and enhancement processes, driven by the development of DNN-based interfaces in recent decades. One of the primary challenges in speech recognition is extracting speaker-specific information from noisy and multi-speaker audio data. To address this, researchers have introduced various solutions, including the incorporation of multiple microphone array architectures. The array of microphones is designed to capture all speech information and unwanted noise, which is then processed by the beamformer to enhance the speech signal [51,95]. The benefits of this approach are evident in applications such as smart home automation, where accurate speech recognition is pivotal for effective voice control.

Fig. 6 illustrates the fundamental concept of distance speech enhancement through beamforming and the application of deep neural networks for speech recognition. Initially, a microphone array captures speech and background noise signals, which are then processed using a beamforming algorithm to improve the speech signal. The enhanced signal is subsequently used for DNN training and testing to identify the speech signal of the dedicated speaker. Additionally, Figs. 7 and 8 showcase the utilization of deep neural networks for single-channel and multi-channel speech signal processing, which are commonly employed by researchers conducting speech-related experiments.

### 3.3.1. *Convolutional neural network (CNN)*

A convolutional neural network (CNN) is a type of discriminative deep neural architecture. Each model in a CNN consists of a convolutional layer and a pooling layer, stacked on top of each other for smooth operation [96,97]. The convolutional layer shares a large amount of weight information, while the pooling layer sub-samples the output from the convolutional layer, decreasing the data rate for faster processing. Choosing the right pooling layer in relation to the convolutional layer results in the invariance properties of the CNN. Some have argued that the limited invariance properties of CNNs hinder their ability to recognize complicated patterns. However, CNNs have consistently delivered the best results for computer vision and image processing tasks. The major advantage of using the CNN model is its specialized linear operation, which consists of three key elements: sparse interactions, parameter sharing, and equivalent representation [26]. These elements prove useful for various applications. In fact, the CNN model has been successfully applied in speech enhancement and speech recognition processes [50,98–100,115,135,145,159,160,168].

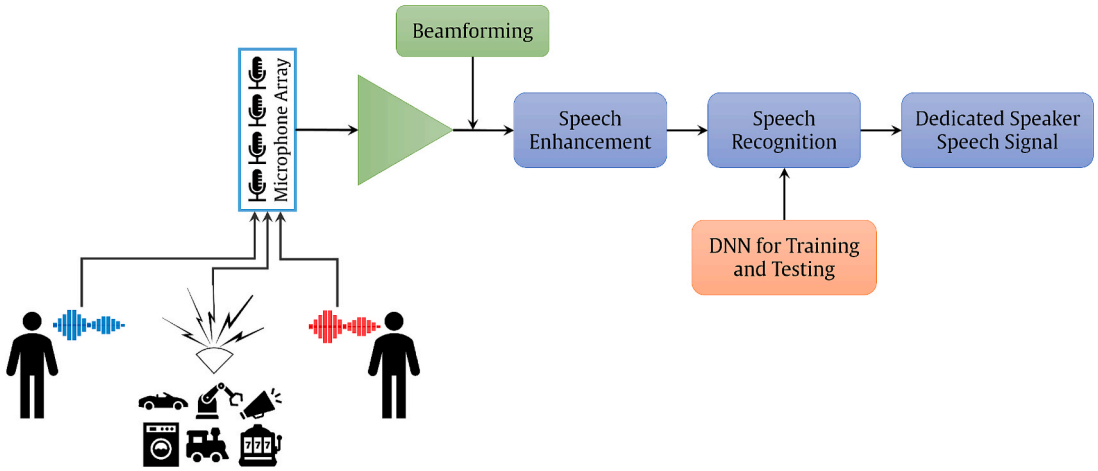



**Fig. 6.** Basic representation of distance speech enhancement using beamforming, speech recognition using deep neural networks.

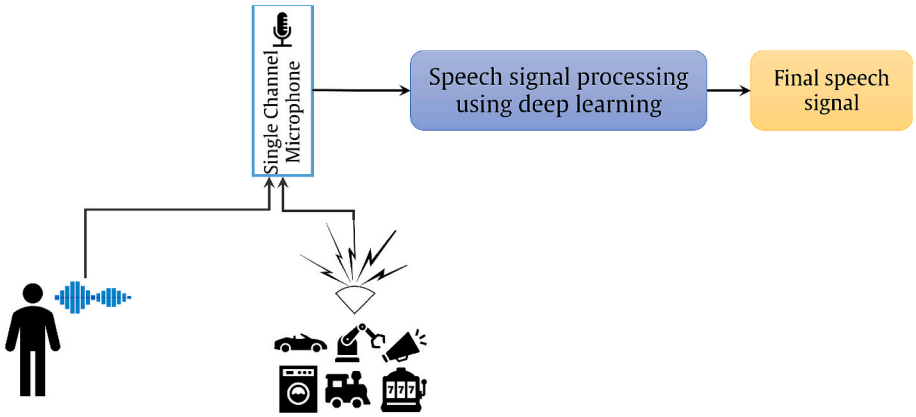



**Fig. 7.** Single channel speech signal processing using deep neural network.

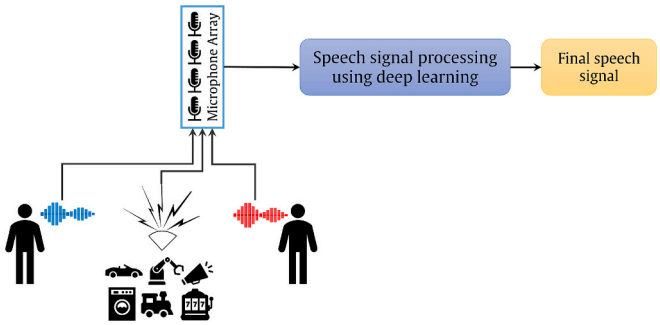



**Fig. 8.** Multi-channel speech signal processing using deep neural network.

### 3.3.2. *Recurrent neural network (RNN)*

A recurrent neural network (RNN) is a type of DNN model used for unsupervised learning, especially when dealing with input data sequences of significant depth. One of its key capabilities is the sharing of parameter across different layers of the network. RNNs are primarily used in scenarios where the same sets of weights are applied recursively over a tree-like structure, which is represented in a topological order. They are commonly employed for predicting future data sequences based on previous data samples and are especially useful for modeling sequence data such as speech or text. However, these types of models are not widely used because they require long-term dependencies and are difficult to train the system; in recent studies, however, they have been employed due to advancements in handling complex data. Moreover, they are integrated into speech recognition applications and speech enhancement processes [146,183], which involve a complex process of processing and analysis [101–103]. This type of RNN is also known as long and short-term memory neural networks due to the processing of large amounts of data for short or long periods according to developers' requirements [104–106,123,130].

### 3.3.3. *Transformer*

The Transformer represents a form of deep neural network that employs self-attention mechanisms to capture long-range dependencies and global context, making it especially suited for sequential data, including speech signals. In contrast to conventional RNNs and LSTMs, Transformers enable parallel processing of input sequences, which enhances computational efficiency and scalability. This functionality is vital for tasks related to speech enhancement and recognition, where it is crucial to accurately model both temporal and spectral aspects of noisy speech for effective performance in difficult acoustic environments [70,72].

Recent studies have shown the effectiveness of Transformer-based models across various speech processing tasks. For instance, the Time-Attention Transformer (TAT) was incorporated into convolutional encoder-decoder frameworks, allowing selective focus on temporal segments of speech, thereby improving both magnitude and phase data. This method resulted in notable enhancements in objective intelligibility and quality benchmarks, such as STOI and PESQ [203]. Similarly, the Dual-Path High-Order Transformer-Style Network (DPHT-ANet) used a hybrid framework that combined Transformers with recursive gated convolutions to capture both local and global dependencies in speech,

resulting in a model that provided substantial speech enhancement while minimizing computational demands, making it ideal for real-time use [206].

Additionally, the Deep Complex Convolution Transformer Network (DCCTN) was tailored specifically for cochlear implant users. By integrating complex-valued U-Net architectures with Transformer bottlenecks, DCCTN effectively tackled phase and harmonic distortions. Its distinct frequency transformation module and self-attention feature allowed it to restore both magnitude and phase components of noisy speech, achieving a marked improvement of up to 40 % in intelligibility and 31 % in quality under various noise conditions. This underscores its potential advantages for cochlear implant users in real-world settings characterized by noise [211].

In addition to its core functionality, Transformers demonstrate superior performance compared to traditional models like CNNs, RNNs, and LSTMs in terms of intelligibility and quality outcomes. They are particularly adept at diminishing non-stationary noise without creating artifacts, making them highly effective for speech enhancement. Furthermore, their ability to process data in parallel and adapt to hybrid architectures positions them as a promising avenue for future developments in both speech recognition and enhancement [69,70,72,202].

### 3.3.4. Hybrid neural network

In current research, there is a comprehensive focus on various aspects of speech processing, including speech enhancement and speech recognition applications using a hybrid form of neural network. The development of hybrid neural networks has led to significant advancements in speech enhancement, particularly in unseen noise conditions, outperforming standalone deep learning models [49,111,113]. The study of hybrid neural networks involves the integration of various types of neural networks [121], including convolutional HMM [107], convolutional recurrent neural networks [108], temporal convolutional neural networks [109], temporal convolutional recurrent neural networks [110], as well as the multiple integration of DNN [46] and object intelligibility integrations with LSTM and RNN [92]. By leveraging hybrid neural networks, researchers have tackled complex tasks that push the boundaries of speech processing. These approaches have successfully handled audio-visual speech enhancement and speech recognition [50,114], multiple noise-invariant speech, and integrated speech-related applications, showcasing the versatility of hybrid neural networks. Moreover, these implementations demonstrate the significance of hybrid neural networks in addressing more complex speech tasks, leading to improved training outcomes and enhanced performance in speech enhancement and speech recognition [190].

Fig. 9 illustrates the PRISMA flow chart showing a report of the obtained outcomesin each phase for the current systematic review. Fig. 9 outlines a flowchart detailing the literature search and filtering process. Initially, 302 papers were obtained from an online digital library database, followed by a series of filtering stages to finalize the 187 papers considered for the study. The detailed stepwise filtering process can be found in Fig. 9.

Upon reviewing information from various papers focused on speech enhancement and speech recognition, it becomes evident that extensive research has been conducted over the past few decades. This research has led to significant advancements in the realm of speech technology. In previous decades, the research primarily centered on different types of deep neural network (DNN) architectures and their practical applications. In recent years, hybrid DNN models have been introduced to meet the increasing demand for real-time speech processing, aiming to

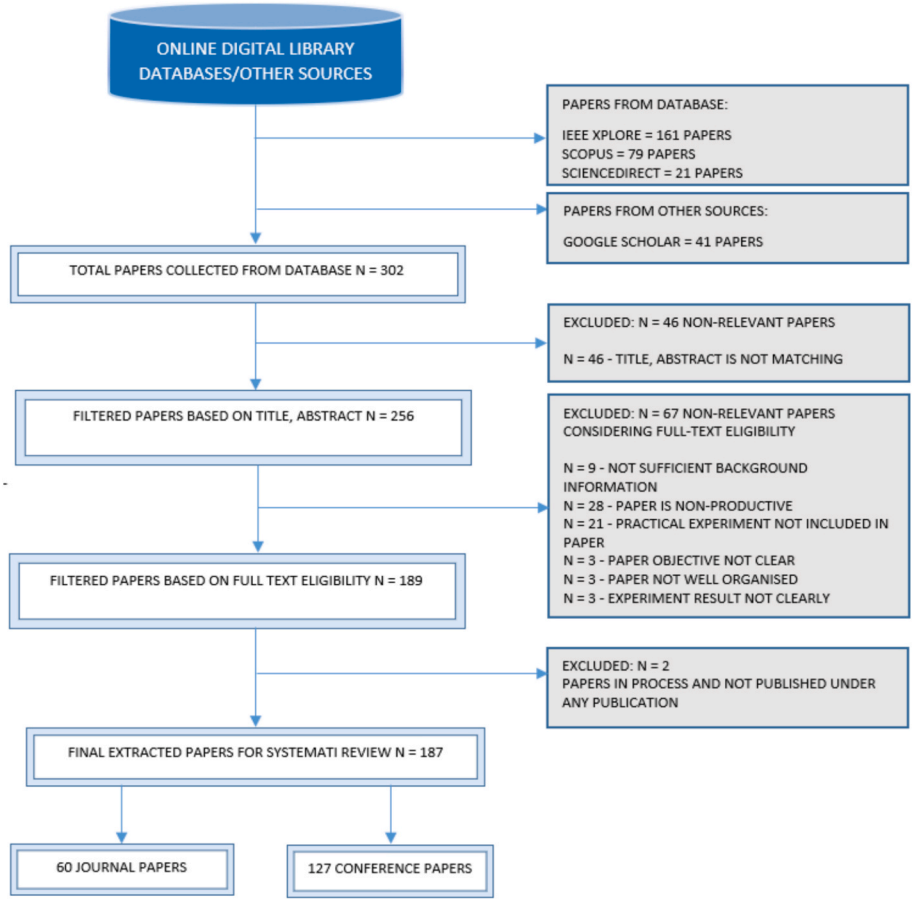



**Fig. 9.** PRISMA flow chart of the literature search and filtering process.

facilitate easier access to speech information for everyday use. These models have been built upon fundamental building blocks such as feature extraction, beamforming blocks for speech enhancement, and speech recognition, as illustrated in Fig. 6. Furthermore, notable progress has been made in multi-speaker analysis, leading to the exploration of a wide range of applications for real-time speech enhancement and speech recognition with significant potential for further development.

## 4. Methodology

Our paper is built upon a comprehensive systematic literature review [112]. The methodology is structured into three distinct phases: planning, conducting, and reporting. Within the planning phase, we have identified six key stages that are crucial to the success of this review. The first stage involves defining the research questions that align with the review's objectives. The second stage involves specifying the search strategy to retrieve relevant papers, determining the search terms and paper selection criteria. The third stage involves developing a rigorous study selection process with inclusion and exclusion rules. The fourth stage involves designing quality assessment criteria to filter out subpar papers. The fifth stage involves creating a data extraction approach to answer the research questions. The sixth and final stage involves synthesizing the extracted data from the research papers. The following subsections provide a detailed overview of the review protocol that was followed to ensure the quality and reliability of our findings.

### 4.1. Research questions

The main objective of this review paper is to answer the most frequently asked questions in speech signals related to deep neural networks in the area of speech enhancement and speech recognition. Based on these factors, the following research questions were identified as follows:

- RQ1: What are the different types of papers considered in the study?
- RQ2: What are the speech types identified in the research papers?
- RQ3: What kinds of databases are frequently used for training and testing algorithms in research papers?
- RQ4: What were the methods used to extract features from speech?
- RQ5: Which techniques were used for evaluation in the research papers?
- RQ6: What types of deep neural network models were used in the research papers?

### 4.2. Search strategy

The research strategy is to conduct a systematic review of research activities. Below is a detailed explanation of the search strategy:

#### 4.2.1. Search terms

The search terms used for this research were identified using three steps:

1. The mentioned questions above are the keys to main research terminology.
2. Some new research terms are from published papers and books.
3. Boolean operators like ANDs and ORs were used to put limits to our search results.

Given below are some examples of the terms we used in our search:

- "deep neural network" AND "speech".
- "deep learning" AND "speech".
- "DNN AND speech".
- "deep neural network" AND "speech recognition".
- "deep neural network" AND "speech enhancement".
- "beamforming" AND "speech".
- "deep neural network" OR "deep neural networks" OR DNN AND speech OR "beamforming" AND "speech".

#### 4.2.2. Survey resources

For the overview, large numbers of digital libraries and other sources were used as follows:

- IEEE Explorer;
- Science Direct;
- Scopus;
- Google Scholar.

#### 4.2.3. Search phases

The research papers were searched based on the aforementioned search terms from the specified digital libraries. The inclusion and exclusion criteria used are explained in detail in the following sections. Finally, 187 papers were included in this review based on the inclusion and exclusion criteria, which were determined through statistical analysis.

### 4.3. Study selection

Originally, 302 papers were identified through our search using the listed search terms. To ensure that only relevant papers were included in our study, the author carefully filtered and refined the list to select only those papers that are pertinent to this systematic review. The details of the filtration method are outlined below:

Step 1: All research papers were downloaded from a digital library database and other relevant sources.
Step 2: Each paper was filtered based on its relevance to our research subject, using the title and abstract as criteria.
Step 3: Papers filtered in Step 2 underwent further assessment based on quality criteria to ensure that they provided the best answers to our research questions. Full-text eligibility criteria were applied to remove papers that did not meet the quality assessment rule.
Step 4: Papers that met the full-text eligibility criteria underwent an additional filter to ensure that only those with proper publication details were considered and selected. This systematic approach was used to ensure that only the most relevant and high-quality papers were included in our review.

The Following are the inclusion and exclusion criteria used in this review:
Inclusion criteria:

- Conference papers or journal articles that investigate deep neural networks in the area of speech.
- Papers that explore deep learning in the context of speech.
- Studies that examine speech recognition and/ or speech enhancement using deep neural networks.
- Research papers that focus on speech recognition, and/ or speech enhancement using deep learning.

Exclusion criteria:

- Papers that contain deep neural networks in areas outside of speech.
- Papers that discuss speech but do not employ deep neural networks in their study.
- Studies related to speech recognition and/ or speech enhancement that do not utilize deep neural networks.
- Papers that lack clear publication information, such as author names, journal title, or date of publication.

### 4.4. Quality assessment rule (QAR) used in this study

To ensure the quality of our research papers, we developed a comprehensive quality assessment rule (QAR) that was applied to each paper. Our evaluation criteria were based on the relevance of each paper to our research questions. We identified ten QARs, each equally weighted, and scored on a scale of 0 to 1. The scoring criteria for each QAR were as follows: fully answered (1), above average answer (0.75), average answer (0.5), below average answer (0.25), and completely not answered (0). The overall score of each paper was calculated by summing the scores from each QAR. Papers with a total score of 6 or less were excluded from this review. The following QARs were used to evaluate the quality of the research papers:

1. **Organizational clarity**: Is the paper well-organized and easy to follow?
2. **Research objectives**: Are the research objectives clearly stated and relevant to the study?
3. **Background information**: Does the paper provide sufficient background information on the topic?
4. **Speech definition**: Is a specific area of speech clearly defined and used in the research paper?
5. **Practical application**: Does the paper involve practical experimentation?
6. **Experiment suitability**: Is the experiment performed suitable and acceptable?
7. **Dataset identification**: Is the dataset used in the paper clearly identified?
8. **Result reporting**: Are the results drawn from the paper clear and reported?
9. Methodological appropriateness: Are the methods used to analyze the results appropriate and well-justified?
10. **Overall productivity**: Is the paper productive and contributes meaningfully to the field?

In this section, we provide an example of how the paper titled 'Deep Casual Speech Enhancement and Recognition Using Efficient Long-Short Term Memory Recurrent Neural Network' was selected for inclusion in our systematic review, illustrating the application of the above-mentioned methodology:

Now, let us walk you through the specific example of the paper, "Deep Casual Speech Enhancement and Recognition Using Efficient Long-Short Term Memory Recurrent Neural Network (published in PLOS ONE journal in 2024) which was included in our review.

- **Title & Abstract**: Initially, we reviewed the title and abstract of this paper to ensure it fit within the scope of our study. The title clearly indicated that it focused on speech enhancement and recognition, and the abstract highlighted the use of Long Short-Term Memory (LSTM), a type of deep neural network, which directly aligned with our research questions.
- **Full Text Eligibility**: Upon reviewing the full text, we found that the paper:
  - Provided sufficient background information on the challenges of speech enhancement and recognition and the application of LSTM networks.
  - Clearly defined research objectives, aimed at improving the performance of speech enhancement and recognition using LSTM-based approaches.
  - Included practical experiments where the authors applied their LSTM model to real-world speech datasets and reported experimental results that demonstrated improvements in both speech enhancement and recognition performance.
  - Was well-organized, with clearly defined sections for methodology, experimental setup, results, and conclusions.
- QAR Process & Scoring: The paper was assessed using the QAR process, and the scores for each of the 10 criteria were as follows:

1. Organizational clarity: 1 (Fully answered).
2. Research Objectives: 1 (Fully answered).
3. Background information: 1 (Fully answered).
4. Speech definition: 0.75 (Above average).
5. Practical application: 1 (Fully answered).
6. Experimental suitability: 1 (Fully answered).
7. Dataset identification: 1 (Fully answered).
8. Result reporting: 1 (Fully answered).
9. Methodological appropriateness:
10. Over productivity: 0.75 (Above average).

Total Score: 9.5 (This paper met the threshold of 6, indicating it was of high quality and passed the QAR process.).

Thus, this paper passed through both our initial filtration process (based on title, abstract, and full-text eligibility) and the QAR process, making it eligible for inclusion in the final set of 187 papers.

### 4.5. Data extraction strategy

To answer the research question posed in section 4.1, data was extracted from the final list of identified papers. The following information was collected: paper title, publication year, publication details, and answers to six specific research questions (RQ1 to RQ5). However, challenges were encountered during the data extraction process. Some papers presented unclear or inconsistent results, making it difficult to accurately extract and analyze the data, as the methodologies used to calculate the results were not adequately explained. Interestingly, not all research papers fully addressed every research question.

### 4.6. Synthesis of extracted data

The data collected for research questions RQ1-RQ3 and RQ5 was organized and presented in tabular form as numerical quantitative data. This data was then subjected to statistical analysis to compare the different findings for each research question. Through an examination of the resulting statistics, specific research patterns and trends from the past decade were revealed, providing valuable insights into key research directions. Regarding RQ4, the qualitative data was descriptively compared to identify the main similarities and differences among the included research papers.

## 5. Results

### 5.1. Research question 1

The study included 187 papers from four major types: conference papers, journal papers, workshop papers, and research institute publications. Fig. 10 illustrates the overall distribution of the extracted research papers.

It is noteworthy that the majority of papers used in this study, 60 %, were published in conferences, as shown in Fig. 10. The remaining 40 % were distributed between journals, workshops, and research entity publications. Specifically, journals contributed 32 % of the papers, while workshops and research entity contributed 7 % and 1 %, respectively. To provide a detailed overview of the extracted papers, statistical data were derived from different conferences and journals, published in Tables 7 and 8, respectively.

Table 7 presents a detailed distribution of papers among the 28 identified conferences. As shown in Table 7, the majority of conference papers, at 40.94 %, were published in ICASSP (IEEE International Conference on Acoustics, Speech and Signal Processing). This was followed by 29.13 % of the papers published in Interspeech, and then 4.72 % of the papers published in ASRU. Additionally, two different conferences, IWAENC and MLSP, had each been published at 2.36 %, while

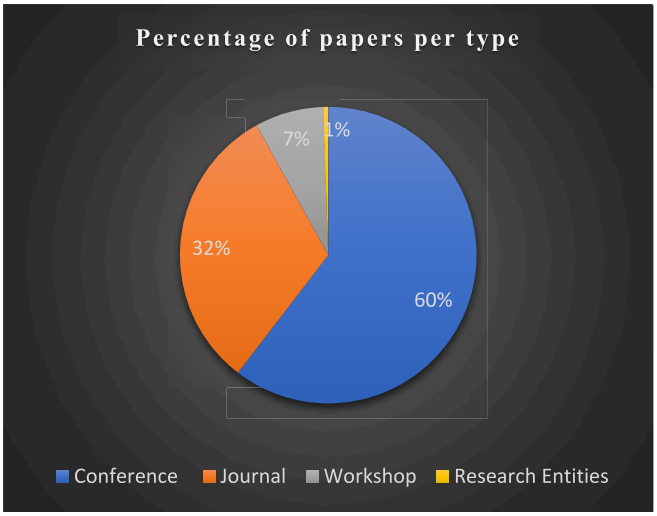



**Fig. 10.** Percentage of papers classified in each type.

**Table 7**
Distribution of conference papers.

| Name of conference | List of extracted papers published in different conferences | Total no. of papers | Percentage |
|---|---|---|---|
| ICASSP | 7, 9, 14, 16, 17, 20, 21, 22, 23, 32, 33, 35, 46, 51, 52, 62, 63, 66, 67, 69, 70, 76, 87, 88, 89, 91, 95, 100, 107, 108, 109, 111, 113, 119, 120, 127, 134, 142, 143, 144, 147, 151, 154, 156, 161, 164, 176, 177, 178, 184, 185, 188 | 52 | 40.94 |
| Interspeech | 3, 10, 12, 13, 18, 29, 30, 31, 36, 48, 49, 53, 65, 71, 72, 73, 78, 80, 90, 97, 101, 115, 116, 124, 125, 128, 135, 138, 146, 149, 150, 152, 155, 163, 166, 183, 186 | 37 | 29.13 |
| ASRU | 11, 47, 68, 122, 148, 165 | 6 | 4.72 |
| IWAENC | 24, 96, 104 | 3 | 2.36 |
| MLSP | 2, 98, 129 | 3 | 2.36 |
| ACII | 79, 121 | 2 | 1.57 |
| APSIPA ASC | 110, 141 | 2 | 1.57 |
| ICML | 94, 181 | 2 | 1.57 |
| ICSP | 6 | 1 | 0.79 |
| ISCSLP | 130 | 1 | 0.79 |
| LVA/ICA | 106 | 1 | 0.79 |
| ChinaSIP | 64 | 1 | 0.79 |
| ECMSM | 145 | 1 | 0.79 |
| HSCMA | 92 | 1 | 0.79 |
| ICCE | 105 | 1 | 0.79 |
| ICCIP | 19 | 1 | 0.79 |
| ICIS | 118 | 1 | 0.79 |
| ICMLA | 137 | 1 | 0.79 |
| ICMTMA | 126 | 1 | 0.79 |
| ICSigSys | 139 | 1 | 0.79 |
| ISQED | 132 | 1 | 0.79 |
| IOP | 99 | 1 | 0.79 |
| ISDFS | 140 | 1 | 0.79 |
| MMSP | 189 | 1 | 0.79 |
| SLT | 123 | 1 | 0.79 |
| DeSE | 191 | 1 | 0.79 |
| ICFTSC | 195 | 1 | 0.79 |
| SAS | 204 | 1 | 0.79 |

three different conferences named ACII, APSIPA ASC, and ICML were each published at 1.57 %. Finally, 0.79 % of the papers were published in each of the remaining 20 conferences. Similarly, Table 8 provides detailed information about the research papers published in journals.

Table 8 depicts a comprehensive breakdown of papers from the 26 identified journals. As shown in Table 8, the majority of the journal

**Table 8**
Distribution of papers over the identified journal.

| Name of Journal | List of extracted papers published in different Journals | Total no. of papers | Percentage |
|---|---|---|---|
| Computer Speech and Language | 1, 75 | 2 | 3.33 |
| Procedia Computer Science | 81, 133 | 2 | 3.33 |
| Speech Communication | 54, 102 | 2 | 3.33 |
| EURASIP Journal on Advances in Signal Processing | 27 | 1 | 1.67 |
| IEEE Access | 34, 77, 169, 170, 207, 208 | 6 | 10.00 |
| IEEE Journal of Selected Topics in Signal Processing | 4 | 1 | 1.67 |
| IEEE Signal Processing Letters | 44, 159, 172 | 3 | 5.00 |
| IEEE Transactions on Emerging Topics in Computational Intelligence | 50 | 1 | 1.67 |
| IEEE Signal Processing Magazine | 28 | 1 | 1.67 |
| IEEE/ACM Transactions on Audio, Speech, and Language Processing | 5, 8, 25, 45, 57, 58, 61, 93, 131, 136, 160, 171, 173, 174, 175, 179, 190, 211 | 18 | 30.00 |
| International Journal of Advanced Computer Science and Applications (IJACSA) | 117 | 1 | 1.67 |
| International Journal of Intelligent Enterprise | 103 | 1 | 1.67 |
| Modern Physics Letters B | 158 | 1 | 1.67 |
| Multimedia Tools and Applications | 153, 182 | 2 | 3.33 |
| Springer, Applied Intelligence | 114 | 1 | 1.67 |
| Springer, EURASIP Journal on Audio, Speech, and Music Processing | 162, 202, 210 | 3 | 5.00 |
| Springer, International Journal of Speech Technology | 157 | 1 | 1.67 |
| Applied Acoustics | 55, 167, 168, 206 | 4 | 6.67 |
| PeerJ Computer Science | 60, 205 | 2 | 3.33 |
| PLOS ONE | 180 | 1 | 1.67 |
| Symmetry | 187 | 1 | 1.67 |
| Journal of Robotics and Mechatronics | 74 | 1 | 1.67 |
| Journal of the American Statistical Association (JASA) | 59 | 1 | 1.67 |
| Neural Networks | 56 | 1 | 1.67 |
| IET Signal Processing | 15 | 1 | 1.67 |
| Digital Signal Processing | 203 | 1 | 1.67 |

papers were selected from IEEE/ACM Transactions on Audio, Speech, and Language Processing at 30 %, followed by 10 % published in IEEE Access. Applied Acoustics journals contributed 6.67 %. Both IEEE Signal Processing Letters and EURASIP Journal on Audio, Speech, and Music Processing contributed 5 % each. Five different journals, Computer Speech and Language, Procedia Computer Science, Speech Communication, Multimedia Tools and Applications, and PeerJ Computer Science, each accounted for 3.33 %. The remaining sixteen journals, including EURASIP Journal on Advances in Signal Processing, IEEE Journal of Selected Topics in Signal Processing, IEEE Transactions on Emerging Topics in Computational Intelligence, IEEE Signal Processing Magazine, International Journal of Advanced Computer Science and Applications, International Journal of Intelligent Enterprise, Modern Physics Letters B, Applied Intelligence, International Journal of Speech Technology, PLOS ONE, Symmetry, Journal of Robotics and

Mechatronics, Journal of the American Statistical Association (JASA), Neural Networks, IET Signal Processing, and Digital Signal Processing were each published at a rate of 1.67 %. This statistic provides a detailed distribution of the extracted papers in Table 8.

### 5.2. Research question 2

Among the 187 research papers, various areas of speech were identified, including speech recognition, speech emotion recognition, distant speech, speech enhancement, and other categories. Table 9 presents statistical information regarding the distribution of these different areas of speech-related papers. The majority of the papers, 44.92 %, fall under the speech enhancement area, followed by 34.22 % for speech recognition. Papers focused on distant speech accounted for 5.35 %, while papers covering both speech enhancement and speech recognition topics also accounted for 5.35 %. The speech emotion recognition category constituted 1.60 % of the papers, while the remaining 8.56 % fell into other categories. These data were obtained from various publications and publishing platforms.

### 5.3. Research question 3

Fig. 11 presents a list of several databases that were utilized in the research papers based on the experimental requirements. The TIMIT database was the most utilized at 17.74 %, followed by the WSJ at 15.60 % and the CHiME database at 11.62 %. The AURORA database accounted for 7.34 % usage, while the NoiseX-92 database was used in 5.20 % of the papers. LibriSpeech, and REVERB databases each contributed 3.98 % and 3.67 % of the papers respectively. This was followed by DEMAND database utilized in 3.06 % of the papers. Five different databases, namely the NOIZEUS database, VCTK, AISHELL, MUSAN and DNS-Challenge dataset, were used in each case at a 1.53 % rate. Furthermore, the own dataset was employed in 1.22 % of the papers. Four different databases, namely the IEEE database, AMI, GRID corpus, and VoiceBank dataset, were used in each case at a 0.92 % rate. The Speech Separation Challenge (SSC), Switchboard, JEIDA noise database, WHAMR!, and Multichannel Impulse Response Database (MIRD) were each used in 0.61 % of the papers.

The remaining 16.21 % of the papers employed different databases, each contributing less than 0.5 %, which are collectively referred to as others in Fig. 11. The specific names of all the additional databases are listed here:, Callhome, Chinese, CIFAR-10, EmoDB, English Broadcast News Speech Corpora, eNTERFACE'05, Berlin database, CRSS-4English-14 corpus, RSR2015 corpus, Fisher, FreeSFX, French database, Gale Mandarin, In house English video dataset, Interactive Emotional Dyadic Motion Capture, Internal Baidu corpora, ITU-T, Japanese database, ST-CMDS, AIDATATANG, McGill TSP speech database, Mixer 6 speech database, MNIST, Multi-Channel Impulse Response Database, NIST LRE 2007, NTT Multilingual Speech Database, HKUST Mandarin telephone speech, PN/NC version 1.0 corpus, RSR 2015 corpus, Litus ROUEN, Fisher-Swbd, DeepSpeech, Bing Mobile, Xbox, SMD, SVHN, SWB, Taiwan MHINT, TCD-TIMIT, APASCI, Euronew, DIRHA, INTERSPEECH 2021 ConferencingSpeech Challenge dataset, AudioSet, spoken English digits, Google Speech Command, Korean, Danish speech corpus Akustiske, Microsft-Internal Voice Search (VS), LibriCSS, JNAS, VOiCES, Turkish Microphone Speech Corpus (METU-1.0), Turkish Speech Corpus (TSC), ESC-50, UrbanSound8k, common voice corpus.

Based on the statistical analysis of the studies reviewed, we discuss the three most commonly used databases in the research literature in this section. It was found that the TIMIT database was the most widely utilized in speech enhancement and recognition, appearing in 58 papers. TIMIT is a well-established, high-quality database that provides a diverse collection of speech samples covering a wide range of phonetic contexts, making it ideal for training and evaluating speech recognition systems. Its carefully annotated phonetic transcriptions, diverse speaker demographics, and consistent recording conditions enable robust model training and benchmarking. As a result, TIMIT remains a widely used resource in speech processing research.

The Wall Street Journal (WSJ) database ranked as the second most frequently utilized resource, cited 51 times. WSJ is frequently used in speech recognition applications because of its extensive collection of high-quality recordings of spoken language, which are especially valuable for developing models applicable in real-life situations. Its vast vocabulary and clear recordings render it appropriate for both speech recognition and speech enhancement tasks that necessitate large datasets for effective generalization.

The CHIME database is ranked third, appearing in 38 studies. It is commonly used in research focused on robust speech recognition in noisy environments. Particularly, it is valuable for speech enhancement studies, as it contains recordings of speech under various noise conditions, providing an ideal testing ground for algorithms aimed at improving speech intelligibility in challenging auditory settings.

### 5.4. Research question 4

In the analysis, it was found that features were extracted from speech signals using various techniques and methods. More details regarding this can be found in Fig. 12. The most commonly used feature is the Mel-Frequency Cepstral Coefficients (MFCC), which appeared in 23.19 % of the papers. The MFCC method was used to extract features from speech signals. Following MFCC, the usage of Log Power Spectra (LPS) accounted for 14.01 % of the papers, while the short-time Fourier

**Table 9**
Different areas of speech the extracted research papers fall under.

| Area | Conference | Journal | Workshop | Research Entity | Total no. of Papers | Percentage |
|---|---|---|---|---|---|---|
| Speech Recognition | 13, 29, 31, 32, 33, 35, 36, 62, 63, 64, 66, 67, 69, 70, 71, 72, 87, 88, 91, 94, 99, 100, 107, 115, 116, 124, 126, 127, 130, 132, 138, 140, 141, 144, 152, 154, 181, 184, 185, 195 | 1, 4, 34, 61, 74, 75, 77, 81, 102, 103, 114, 117, 131, 133, 160, 162, 169, 170, 190, 205 | 68, 122, 165 | 28 | 64 | 34.22 |
| Speech Emotion Recognition | 79, 80, 121 | – | – | – | 3 | 1.60 |
| Distant Speech | 3, 30, 73, 105, 134, 149, 161 | 56, 153, 159 | – | – | 10 | 5.35 |
| Speech Enhancement | 9, 12, 16, 19, 46, 48, 49, 51, 52, 53, 76, 78, 89, 90, 92, 95, 97, 101, 106, 108, 109, 110, 111, 118, 120, 125, 128, 135, 137, 139, 142, 146, 147, 150, 151, 155, 156, 163, 166, 176, 177, 178, 183, 191, 204 | 5, 8, 15, 27, 44, 45, 50, 54, 55, 58, 59, 60, 93, 136, 157, 158, 167, 168, 172, 174, 175, 182, 187, 202, 203, 206, 207, 208, 210, 211 | 2, 24, 47, 96, 104, 123, 129, 145, 189 | – | 84 | 44.92 |
| Speech Enhancement and Speech Recognition | 7, 65, 113, 119, 186 | 25, 57, 179, 180 | 11 | – | 10 | 5.35 |
| Others | 6, 10, 14, 17, 18, 20, 21, 22, 23, 143, 148, 164, 188 | 171, 173 | 98 | – | 16 | 8.56 |

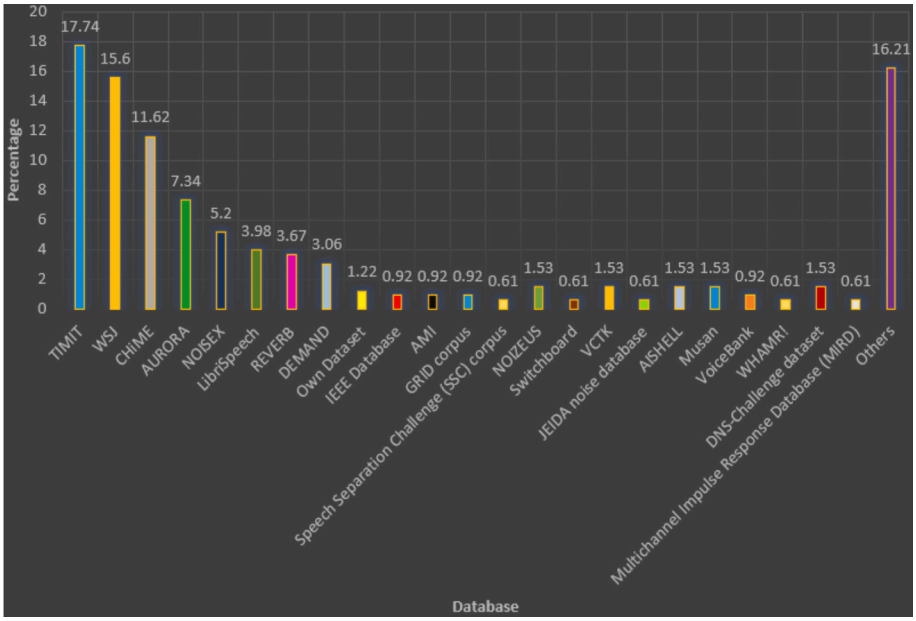



**Fig. 11.** Identified database from extracted list of papers.

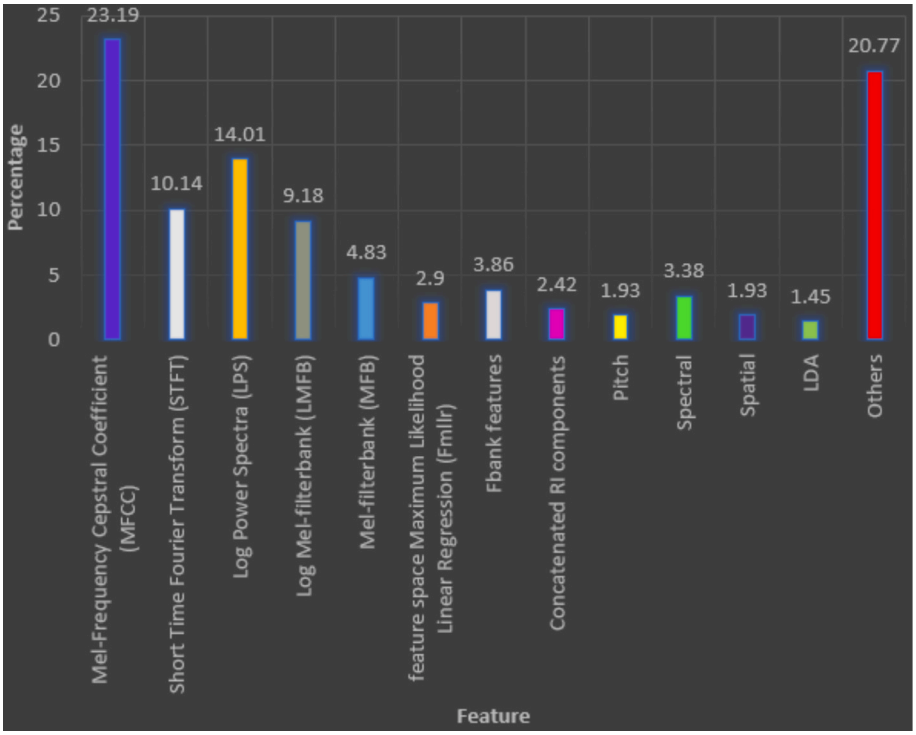



**Fig. 12.** Identified features used in the research papers.

transform was the third most used feature at 10.14 %. Additionally, a Log Mel Filterbank (LMFB) was utilized in 9.18 % of the papers, while the Mel Filterbank (MFB) was used in 4.83 %. Furthermore, Fbank Features were applied in 3.86 % of the papers, and spectral features were used in 3.38 % of the papers. The feature space Maximum Likelihood Linear Regression (Fmllr) was featured in 2.90 % of the papers, with 2.42 % utilizing Concatenated RI components details as features. Additionally, two features – Pitch, and Spatial – each contributed 1.93 %. LDA feature was used in 1.45 % of the papers. On the other hand, 20.77 % of the papers were classified as others because the features utilized by each paper scored less than 1 %. Some of the features falling under this category include Mel scale features, perceptual linear predictive (PLP), PSD, mel-frequency spectra, bark-frequency cepstrum coefficients (BFCC), cepstral features, DCT, FDLPM, GFE, IRM, log FFT, LDA, mel-features, MGDCC, MLLT, raw signal, SAT, spatial covariance matrices, spectrogram, mel-spectrogram, local and global features, discrete Charlie transform, complex ratio masking, complex compressed spectrum, and others.

This section presents a discussion on the three features most frequently cited in the research literature, derived from a statistical analysis of the examined studies. Mel-Frequency Cepstral Coefficients (MFCC) emerged as the most prevalent feature, appearing in 23.19 % of

the analyzed papers. This prevalence is expected, given that MFCC is well-established for its efficacy in both speech recognition and enhancement. In the field of speech recognition, MFCC effectively captures the spectral characteristics of speech, aligning closely with how the human auditory system perceives sound, which is vital for accurate phoneme identification. Likewise, in speech enhancement, MFCC maintains critical aspects of the speech signal while mitigating noise, thus serving as a powerful tool in noise reduction strategies.

Following MFCC, LPS and STFT ranked as the second and third most frequently used features, appearing in 14.01 % and 10.14 % of the papers, respectively. LPS is often deployed due to its effectiveness in representing spectral energy, which is crucial for various speech enhancement methods, including denoising. STFT, with its capability to provide time–frequency representations, is particularly well-suited for real-time analysis of speech signals, making it a valuable asset in enhancing speech within dynamic environments.

### 5.5. Research question 5

Several evaluation techniques were utilized in the research papers to assess the overall performance of the developed system. The different identified evaluation techniques are listed in Fig. 13. Notably, 22.12 % of the papers employed the word error rate (WER) to evaluate the system's performance, making it the most frequently used evaluation technique. Perceptual Evaluation of Speech Quality (PESQ) followed as the second most used technique, with 21.83 % of the papers utilizing it. Subsequently, STOI was utilized in 14.16 % of the papers, followed by signal-to-distortion ratio (SDR) and source-to-distortion ratio, each contributing 7.96 % and 1.47 % respectively. The accuracy metric accounted for 4.13 % of the research papers examined. Phoneme Error Rate (PER) contributed 3.83 % of the papers. Signal-to-Noise Ratio (SNR) metric was used in 3.54 % of the papers followed by Character Error Rate (CER) metric contributed 2.65 % of the papers. Extended Short Time Objective Intelligibility (ESTOI) metric was used in 2.36 % of the papers. Furthermore, 1.77 % of papers used Signal-to-Artifacts Ratio (SAR), while Log Spectral Distance (LSD) and Signal-to-Interference Ratio (SIR) were each accounted for 1.47 %. 1.18 % of the papers used Mean Square Error (MSE) as evaluation metrics. On the contrary, 10.03 % of the papers were classified as "others" as the techniques utilized in each paper scored less than 1 %. This category includes techniques such as WRR, precision, spectrogram, WCR, absolute gain, ceptral distance, emotion recognition accuracy, ΔSINR, ΔPESQ, ΔSTOI, EER, RMSE, COM, CSED, CQS, POLQA, RERR, LE, SER, correlation, spectrogram, recall, f1-score, hearing-aid speech perception index, hearing-aid speech quality index, alarm rates, mean opinion score (MOS) listening quality objective (LQO), noise hit rate, overall perceptual score, speech distortion index, sentence error rate, speech hit rate, scale-invariant source-to-noise ratio, speech-to-reverberation modulation energy ratio, speech reception threshold, Bak, Csig, and Covl.

This section analyzes the three most commonly used evaluation metrics identified through a statistical analysis of the reviewed studies. The analysis showed that WER and PESQ are the leading metrics, with 75 occurrences (22.12 % of papers) for WER and 74 occurrences (21.83 % of papers) for PESQ respectively. WER serves as the standard measure for assessing the performance of ASR systems by comparing the accuracy of transcriptions against reference outputs. This metric is essential for

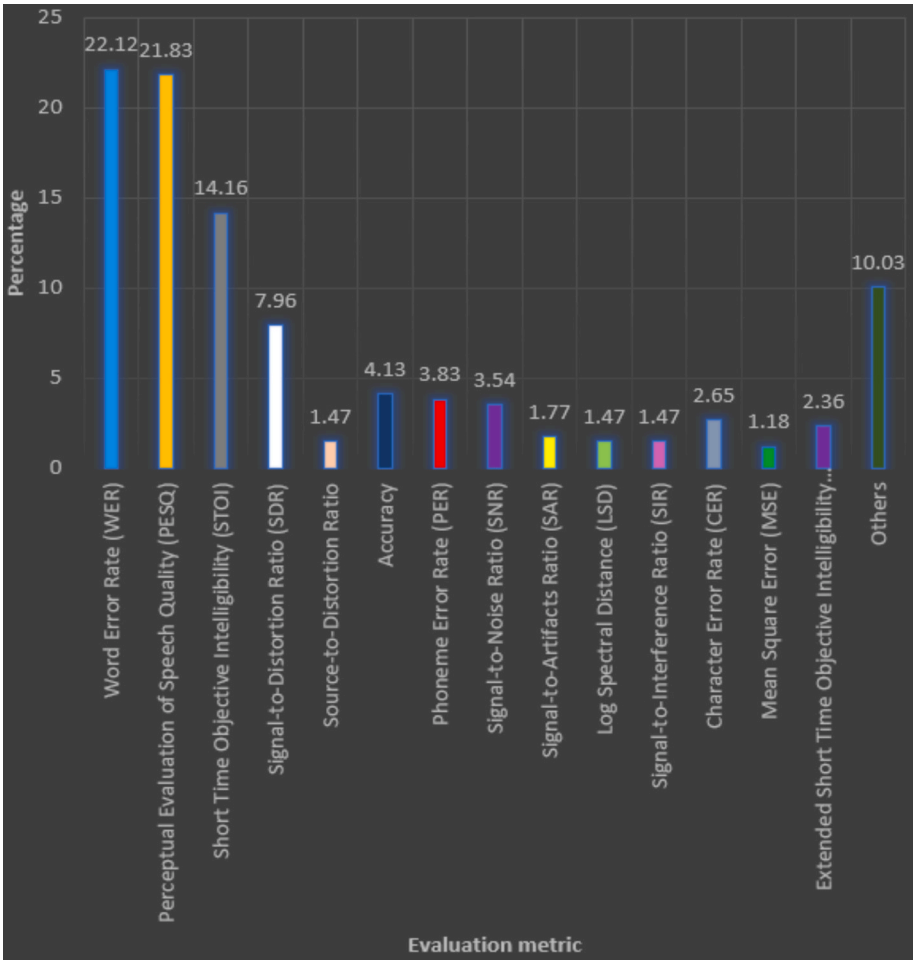



**Fig. 13.** Identified evaluation techniques used in the research papers.

evaluating ASR systems, where minimizing recognition errors is of utmost importance.

PESQ is a crucial metric in research on speech enhancement because it offers an objective assessment of the quality of enhanced speech based on perceptual speech quality, which closely correlates with human assessments. PESQ is particularly significant for evaluating improvements made in noisy conditions, where clarity of speech is vital. Its widespread adoption underscores its importance in determining the effectiveness of noise suppression and speech enhancement algorithms in enhancing audio quality.

Short-Time Objective Intelligibility (STOI) is the third most commonly utilized evaluation metric, appearing in 48 studies (14.16 % of papers), and is particularly relevant for speech enhancement. STOI assesses the intelligibility of speech, making it vital in scenarios where clarity and comprehensibility are the main objectives, such as in hearing aids or mobile communication systems operating in noisy environments.

### 5.6. Research question 6

Different types of deep neural networks were utilized in this review paper. Out of 187 papers, 144 employed DNN models as standalone models while 43 papers utilized hybrid models incorporating two or more models. Figs. 14 and 15 present the standalone and hybrid models, respectively. As illustrated in Fig. 14, majority (51.39 %) of the papers employed the DNN standalone model, followed by LSTM at 15.28 %. Furthermore, 13.19 % of the papers were based on CNN, followed by RNN at 5.56 %. The auto-encoder-based model was used in 4.17 % of the papers. Transformer model contributed at 2.08 % of the papers followed by Fully Convolutional Network (FCN) at 1.39 %. The remaining 6.94 % of the papers were based on various standalone models, each covering 0.69 %. Some examples of these standalone models include Deep convolutional neural network (DCNN), Deep Recurrent Neural Network (DRNN), Deep Denoising Autoencoder (DDA), Multi-Stream HMM (MSHMM), Recurrent Deep Neural Network (RDNN), Temporal Convolutional Neural Network (TCNN), Projected minimal Gated Recurrent Unit (PmGRU), Complex-Valued Spatial Autoencoder (COSPA), Time Domain Convolutional Neural Network (TDCNN), and Spiking Neural Network (SNN); additional standalone model names can be found in Fig. 14.

Based on the statistical evaluation of the reviewed literature, this section discusses the three standalone deep learning models that appear most prominently in the research. The analysis of the standalone deep learning models highlighted that Deep Neural Networks (DNN) were the most frequently employed, appearing in 74 studies (51.39 % of papers). DNNs are prevalent in tasks related to speech recognition and enhancement due to their capacity to identify complex data patterns, rendering them effective for tasks like phoneme classification and noise suppression. Their ability to model nonlinear relationships allows DNNs to establish intricate connections between input speech signals and desired outputs, making them a popular choice in contemporary speech processing frameworks.

Long Short-Term Memory (LSTM) networks, mentioned in 22 studies (15.28 % of papers), were the third most utilized model. LSTMs, a specific type of Recurrent Neural Network (RNN), excel at managing sequential data, which is crucial for speech recognition tasks due to the temporal dependencies inherent in phonemes. The memory architecture of LSTM networks enables them to retain long-term dependencies in speech signals, significantly enhancing the accuracy of recognizing continuous speech sequences.

Convolutional Neural Networks (CNN) ranked second, being referenced in 19 papers (13.19 % of papers). CNNs are especially effective for feature extraction in speech enhancement, as they can autonomously learn significant spectral and temporal features from raw speech data. Their design allows for spatial hierarchy processing, making them well-suited for recognizing intricate patterns within speech signals.

Fig. 15 illustrates the percentage of papers that employed various hybrid models in this review. A total of 13.95 % of the papers utilized the DNN-HMM hybrid model, followed by the CD-DNN-HMM model, which

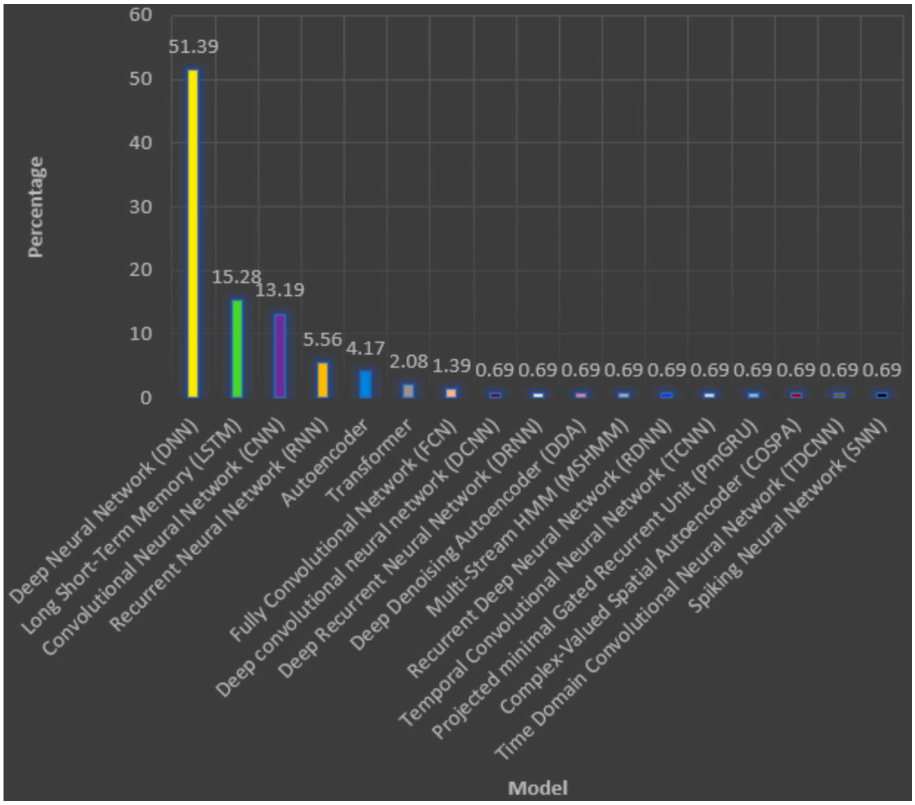



**Fig. 14.** Identified research papers based on standalone deep neural networks.

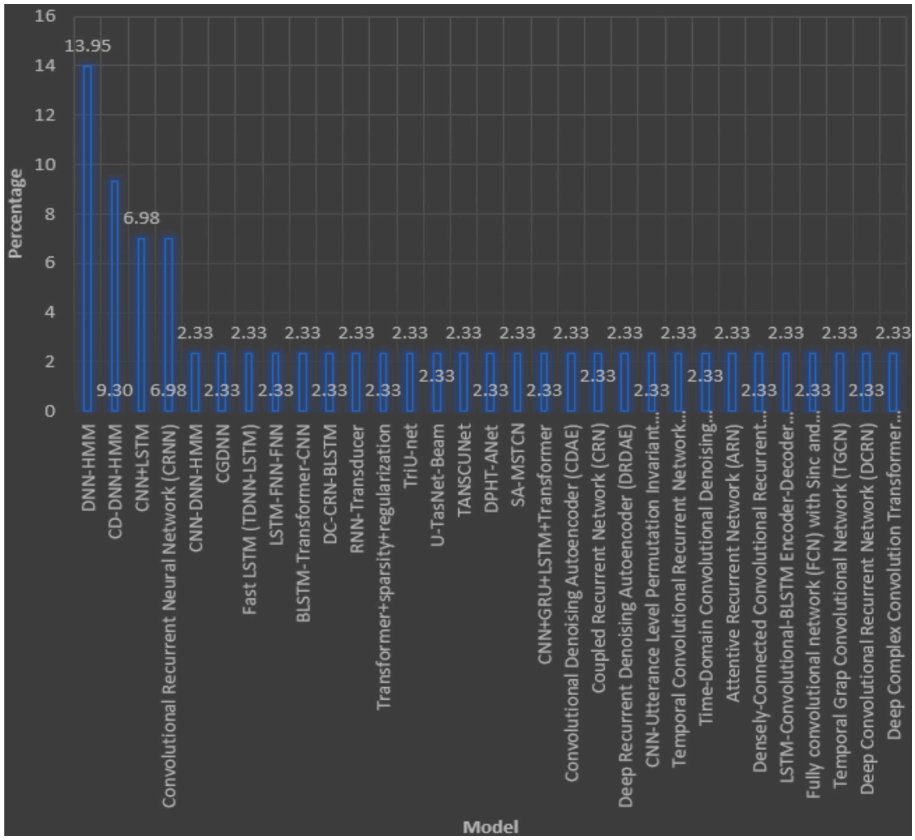



**Fig. 15.** Identified research papers based on hybrid deep neural networks.

was used in 9.30 % of the papers. Furthermore, both the Convolutional Recurrent Neural Network (CRNN) and the CNN + LSTM hybrid model accounted for 6.98 % of the papers each, indicating their equal representation in the research. The remaining 62.79 % of the papers employed different hybrid models, with each hybrid model covering 2.33 %. Examples of such hybrid models include CNN-DNN-HMM, CGDNN, Fast LSTM (TDNN-LSTM), LSTM-FNN-FNN, BLSTM-Transformer-CNN, DC-CRN-BLSTM, RNN-Transducer, Transformer + sparsity + regularization TriU-net, U-TasNet-Beam, TANSCUNet, DPHT-ANet, SA-MSTCN, and CNN + GRU + LSTM + Transformer, Convolutional Denoising Autoencoder (CDAE), Coupled Recurrent Network (CRN), Deep Recurrent Denoising Autoencoder (DRDAE), CNN-Utterance Level Permutation Invariant Training (CNN-uPIT), Temporal Convolutional Recurrent Network (TCRN), Time-Domain Convolutional Denoising Autoencoder (TCDAE), Attentive Recurrent Network (ARN), Densely-Connected Convolutional Recurrent Network (DCCRN), LSTM-Convolutional-BLSTM Encoder-Decoder (LCLED), Fully convolutional network (FCN) with Sinc and dilated convolutional layers (SDFCN), Temporal Grap Convolutional Network (TGCN), Deep Convolutional Recurrent Network (DCRN), Deep Complex Convolution Transformer Network (DCCTN); details of all other hybrid model names can be available in Fig. 15.

This section reviews the three hybrid models most frequently utilized in the research literature, as determined by a statistical analysis of the examined studies. Among hybrid models, LSTM-RNN emerged as the most widely used combination, as observed in numerous studies on speech enhancement and recognition. LSTM-RNNs integrate the strengths of RNNs in managing sequential data with the memory capabilities of LSTMs, making them highly effective for complex applications such as robust speech recognition in noisy environments.

The second most prevalent hybrid model was DNN-HMM (Hidden Markov Models), which appeared in several studies. This combination leverages the strengths of DNNs to model speech characteristics while employing HMMs to capture the temporal dependencies present in speech sequences, thereby improving the performance of recognition systems.

## 6. Discussion

Based on a comprehensive literature analysis of various deep learning approaches presented in the related work section, we have outlined the limitations of different deep learning models in Table 1 through Table 6, specifically concerning speech enhancement and speech recognition. Drawing from these findings, we aim to provide a stronger synthesis and highlight the broader implications for the fields of speech enhancement and recognition. The trends observed across various deep learning models underscore essential directions for future research and development.

For example, the limitations identified in traditional models such as DNNs and CNNs, particularly their slow convergence rates and high computational costs, point to the necessity for more efficient training techniques and model architectures. These challenges signal an increasing demand for the development of more resource-efficient models that can achieve faster convergence and operate effectively in real-time applications. Such advancements could significantly benefit industries that rely on speech processing, including consumer technology (e.g., smart devices and hearing aids), where low latency and computational efficiency are crucial.

Furthermore, the integration of hybrid models and advanced mechanisms, such as attention in architectures like RNNs and LSTMs, represents a broader shift towards combining the strengths of different approaches. This trend reflects the increasing complexity of speech processing tasks, particularly in noisy and dynamic environments. By blending various model architectures, researchers can enhance the robustness and flexibility of speech enhancement and recognition systems. These developments are vital for addressing real-world challenges,

such as speech enhancement in noisy settings and speech recognition in varied acoustic conditions, where no single model type may excel.

The rising adoption of transformer-based models further illustrates the evolving landscape of speech recognition, as these models have demonstrated superior performance in handling long-range dependencies and noisy inputs. As the field continues to embrace multimodal, hybrid, and attention-based approaches, we anticipate significant advancements in the ability of models to generalize across tasks and domains, making them more adaptable to diverse applications.

By synthesizing these trends, we can observe that the field is undergoing a substantial transformation. As deep learning models continue to evolve, we anticipate a shift towards architectures that combine efficiency, adaptability, and robustness, positioning themselves to better meet the demands of real-time, resource-constrained applications. Future research will likely focus on optimizing these models, addressing current limitations, and further integrating hybrid approaches to meet the diverse needs of both academic and industrial applications.

Ultimately, these findings not only provide insights into the current state of speech enhancement and recognition technologies but also pave the way for future innovations that will enhance their relevance and application in the broader field of speech technology.

## 7. Conclusions

This paper aims to compile the existing scientific knowledge about deep neural networks and their development in speech enhancement and recognition. By examining 187 papers published between 2012 and 2024, this study explores the statistical overview of speech applications. Notably, 153 standalone papers focused on DNN, while 34 explored hybrid models incorporating CNN, LSTM, and RNN. The majority of the papers (60 %) were conference papers, with a significant portion (40.94 %) originating from ICASSP. Additionally, 30 % of journal papers were published in IEEE/ACM Transactions on Audio, Speech, and Language Processing. The investigation emphasizes speech enhancement and speech recognition, revealing that 44.92 % of the papers focused on the former and 34.22 % on the latter. One interesting aspect to highlight is the use of diverse open-source databases, primarily in English, obtained from repositories such as TIMIT, WSJ, and LibriSpeech. TIMIT emerges as the most commonly used database, featured in 17.74 % of the papers. Furthermore, the recordings in these repositories were conducted in peaceful, natural settings with minimal background noise. Evaluation predominantly employed the word error rate (WER) as a performance metric, encompassing 22.12 % of the papers. This study not only serves as a valuable resource for current researchers but also aims to inspire future exploration and innovation in this field. Despite the advancements, it is worth noting the sustained reliance on traditional techniques like MFCCs for feature extraction, as indicated by their use in 23.19 % of the papers. Standalone deep neural network models have dominated the field, covering 81.82 % of the papers, with hybrid model-based papers comprising only 18.18 %. Significantly, a hybrid model has shown substantial improvement in speech recognition accuracy, achieving a 30 % relative word error rate (WER) in noisy and reverberant conditions, demonstrating its potential for effective speech recognition in challenging scenarios [57]. Another experiment has shown the advantage of using a hybrid model for speech enhancement tasks where the multi-task learning approach was applied, which uses the LSTM-RNN model, outperformed conventional DNN in unseen noise conditions to achieve significant improvement in both speech quality and intelligibility [92] and significant improvement in speech recognition using LSTM-RNN model [169]. Recent research has presented compelling evidence of substantial advancements in speech enhancement through the utilization of deep complex convolution transformers with frequency transformation. These developments have demonstrated practical applications, particularly in the context of Cochlear Implant Recipients [211]. Furthermore, the hybrid model has demonstrated considerable advancements in speech recognition tasks by outperforming standalone models such as GRU, LSTM, and Transformer [205].

In conclusion, the findings of this study underscore the continuous demand for innovation in the fields of speech enhancement and recognition technologies. Future research can focus on investigating the fusion of sophisticated deep learning frameworks, including Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) and Transformer models, with various other artificial intelligence methods like Convolutional Neural Networks (CNNs), attention mechanisms, and Generative Adversarial Networks (GANs). This fusion aims to tackle significant challenges in speech enhancement, such as mitigating noise and dereverberation, thereby enhancing the robustness and accuracy of speech recognition systems, particularly in environments with substantial noise and reverberation.

A promising path for future investigation could entail examining the synergistic combination of these deep learning architectures to exploit their respective strengths. For instance, integrating LSTM-RNNs with Transformer models could create a robust system that merges the sequential processing advantages of LSTMs with the attention-based methodologies offered by Transformers. This configuration has shown effectiveness in systems such as ChatGPT, which adeptly manage long-range dependencies and contextual nuances crucial for interpreting noisy or reverberant speech.

Another potential research avenue involves the inclusion of CNNs in these hybrid architectures to improve feature extraction capabilities, particularly in acoustically challenging environments. The proficiency of CNNs to capture essential acoustic features could greatly enhance the accuracy of speech recognition and the overall audio quality. Moreover, exploring attention mechanisms—either when combined with Transformer models or utilized independently—may provide effective strategies to focus on key speech components, thereby enhancing speech recognition quality by minimizing background noise and irrelevant information.

Additionally, the use of Generative Adversarial Networks (GANs) holds significant potential, as these networks can produce clean, noise-free audio from compromised input while maintaining the natural characteristics of spoken language. Merging GANs with other models is anticipated to deliver superior performance in noise reduction and reverberation suppression, thereby increasing adaptability in diverse environments.

Finally, utilizing recent advancements in self-supervised and transfer learning can open pathways to develop models necessitating fewer labeled datasets without compromising performance. Techniques such as domain adaptation and few-shot learning can enhance model durability and generalization when faced with unfamiliar acoustic scenarios.

Building on these methodological advancements, it is important to note that the predominant metric for evaluating speech recognition systems has been the Word Error Rate (WER). In future studies, integrating additional evaluation metrics including Phoneme Error Rate (PER), accuracy, precision, character error rate (CER), recall, and F1-score can yield more profound insights into the strengths and weaknesses of various models. Adopting a diverse array of assessment metrics will facilitate a more thorough understanding of system performance, enabling informed choices during the development and optimization of speech recognition technologies. Addressing these proposed research directions promises to advance speech recognition technology, broaden its applicability across different languages and contexts, and ultimately enhance user experiences across a wide range of speech-related applications.

## CRediT authorship contribution statement

**Sureshkumar Natarajan:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Syed Abdul Rahman Al-Haddad:** Visualization, Validation, Supervision. **Faisul Arif Ahmad:** Visualization, Validation, Supervision. **Raja Kamil:** Visualization, Validation,

Supervision. **Mohd Khair Hassan:** Visualization, Validation, Supervision. **Syaril Azrad:** Visualization, Validation, Supervision. **June Francis Macleans:** Project administration, Funding acquisition. **Sadiq H. Abdulhussain:** Writing – review & editing, Validation. **Basheera M. Mahmmod:** Writing – review & editing, Validation. **Nurbek Saparkhojayev:** Visualization, Validation, Conceptualization. **Aigul Dauitbayeva:** Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This research was funded by Research Management Centre, Universiti Putra Malaysia, Grant Putra 9695500.

## References

[1] Hori T, et al. Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend. *Elsevier, Computer Speech and Language* 2017;46:401–18.
[2] Ceolini E, Liu Shih-Chii. Combining deep neural networks and beamforming for real-time multi-channel speech enhancement using a wireless acoustic sensor network. In: Proc. IEEE MLSP; 2019. p. 13–6.
[3] Mirsamadi Seyedmahdad, Hansen John HL. "A study on deep neural network acoustic model adaptation for robust far-field speech recognition". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2015. p. 2430–4.
[4] Bo Wu, Li K, Ge F, Huang Z, Yang M, Siniscalchi SM, Lee Chin-Hui. "An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition". IEEE J Sel Top Signal Process 2017;11 (8):1289–300.
[5] Valentini-Botinhao C, Yamagishi J. Speech enhancement of noisy and reverberant speech for text-to-speech. *IEEE/ACM Trans Audio Speech Lang Process* 2018;26 (08):1420–33.
[6] Tu Yanhui Du, Yong Jun, Xu, Lirong Dai, Chin-Hui Lee. Deep neural network based speech separation for robust speech recognition. In: Proc. ICSP: IEEE; 2014. p. 532–6.
[7] Vu Thanh T, Bigot B, Siong Chng E. "Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2016. p. 499–503.
[8] Kolbæk M, Tan ZH, Jensen J. Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. *IEEE/ACM Trans Audio Speech Lang Process* 2017;25(01):153–67.
[9] Ming Tu, Zhang X. "Speech enhancement based on deep neural networks with skip connections". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2017. p. 5565–9.
[10] Bo Li TN, Sainath RJ, Weiss KW, Wilson, Bacchiani, M. "Neural network adaptive beamforming for robust multichannel speech recognition". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2016. p. 1976–80.
[11] Sivasankaran S, et al. "Robust ASR using neural network based speech enhancement and feature simulation". In: *Proc. IEEE Work- Shop Autom. Speech Recognit. Understand. (ASRU)*; 2015. p. 482–9.
[12] Wang ZQ, Wang D. "All-neural multi-channel speech enhancement". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2018. p. 3234–8.
[13] He Weipeng, Motlicek Petr, Odobez Jean-Marc. "Joint localization and classification of multiple sound sources using a multi-task neural network". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2018. p. 312–6.
[14] Qian K, Zhang Y, Chang S, Yang X, Florencio D, Mark HJ. "Deep learning based speech beamforming". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2018. p. 5389–93.
[15] Malek J, Koldovský Z, Bohac M. Block-online multi-channel speech enhancement using deep neural network supported relative transfer function estimates. IEEE, IET Signal Processing 2020;14(3):124–33.
[16] Zhao Yan, Buye Xu, Giri Ritwik, Zhang Tao. "Perceptually guided speech enhancement using deep neural networks". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2018. p. 5074–8.
[17] Meng Z, Watanabe S, Hershey JR, Erdogan H. "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2017. p. 271–5.
[18] Erdogan H, Hershey J, Watanabe S, Mandel M, Le Roux J. "Improved MVDR beamforming using single-channel mask prediction networks,". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2016. p. 1981–5.
[19] Shi W, Zhang X, Zou X, Sun M. "Experimental study on speech enhancement using DNN with perceptual weighting". In: *Proc. ICCIP*; 2018. p. 309–12.
[20] Zohrer M, Pfeifenberger L, Schindler G, Froning H, Pemkopf F. "Resource efficient deep eigenvector beamforming". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2018. p. 3354–8.
[21] Pfeifenberger L, Zohrer M, Pernkopf F. "DNN-based speech mask estimation for eigenvector beamforming". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2017. p. 66–70.
[22] Xiao Xiong, Zhao Shengkui, Jones Douglas L, Chng Eng Siong, Li Haizhou. "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2017. p. 3246–50.
[23] Xiao X, et al. "Deep beamforming networks for multi-channel speech recognition". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2016. p. 5745–9.
[24] Chazan SE, Gannot S, Goldberger J. "A phoneme-based pre-training approach for deep neural network with application to speech enhancement". *Proc. IEEE, IWAENC*. 2016.
[25] Bo Wu, Li K, Yang M, Lee C-H. A reverberation-time-aware approach to speech dereverberation based on deep neural networks. *IEEE/ACM Trans Audio Speech Lang Process* 2017;25(1):102–11.
[26] Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K. "Speech recognition using deep neural networks: a systematic review". *IEEE Access* 2019:19143–65.
[27] Xiao X, Zhao S, Ha Nguyen DH, Zhong X, Jones DL, Chng Eng Siong, Li Haizhou. "Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation". *EURASIP Journal on Advances in Signal Processing* 2016;2016(4):1–18.
[28] Hinton G, Deng L, Yu D, Dahl G, Abdel-rahman Mohamed N, Jaitly A, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Processing Magazine, Nov 2012;29(6):82–97.
[29] Zeyer A, Irie K, Schluter R, Ney H. "Improved training of end-to-end attention models for speech recognition". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2018. p. 7–11.
[30] Chen Szu-Jui, Subramanian Aswin Shanmugam, Hainan Xu, Watanabe Shinji. "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2018. p. 1571–5.
[31] Deng Li, Platt John C. "Ensemble deep learning for speech recognition". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2014. p. 1915–9.
[32] Weng C, Yu D, Watanabe S, Fred Juang BH. Recurrent deep neural networks for robust speech recognition. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process (ICASSP); 2014. p. 5532–6.
[33] Palaz D, Magimai M, Collobert DR. Convolutional neural networks-based continuous speech recognition using raw speech signal. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2015. https://doi.org/10.1109/ICASSP.2015.7178781.
[34] Jiang W, Wen F, Liu P. Robust beamforming for speech recognition using DNN-based time-frequency masks estimation. IEEE Access 2018;6:52385–92.
[35] Deng L, Hinton G, Kingsbury B. New types of deep neural network learning for speech recognition and related applications: an overview. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process (ICASSP); 2013. p. 8599–603.
[36] Abdel-Hamid O, Deng Li, Dong Yu. Exploring convolutional neural network structures and optimization techniques for speech recognition. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH); 2013. p. 3366–70.
[37] Shrestha A, Mahmood Ausif. Review of deep learning algorithms and architectures. In: Proc IEEE Access; 2019. p. 53040–65. https://doi.org/10.1109/ACCESS.20192912200.
[38] Ibrahim H, Varol A. A study on automatic speech recognition systems. In: International Symposium on Digital Forensics and Security (ISDFS); 2020. p. 1–5. https://doi.org/10.1109/ISDFS49300.2020.9116286.
[39] Lee W, Seong JJ, Ozlu B, Shim BS, Marakhimov A, Lee S. Biosignal sensors and deep learning-based speech recognition: a review. Sensors 2021;21(4):1399. https://doi.org/10.3390/s21041399.
[40] Alharbi S, Alrazgan M, Alrashed A, Alnomasi T, Almojel R, Alharbi R, et al. Automatic speech recognition: systematic literature review. IEEE Access 2021;9: 131858–76. https://doi.org/10.1109/ACCESS.2021.3112535.
[41] Dhanjal Amandeep Singh, Singh Williamjeet. A comprehensive survey on automatic speech recognition using neural networks. Multimedia Tools Applications 2023;83:23367–412. https://doi.org/10.1007/s11042-023-16438-y.
[42] Kheddar H, Hemis M, Himeur Y. Automatic speech recognition using advanced deep learning approaches: a survey. Inf Fusion 2024;109(102422):1–19. https://doi.org/10.1016/j.inffus.2024.102422.
[43] Shaughnessy Douglas O. Speech enhancement—a review of modern methods. IEEE Trans Hum-Mach Syst 2024;54(1):110–20. https://doi.org/10.1109/THMS.2023.3339663.
[44] Xu Y, Jun Du, Dai Li-Rong, Lee Chin-Hui. "An experimental study on speech enhancement based on deep neural networks". IEEE Signal Process Lett 2014;21 (1):65–8.
[45] Yong Xu, Jun Du, Dai L-R, Lee C-H. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Trans Audio Speech Lang Process* 2015; 23(1):7–19.
[46] Karjol P, Kumar A, Ghosh PK. Speech enhancement using multiple deep neural networks. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process (ICASSP); 2018. p. 5049–53.
[47] Vu Thanh T, Bigot B, Chng ES. Speech enhancement using beamforming and non negative matrix factorization for robust speech recognition in the CHIME-3

challenge. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understand (ASRU); 2015. p. 423–9. 13–17.
[48] Kumar A, Florencio D. Speech enhancement in multiple-noise conditions using deep neural networks. In: Proc. Annu. Conf. Int. Speech Commun. Assoc (INTERSPEECH); 2016. p. 3738–42.
[49] Nie Shuai, Liang Shan, Liu Bin, Zhang Yaping, Liu Wenju, Tao Jianhua. Deep noise tracking network: a hybrid signal processing/deep learning approach to speech enhancement. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH); 2018. p. 3219–23.
[50] Hou JC, Wang SS, Lai YH, Tsao Y, Chang HW, Wang HM. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Trans Emerging Top Comput Intell* April 2018;02(02):117–28.
[51] Furnon N, Serizel R, Illina I, Essid S. DNN-based distributed multichannel mask estimation for speech enhancement in microphone arrays. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP); 2020. p. 4672–6.
[52] Yemini Yochai, Chazan Shlomo E, Goldberger Jacob, Gannot Sharon. A composite DNN architecture for speech enhancement. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2020. p. 841–5.
[53] Bando Y, Sekiguchi K, Yoshii K. Adaptive neural speech enhancement with a denoising variational autoencoder. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH); 2020. p. 2437–41.
[54] Yuan W. A time–frequency smoothing neural network for speech enhancement. Speech Comm 2020;124:75–84.
[55] Cui X, Chen Z, Yin F. Multi-objective based multi-channel speech enhancement with BiLSTM network. Appl Acoust 2021;177(107927):1–13. https://doi.org/10.1016/j.apacoust.2021.107927.
[56] Li G, Liang S, Nie S, Liu W, Yang Z. Deep neural network-based generalized sidelobe canceller for dual-channel far-field speech recognition. Neural Netw 2021;141(2021):225–37.
[57] Zhang W, Chang X, Boeddeker C, Nakatani T, Watanabe S, Qian Y. End-to-end dereverberation beamforming, and speech recognition in a cocktail party. IEEE/ACM Trans Audio Speech Lang Process 2022;30(27):3173–88. https://doi.org/10.1109/TASLP.2022.3209942.
[58] Tesch K, Gerkmann T. Insights into deep non-linear filters for improved multi-channel speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 2022;31(10):563–75.
[59] Kuang K, Yang F, Li J, Yang J. Three-stage hybrid neural beamformer for multi-channel speech enhancement. J Acoust Soc Am 2023;153(6):3378. https://doi.org/10.1121/10.0019802.
[60] Cherukuru P, Mustafa MB. CNN-based noise reduction for multi-channel speech enhancement system with discrete wavelet transform (DWT) preprocessing. PeerJ Comput Sci 2024;10:1–34. https://doi.org/10.7717/peerj-cs.1901. PMC10909157.
[61] Dahl GE, Yu D, Alex Acero LD. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech Lang Process 2012;20(1):30–42.
[62] Seltzer ML, Yu D, Wang Y. An investigation of deep neural networks for noise robust speech recognition. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2013. p. 7398–402.
[63] Yu D, Yao K, Su H, Li G, Seide F. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2013. p. 7893–7.
[64] Meng X, Liu C, Zhang Z, Wang D. Noisy training for deep neural networks. In: Proc. IEEE, ChinaSIP; 2014. p. 16–20.
[65] Du J, Wang Q, Gao T, Xu Y, Dai L, Lee CH. Robust speech recognition with speech enhanced deep neural networks. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH); 2014. pp. 616–620.
[66] Gao Tian, Jun Du, Dai Li-Rong, Lee Chin-Hui. Joint training of front-end and back-end deep neural networks for robust speech recognition. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2015. p. 4375–9.
[67] Giri R, Seltzer ML, Droppo J, Yu D. Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2015. p. 5014–8.
[68] Battenberg E, Chen J, Child R, Coates A, Li YGY, Liu H, et al. Exploring neural transducers for end-to-end speech recognition. In: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), December 2017, Pp. 206–213. Okinawa, Japan; 2017. p. 206–13. https://doi.org/10.1109/ASRU.2017.8268937.
[69] Chang X, Zhang W, Qian Y, Le Roux J, Watanabe S. End-to-end multi-channel multi-speaker speech recognition with Transformer". In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP); 2020. p. 6134–8.
[70] Wang Y, et al. Transformer-based acoustic modeling for hybrid speech recognition. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, Pp. 6874–6878. Barcelona, Spain; 2020. p. 6874–8. https://doi.org/10.1109/ICASSP40776.2020.9054345.
[71] Lee M-H, Lee S-E, Seong J-S, Chang J-H, Kwon H, Park C. Regularizing transformer-based acoustic models by penalizing attention weights for robust speech recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech 18–22 September; 2022. p. 56–60. 10.21437/Interspeech.2022-362.
[72] Li J, Su R, Xie X, Wang L, Yan N. A multi-level acoustic feature extraction framework for transformer based end-to-end speech recognition. Proc Interspeech, International Speech Communication Association 2022;18–22: 3173–7. https://doi.org/10.21437/Interspeech.2022-915.
[73] Zhang W, et al. End-to-end far-field speech recognition with unified dereverberation and beamforming. In: Proc. ISCA Interspeech; 2020. p. 324–8.
[74] Nagano D, Nakazawa K. Simultaneous execution of dereverberation, denoising, and speaker separation using a neural beamformer for adapting robots to real environments. J Rob Mechatronics 2022;34(6):1399–410. https://doi.org/10.20965/jrm.2022.p1399.
[75] Falavigna D, Matassoni M, Jalalvand S, Negri M, Turchi M. DNN adaptation by automatic quality estimation of ASR hypotheses. *Elsevier, Computer Speech and Language* 2017;46:585–604.
[76] Zhang X, Wang Z, Wang D. A speech enhancement algorithm by iterating single and multi-microphone processing and its application to robust ASR. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process (ICASSP); 2017. p. 276–80.
[77] Sehga A, Kehtarnavaz N. A convolutional neural network smartphone app for real-time voice activity detection. IEEE Access 2018;6:9017–26.
[78] Guzewich P, Zahorian S, Chen X, Zhang H. "Cross-corpora convolutional deep neural network dereverberation preprocessing for speaker verification and speech enhancement. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH); 2018. p. 1329–33.
[79] Zheng WQ, Yu JS, Zou YX. An experimental study of speech emotion recognition based on deep convolutional neural networks. IEEE, ACII; 2015. p. 827–31.
[80] Guo L, Wang L, Dang J, Zhang L, Guan H, Li X. Speech emotion recognition by combining amplitude and phase information using convolutional neural network. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH); 2018. p. 1611–5.
[81] Wang W, Song W, Chen C, Zhang Z, Xin Yi. "I-vector features and deep neural network modeling for language recognition". *Elsevier bv, Procedia Computer Science* 2019;147:36–43.
[82] Sarker IH. "Machine learning algorithms, real-world applications and research directions". Springer Nature, SN Computer Science 2021;2(160):1–21.
[83] Alpaydin E. Introduction to machine learning. 3rd ed. Cambridge, MA, USA: MIT Press; 2015.
[84] Vladimir Nasteski. An overview of the supervised machine learning methods. 2017. 1-11.
[85] Rojas R. Unsupervised learning and clustering algorithms. Neural Networks Springer 1996:99–121. https://link.springer.com/chapter/10.1007/978-3-642-61068-4_5.
[86] Cho Y, Saul LK. 'Kernel methods for deep learning,'. Proc Adv Neural Inf Process Syst (NIPS) 2009;22:342–50.
[87] Li Deng J, Li JT, Huang K, Yao D, Yu F, Seide M Seltzer, GeojjZweig Xiaodong He, Williams J, Gong Y, Acero A. "Recent Advances in deep learning for speech research at microsoft". In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2013. p. 8604–8.
[88] Thomas S, Seltzer ML, Church K, Hermansky H. Deep neural network features and semi-supervised training for low resource speech recognition. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2013. p. 6704–8.
[89] Masuyama Y, Togami M, Komatsu T. Consistency-aware multi-channel speech enhancement using deep neural networks. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2020. p. 821–5.
[90] Liu D, Smaragdis P, Kim M. Experiments on deep learning for speech denoising. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH); 2014. p. 2685–9.
[91] Graves Alex, Mohamed Abdel-rahman, Hinton Geoffrey. Speech recognition with deep recurrent neural networks. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2013. p. 6645–9.
[92] Sun L, Du J, Dai LR, Lee CH. Multiple-target deep learning for LSTM-RNN based speech enhancement. In: Proc. Hands-Free Speech Commun. Microphone Arrays (HSCMA); 2017. p. 136–40.
[93] Yan-Hui Tu, Jun Du, Lee C-H. Speech enhancement based on teacher-student deep learning using improved speech presence probability for noise-robust speech recognition. *IEEE/ACM Trans Audio Speech Lang Process* Dec. 2019;27(12): 2080–91.
[94] T. Ochiai, S.i Watanabe, T. Hori and J. R. Hershey. Multichannel End-to-end Speech Recognition. Int. Conf. Mach. Learn. (ICML), Sydney. 2017. Australia, PMLR 70.
[95] Tan K, Zhang X, Wang D. Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2019. p. 5751–5.
[96] Chakrabarty S, Wang D, Habets EAP. "Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks". In: Proc. IEEE, IWAENC; 2018. p. 476–80.
[97] Gabbay A, Shamir A, Peleg S. Visual speech enhancement. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH); 2018. p. 1170–4.
[98] Kolbeak M, Yu D, Tan ZH, Jensen J. Joint separation and denoising of noisy multi-talker speech using recurrent neural networks and permutation invariant training. Proc. IEEE, MLSP. 2017.
[99] Poliyev AV, Korsun ON. Speech recognition using convolution neural network on small training sets. In: *Proc. IOP Conference Series: Material Science and Engineering, Workshop on Materials and Engineering in Aeronautics*; 2019. p. 1–5.
[100] Huang JT, Li J, Gong Yifan. An analysis of convolutional neural networks for speech recognition. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2015. p. 4989–93.
[101] Maas Andrew L, Le QV, ONeil TM, Vinyals O, Nguyen P, Ng Andrew Y. Recurrent neural networks for noise reduction in robust ASR. Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH). 2012.

[102] Tao F, Busso C. End-to-end audiovisual speech activity detection with bimodal recurrent neural models. *Elsevier, Speech Communication* Oct 2019;113:25–35.
[103] James PE, Kit MH, Vaithilingam CA. Recurrent neural network based speech recognition using MATLAB. International Journal of Intelligent Enterprise 2020;7 (1/2/3):56–66.
[104] Xiang Y, Bao C. Speech enhancement via generative adversarial LSTM networks. In: Proc. IEEE, IWAENC; 2018. p. 46–50.
[105] Trianto R, Tai TC, Wang JC. Fast LSTM acoustic model for distant speech recognition. *Proc. IEEE Conference on Consumer Electronics (ICCE)*. 2018.
[106] Weninger F, Erdogan H, Watanabe S, Vincent E, Le Roux J, Hershey JR, Schuller B. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In: Proc. HAL, International Conference on Latent Variable Analysis and Signal Separation; 2015. p. 1–9.
[107] Abdel-Hamid O, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2012. p. 4277–80.
[108] Zhao H, Zarar S, Tashev I, Lee Chin-Hui. Convolutional-recurrent neural networks for speech enhancement. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2018. p. 2401–5.
[109] Pandey A, Wang D. "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain". In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2019. p. 6875–9.
[110] Li J, Zhang H, Zhang X, Li C. "Single channel speech enhancement using temporal convolutional recurrent neural networks". In: *Proc. IEEE, APSIPA ASC*; 2019. p. 18–21.
[111] Kolbæk M, Tan ZH, Jensen J. "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2018. p. 505915–6320.
[112] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," Technical Report EBSE-2007-01, 9 July 2007.
[113] Yan-Hui Tu, et al. "A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2018. p. 2531–5.
[114] Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T. Audio-visual speech recognition using deep learning. Appl Intell 2014;42:722–37.
[115] Parcollet T, et al. "Quaternion convolutional neural networks for end-to-end automatic speech recognition". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2018. p. 22–6.
[116] Karita S, Watanabe S, CIwata T, Ogawa A, Delcroix M. "Semi-supervised End-to-End speech recognition". In: *Proc. Annu. Conf. Int. Speech Commun.assoc. (INTERSPEECH)*; 2018. p. 2–6.
[117] Dridi H, Ouni K. Towards robust combined deep architecture for speech recognition: experiments on TIMIT. *Int J Adv Comput Sci Appl* 2020;11(04): 525–34.
[118] Saleem N, Irfan M, Chen X, Ali M. "Deep neural network based supervised speech enhancement in speech-babble noise". In: Proc. IEEE, ICIS; 2018. p. 871–4.
[119] Feng X, Zhang Y, Glass J. "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2014. p. 1759–63.
[120] Araki S, Hayashi T, Delcroix M, Masakiyo Fujimoto, Takeda K, Nakatani T. "Exploring multi-channel features for denoising-autoencoder-based speech enhancement". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2015. p. 116–20.
[121] Li L, Zhao Y, Jiang D, Zhang Y, Wang F, Gonzalez I, Valentin Enescu, Sahli Hichem. "Hybrid Deep Neural Network–Hidden Markov Model (DNN-HMM) based speech emotion recognition". In: *Proc. IEEE, ACII*; 2013. p. 312–7.
[122] Pironkov G, Dupont S, Dutoit T. "Investigating sparse deep neural networks for speech recognition". In: *Proc. IEEE Work- Shop Autom. Speech Recognit. Understand. (ASRU)*; 2015. p. 124–9.
[123] Kolbæk M, Tan Z, Jensen J. "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification". In: *Proc. IEEE, Spoken Language Technology Workshop*; 2016. p. 305–11.
[124] Kondo K, Taira K, Kobayashi Y. "Binaural speech intelligibility estimation using deep neural networks". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2018. p. 1858–62.
[125] Llombart J, Ribas D, Miguel A, Vicente L, Ortega A, Lleida E. "Progressive speech enhancement with residual connections". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2019. p. 3193–7.
[126] Ling Z. "An acoustic model for english speech recognition based on deep learning". In: *Proc. IEEE, ICMTMA*; 2019. p. 610–4.
[127] Zhang Y, Chan W, Jaitly N. "Very deep convolutional networks for end-to-end speech recognition". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2017. p. 4845–9.
[128] Pandey A, Wang D. "A new framework for supervised speech enhancement in the time domain". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2018. p. 1136–40.
[129] Fu SW, Hu T, Tsaa Yu, Xugang Lu. "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning". In: *Proc. IEEE, MLSP*; 2017. p. 25–8.
[130] Chien JT, Misbullah A. "Deep long short-term memory networks for speech recognition". *Proc. IEEE, Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*. 2017.
[131] Li Bo, Khe Chai Sim. A spectral masking approach to noise-robust speech recognition using deep neural networks. *IEEE Trans Audio Speech Lang Process* 2014;22(8):1296–305.
[132] An Q, Bai K, Zhang M, Yi Yang, Liu Y. "Deep neural network based speech recognition systems under noise perturbations". In: *Proc. IEEE, ISQED*; 2020. p. 377–82.
[133] Rebai I, Ayed YB, Mahdi W, Lorre JP. "Improving speech recognition using data augmentation and acoustic model fusion". *Elsevier, Procedia Computer Science* 2017;112:316–22.
[134] Yoshioka T, Karita S, Nakatani T. "Far-field speech recognition using CNN-DNN-HMM with convolution in time". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2015. p. 4360–4.
[135] Park Se Rim, Lee Jin Won. "A fully convolutional neural network for speech enhancement". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2017. p. 3234–8.
[136] Zhao Y, Wang ZQ, Wang D. Two-stage deep learning for noisy-reverberant speech enhancement. *IEEE/ACM Trans Audio Speech Lang Process* 2019;27(1):53–62.
[137] Odelowo BO, Anderson DV. "A noise prediction and time-domain subtraction approach to deep neural network based speech enhancement". In: *16th IEEE International Conference on Machine Learning and Applications*; 2018. p. 372–7.
[138] Mack W, Chakrabarty S, Stoter FR, Braun S, Edler B, Habets EAP. "Single-channel dereverberation using direct MMSE optimization and bidirectional LSTM networks". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2018. p. 1314–8.
[139] Yang H, Choe S, Kim K, Kang HG. "Deep learning-based speech presence probability estimation for noise PSD estimation in single-channel speech enhancement". In: *Proc. IEEE, International Conference on Signals and Systems (ICSigSys)*; 2018. p. 267–70.
[140] A. H. Abdulqader, S. A. R. Al-Haddad, S. Abdo, A. Abdulghani, S. Natarajan, "Hybrid Feature Extraction MFCC and Feature Selection CNN for Speaker Identification using CNN: A Comparative Study," *2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, Ibb, Yemen, pp. 1–6, 2022. pp. 1–5. Beirut, Lebanon. DOI 10.1109/ ISDFS49300.2020.9116286.
[141] Zhao M, Wang D, Zhang Z, Zhang X. "Music removal by convolutional denoising autoencoder in speech recognition". In: *Proc. IEEE, APSIPA ASC*; 2015. p. 338–41.
[142] Kim M. "Collaborative deep leaning for speech enhancement: a run-time model selection method using autoencoders". In: *Proc. IEEE, ICASSP*; 2017. p. 76–80.
[143] Weng C, Yu D, Seltzer ML, Droppo J. "Single-channel mixed speech recognition using deep neural networks". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2014. p. 5632–6.
[144] Narayanan A, Wang D. "Joint noise adaptive training for robust automatic speech recognition". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2014. p. 2504–8.
[145] Kounovsky T, Malek J. "Single channel speech enhancement using convolutional neural network". *Proc. IEEE, International Workshop of Electronics, Control, Measurement, Signals and their Applications to Mechatronics (ECMSM)*, 24-26. 2017.
[146] Tan Ke, Wang DeLiang. "A convolutional recurrent neural network for real-time speech enhancement". In: *Proc. Interspeech, ISCA*; 2018. p. 3229–33.
[147] Chai Li, Jun Du, Lee Chin-Hui. "Error modeling via asymmetric laplace distribution for deep neural network based single-channel speech enhancement". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2018. p. 269–3273.
[148] Barker Jon, Marxer Ricard, Vincent Emmanuel, Watanabe Shinji. "The third 'chime' speech separation and recognition challenge: dataset, task and baselines". In: *2015, IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, hal-01211376; 2015. p. 504–11.
[149] Ravanelli M, Omologo M. "Contaminated speech training methods for robust DNN-HMM distant speech recognition". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2015. p. 756–60.
[150] Wang X, Bao C. "Masking estimation with phase restoration of clean speech for monaural speech enhancement". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2019. p. 3188–92.
[151] Bando Y, Mimura M, Itoyama K, Yoshii K, Kawahara T. "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2018. p. 716–20.
[152] Ko Tom, Peddinti Vijayaditya, Povey Daniel, Khudanpur Sanjeev. "Audio augmentation for speech recognition". *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. 2015.
[153] Oo Zeyan, et al. Phase and reverberation aware DNN for distant-talking speech enhancement. Springer, Multimedia Tools and Applications. 2018;77:18865–80.
[154] Bahdanau D, Chorowski J, Serdyuk D, Brakel P, Bengio Y. "End-to-end attention-based large vocabulary speech recognition". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2016. p. 4945–9.
[155] Tawara NT, Kobayashi T. Ogawa "Multi-channel speech enhancement using time-domain convolutional denoising autoencoder". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2019. p. 86–90.
[156] Hao X, Shan C, Xu Y, Sun S, Xie L. "An attention-based neural network approach for single channel speech enhancement". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2019. p. 6895–9.
[157] S. Raju C. and M. Tripathy, "Low SNR Speech Enhancement with DNN based Phase Estimation," *Springer International Journal of Speech Technology*, 22, pp. 283-292, 23 Feb. 2019.
[158] Shi W, Zhang X, Zou X, Han W. Deep neural network and noise classification-based speech enhancement. *Mod Phys Lett B* 2017;31(19–21).
[159] Swietojanski P, Ghoshal A, Renals S. Convolutional neural networks for distant speech recognition. *IEEE Signal Process Lett* 2014;21(9):1120–4.

[160] Abdel-Hamid O, et al. Convolutional neural networks for speech recognition. *IEEE/ACM Trans Audio Speech Lang Process* 2014;22(10):1533–45.

[161] Ravanelli M, Brakel P, Omologo M, Bengio Y. "A network of deep neural networks for distant speech recognition". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2017. p. 4880–4.

[162] S. Yin, C. Liu, Z. Zhang, Y. Lin, D. Wang, J. Tejedor, T. F. Zheng and Yinguo Li, "Noisy Training for Deep neural networks in Speech Recognition," *Springer, EURASIP Journal on Audio, Speech, and Music Processing, Article Number:2 (2015)*, pp. 1-14, 20 Jan. 2015.

[163] Xu Y, Du J, Dai Li-Rong, Lee Chin-Hui. "Dynamic noise aware training for speech enhancement based on deep neural networks". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2014. p. 2670–4.

[164] Yan-Hui Tu, Jun Du, Dai Li-Rong, Lee Chin-Hui. "Speech separation based on signal-noise-dependent deep neural networks for robust speech recognition". In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*; 2015. p. 61–5.

[165] Li Bo, Sim Khe Chai. "improving robustness of deep neural networks via spectral masking for automatic speech recognition". In: *Proc. IEEE Work- shop Autom. Speech Recognit. Understand. (ASRU)*; 2013. p. 279–84.

[166] Xugang Lu, Tsao Yu, Matsuda Shigeki, Hori Chiori. "Speech enhancement based on deep denoising autoencoder". In: *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*; 2013. p. 436–40.

[167] Wang Z, Zhang T, Shao Y, Ding B. LSTM-convolutional-BLSTM encoder-decoder network for minimum mean-square error approach to speech enhancement. Appl Acoust 2021;172:1–7. https://doi.org/10.1016/j.apacoust.2020.107647. 107647.

[168] Zhu Y, Xu X, Ye Z. FLGCNN: a novel fully convolutional neural network for end-to-end monaural speech enhancement with utterance-based objective functions. Appl Acoust 2020;170:1–9. https://doi.org/10.1016/j.apacoust.2020.107511. 107511.

[169] Oruh J, Viriri S, Adegun A. Long short-term memory recurrent neural network for automatic speech recognition. IEEE Access 2022;10(14):30069–79. https://doi.org/10.1109/ACCESS.2022.3159339.

[170] Feng R, Jiang W, Yu N, Wu Y, Yan J. 'Projected minimal gated recurrent unit for speech recognition'. IEEE Access 2020;8(01):215192–201. https://doi.org/10.1109/ACCESS.2020.3041477.

[171] Tan K, Wang ZQ, Wang DL. Neural spectrospatial filtering. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing; 2022. p. 605–21.

[172] Kim H, Kang K, Shin JW. Factorized MVDR deep beamforming for multi-channel speech enhancement. IEEE Signal Process Lett 2022;29(22):1898–902. https://doi.org/10.1109/LSP.2022.3200581.

[173] Quan C, Li X. SpatialNet: Extensively learning spatial information for multichannel joint speech separation. Denoising and Dereverberation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 2024;32(07):1310–23. https://doi.org/10.1109/TASLP.2024.3357036.

[174] Chau HN, Bui TD, Nguyen HB, Duong TTH, Nguyen QC. A novel approach to multi-channel speech enhancement based on graph neural networks. IEEE/ACM Trans Audio Speech Lang Process 2024;32(10):1133–44. https://doi.org/10.1109/TASLP.2024.3352259.

[175] Liu CL, Fu SW, Li YJ, Huang JW, Wang HM, Tsao Y. Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks. IEEE/ACM Trans Audio Speech Lang Process 2020;28(26):1888–900. https://doi.org/10.1109/TASLP.2020.2976193.

[176] Halimeh MM, Kellermann W. Complex-valued spatial autoencoders for multichannel speech enhancement. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022. p. 261–5. https://doi.org/10.1109/ICASSP43922.2022.9747528.

[177] Pandey A, Xu B, Kumar A, Donley J, Calamia P, Wang D. TPARN: triple-path attentive recurrent network for time-domain multichannel speech enhancement. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022. p. 6497–501. 10.1109/ICASSP43922.2022.9747373.

[178] Wang ZQ, Wang D. Multi-microphone complex spectral mapping for speech dereverberation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020. p. 486–90. https://doi.org/10.1109/ICASSP40776.2020.9053610.

[179] Wang Z-Q, Wang P, Wang DeLiang. Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR. IEEE/ACM Trans Audio Speech Lang Process 2020;28:1778–87.

[180] Li Zhenqing, Basit Abdul, Daraz Amil, Jan Atif. Deep causal speech enhancement and recognition using efficient long-short term memory Recurrent Neural Network. PLoS One 2024;19(1):1–19. https://doi.org/10.1371/journal.pone.0291240. e0291240.

[181] Amodei D, et al. Deep speech 2: end-to-end speech recognition in english and mandarin. In: Proc. Workshop Track ICML; 2016. p. 1–12.

[182] Garg A. Speech enhancement using long short term memory with trained speech features and adaptive wiener filter. Multimed Tools Appl 2023;82:3647–75. https://doi.org/10.1007/s11042-022-13302-3.

[183] Kuang K, Yang F, Li J, Yang J. DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement. J Acoust Soc Am 2023;153(6):3378. https://doi.org/10.1121/10.0019802.

[184] R. Kumar, A. Purushothaman, A. Sreeram, and S. Ganapathy. End-to-end speech recognition with joint dereverberation of sub-band autoregressive envelopes. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 2022. 6057–6061.

[185] Purushothaman A, Sreeram A, Ganapathy S. 3-D acoustic modeling for far-field multi-channel speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020. p. 6964–8. https://doi.org/10.1109/ICASSP40776.2020.9054481.

[186] Astudillo Ramon F, Correia Joana, Trancoso Isabel. "Integration of DNN based speech enhancement and ASR. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH; 2015. p. 3576–80.

[187] Liu W, Li A, Wang X, Yuan M, Chen Yi, Zheng C, et al. A neural beamspace-domain filter for real-time multi-channel speech enhancement. Symmetry 2022;14(1081):1–17. https://doi.org/10.3390/sym14061081.

[188] Nakatani T, Takahashi R, Ochiai T, Kinoshita K, Ikeshita R, Delcroix M, Araki S. DNN-supported mask-based convolutional beamforming for simultaneous denoising, dereverberation, and source separation. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP); 2020. p. 6399–403.

[189] Valin JM. A hybrid DSP/deep learning approach to real-time full-band speech enhancement. Proc. IEEE, MMSP. 2018.

[190] Cui X, Goeland V, Kingsbury B. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Trans Audio Speech Lang Process* 2015;23(9):1469–77.

[191] Jerjees SA, Mohammed HJ, Radeaf HS, Mahmmod BM, Abdulhussain SH. Deep learning-based speech enhancement algorithm using charlier transform. In: In 2023 15th International Conference on Developments in eSystems Engineering (DeSE); 2023. p. 100–5.

[192] Al-Zubaidi AS, Mahmmod BM, Abdulhussain SH, Naser MA, Hussain A. Low-distortion MMSE estimator for speech enhancement based on hahn moments. In: 2023 15th International Conference on Developments in eSystems Engineering (DeSE); 2023. p. 322–7.

[193] Mahmmod BM, Ramli AR, Baker T, Al-Obeidat F, Abdulhussain SH, Jassim WA. Speech enhancement algorithm based on super-gaussian modeling and orthogonal polynomials. IEEE Access 2019;7:103485–504.

[194] Mahmmod BM, Ramli AR, Abdulhussian SH, Al-Haddad SAR, Jassim WA. Low-distortion MMSE speech enhancement estimator based on laplacian prior. IEEE Access 2017;5(1):9866–81.

[195] Natarajan S, et al. Comparative analysis of different parameters used for optimization in the process of speaker and speech recognition using deep neural network. In: 2022 International Conference on Future Trends in Smart Communities, ICFTSC; 2022. p. 12–7. https://doi.org/10.1109/ICFTSC57269.2022.10040065.

[196] Wang H, Pandey A, Wang D. A systematic study of DNN based speech enhancement in reverberant and reverberant-noisy environments. Comput Speech & Lang 2024;vol. 89(2025):1–12. https://doi.org/10.1016/j.csl.2024.101677. 101677.

[197] Delic V, et al. Speech technology progress based on new machine learning paradigm. Comput Intell Neurosci, 2019;2019(1):4368036.

[198] Koblah D, et al. A survey and perspective on artificial intelligence for security-aware electronic design automation. ACM Trans Des Autom Electron Syst, 2023;28(2):1–57.

[199] Taye MM. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. Computers 2023;12(5):91. https://doi.org/10.3390/computers12050091.

[200] Bhangale KB, Kothandaraman M. Survey of deep learning paradigms for speech processing. Wirel Pers Commun, 2022;125(2):1913–49.

[201] Natarajan S, et al. A comprehensive review of beamforming-based speech enhancement techniques, IoT, and smart city applications. In: 2023 IEEE 2nd Industrial Electronics Society Annual on-Line Conference (ONCON); 2023. p. 1–6. https://doi.org/10.1109/ONCON60463.2023.10431158.

[202] Yecchuri S, Vanambathina SD. Sub-convolutional U-Net with transformer attention network for end-to-end single-channel speech enhancement. EURASIP Journal on Audio, Speech, and Music Processing 2024;8:1–15. https://doi.org/10.1186/s13636-024-00331-z.

[203] Saleem N, Gunawan TS, Dhahbi S, Bourouis S. Time domain speech enhancement with CNN and time-attention transformer. Digital Signal Process 2024;147:1–12. 104408.

[204] Sljubura N, Simic M, Bilas V. Deep learning based speech enhancement on edge devices applied to assistive work equipment. IEEE Sensors Applications Symposium (SAS) 2024:1–16. https://doi.org/10.1109/SAS60918.2024.10636511.

[205] Gormez Y. Customized deep learning based turkish automatic speech recognition system supported by language model. PeerJ Comput Sci 2024:1–22. https://doi.org/10.7717/peerj-cs.1981.

[206] Saleem N, Bourouis S, Elmannai H, Algarni AD. DPHT-ANet: Dual-path high-order transformer-style fully attentional network for monaural speech enhancement. Appl Acoust 2024;224:110131. https://doi.org/10.1016/j.apacoust.2024.110131.

[207] Alohali MA, Saleem N, Rhouma D, Medani M, Elmannai H, Bourouis S. Temporally dynamic spiking transformer network for speech enhancement. IEEE Access 2024;15:1–14. https://doi.org/10.1109/ACCESS.2024.3444596.

[208] Ali J, Saleem N, Bourouis S, Alabdulkreem E, El Mannai H, Dhahbi S. Spatio-temporal features representation using recurrent capsules for monaural speech enhancement. IEEE Access 2024;12(1):21287–303. https://doi.org/10.1109/ACCESS.2024.3444596.

[209] Al-Fraihat D, Sharrab Y, Alzyoud F, Qahmash A, Tarawneh M, Maaita A. Speech recognition utilizing deep learning: a systematic review of the latest developments. HCIS 2024;14(15):1–33. https://doi.org/10.22967/HCIS.2024.14.015.

[210] Zhang Z, Zhang L, Zhuang X, Qian Y. Supervised attention multi-scale temporal convolutional network for monoaural speech enhancement. EURASIP Journal on

Audio, Speech, and Music Processing 2024;20:1–16. https://doi.org/10.1186/s13636-024-00341-x.
[211] Mamun N, Hansen JHL. Speech enhancement for cochlear implant recipients using deep complex convolution transformer with frequency transformation. IEEE/ACM Trans Audio Speech Lang Process 2024;32:2616–29. https://doi.org/10.1109/TASLP.2024.3366760.
[212] Natarajan S, et al. "Revolutionizing speech clarity: unveiling a novel approach with hybrid coherent-to-diffuse power ratio and recursive least square algorithm for reverberation removal". 8th International Conference on Digital Signal Processing (ICDSP 2024), 23-25 February 2024, Hangzhou, China. 2024.

**Sureshkumar Natarajan** received his Bachelor and Master of Engineering degree in Electronics Engineering from Pillai College of Engineering (PCE), University of Mumbai, India in 2007 and 2012 respectively, where he worked as a lecturer from 2009 to 2012. He has submitted his final Ph.D. thesis in Computer and Communication Systems Engineering at Universiti Putra Malaysia, Malaysia. He achieved the best teacher award in 2017 from Vishwaniketan's Institute of Management Entrepreneurship and Engineering Technology (ViMEET), University of Mumbai, where he worked as an assistant professor from 2014 to 2018. He completed 45 Days tutoring program for students at the Technical University of Sofia, Bulgaria in 2016. His research interests include speech enhancement, speech recognition, speaker recognition, adaptive algorithms, deep learning and image processing.

**Syed Abdul Rahman Al-Haddad** is Professor and IEEE senior member. PhD graduated in Electrical, Electronic and Systems Engineering from National University Malaysia. His specialized in Human and Animal Sound Processing, Bio Data Processing, Media Security and Biometric. Lecturer at Department of Computer and Communications Systems Engineering, Universiti Putra Malaysia since 1997 and promoted as Associate Professor and Professor year 2012 and 2020. He taught students for undergraduate and graduate and managed to get International and national grants. Further than that, he has few patents and copyrights and actively join professional society such as Malaysia IEEE Systems Man and Cybernetics as Malaysia Chapter Past Chair.

**Faisul Arif Ahmad** (Member, IEEE) received the B.Eng. degree in information engineering from Muroran Institute of Technology, Muroran,Hokkaido, Japan, in 2001, the M.Eng. degree in electrical engineering from Universiti Teknologi Malaysia, in 2009, and the Ph.D. degree from Universiti Putra Malaysia (UPM), in 2016. He is currently a Senior Lecturer with the Department of Computer and Communication Systems Engineering, Faculty of Engineering, UPM, where he has served as a tutor and appointed as a Senior Lecturer, in February 2017. His research interests include robotic intelligent systems, swarm intelligence systems, embedded and real-time systems, and artificial intelligence systems. In 2001, he was an Engineer with Panasonic AVC Network Johor, Malaysia (formerly known as Matsushita Audio Video Sdn. Bhd.).

**Raja Kamil** received the B.Eng. degree in electrical engineering from the University of Southampton, UK, and the Ph.D. degree in control engineering from the University of Sheffield, UK. He is now an associate professor with the Department of Electrical and Electronic Engineering, a core member of the Control System and Signal Processing Research Center (CSSP) and a research associate of the Institute of Mathematical Research (INSPEM), Universiti Putra Malaysia. His research interests include active noise control, adaptive algorithms, nonlinear systems, machine learning, and the application of control and signal processing to biomedical engineering.

**Mohd Khair Hassan** received his BEng (Hons) degree in Electrical and Electronics from University of Portsmouth, United Kingdom in 1998 and MEng degree in Electrical Engineering from University of Technology Malaysia (UTM). He later completed his Ph.D. degree specializing in Automotive Engineering from University Putra Malaysia (UPM) in 2010. Currently, he is the Head of the Department of Electrical and Electronic Engineering, University Putra Malaysia and a registered Professional Engineer in the field of Electronic under the Board of Engineers Malaysia (BEM). His area of interest includes control system, automotive control, automation system, and AI applications. Currently, his research team is working on Electric Vehicle technology. The research emphasis on x-by-wire i.e. steer-by-wire, and brake-by-wire, energy management, battery modeling, battery balancing and regenerative braking.

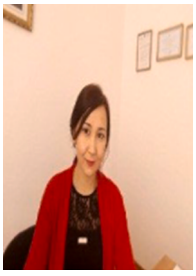
**Syaril Azrad** is currently a senior lecturer in the Department of Aerospace Engineering, Faculty of Engineering at Universiti Putra Malaysia, UPM. Prior to his appointment, he received his PhD in Mechanical Engineering from Chiba University in 2012. He completed his Bachelor and Master of Engineering in Mechanical Engineering from Tokyo University of Science in 2000 and 2002. His research interests are unmanned aerial vehicle control systems, high altitude balloons and platforms, cube-sat systems, vision-based control and systems engineering. He is engaged in various international and national projects and collaborations funded by the Ministry of Higher Education Malaysia, international universities, and companies. He is the regional facilitator for hepta-sat training for Malaysia and a member of University Space Engineering Consortium, UNISEC-Global based in Japan, which is envisioned to enable students from all countries to participate in space-related projects by 2030. He is also a committee member for Southeast Asia Network in Aerospace Engineering (SANAE), which organizes yearly workshops and research collaboration between aerospace-related education and research institutions. He is the contact point for Belt and Road Aerospace Innovation Alliance (BRAIA) for UPM.

**June Francis Macleans** received her Master of Computer Applications degree in 2005 from Shri Chimanbhai Patel Post Graduate Institute of Computer Applications from Gujarat University. She is an IT Professional with 15+ years of experience in leading banks in the Middle East, Information Technology companies in India and Reinsurance sector in South Asia. Currently, she holds the Director position of a Reinsurance Broking firm in Malaysia. She has extensive experience of working in software project analysis, designing, development, implementation and project management. She possesses impeccable project management and leadership skills. Her area of interest includes speech enhancement, speech recognition, deep learning and AI applications.

**Sadiq H. Abdulhussain** received the B.S. degree in electrical engineering from University of Baghdad in 1998 and the M.S. degree in electronics and communication engineering from University of Baghdad in 2001, and the Ph.D. degree in computer and embedded system engineering from Universiti Putra Malaysia, in 2018. Since 2005, he has been a staff member with the Department of Computer Engineering, Faculty of Engineering, University of Baghdad. His research interests include computer vision, signal processing, speech and image processing, and communication.

**Basheera M. Mahmmod** received the B.S. degree in electrical engineering from University of Baghdad in 1998 and the M.S. degree in electronics and communication engineering from University of Baghdad in 2012, and the Ph.D. degree in computer and embedded system engineering from Universiti Putra Malaysia, in 2018. Since 2007, she has been a staff member with the Department of Computer Engineering, Faculty of Engineering, University of Baghdad. Her research interests include speech enhancement, signal processing, computer vision, RFID, and communication.

**Nurbek Saparkhojayev** is Associate Professor and IEEE member. In 2006-2008. he was awarded the Bolashak International Scholarship of the President of the Republic of Kazakhstan and received a Master's degree in Computer Science and Computer Engineering at the University of Arkansas, USA. He is the holder of European Union Erasmus Mundus Fellowship. He did his PhD majoring in 6D070300 - Information Systems. He is member of the International Organization IEEE, IOT, SCIEI. He taught students for undergraduate and graduate and managed to get International and national grants. He specializes in Classification methods and algorithms, Supercomputing and High-Performance Computing, RFID fingerprinting and RFID Applications, Algorithms and networks, Sensor smart systems, Data science, Big Data, Data Analysis and Processing. Currently he works as Vice-rector for Research and International Relations at Karaganda Industrial University.



**Aigul Dauitbayeva** is a candidate of technical sciences, graduated from the Korkyt Ata Kyzylorda State University with a degree in Applied Mathematics. In 2011, she successfully defended her PhD thesis in the direction of system analysis, management and information processing. Senior lecturer, Candidate of technical sciences of the Department of "Computer Science" of the Korkyt Ata Kyzylorda University since 2008. She teaches lectures for undergraduate and graduate students. She is a member of the editorial board of the journal of the Korkyt Ata Kyzylorda University.