Jurnal Elektronika dan Telekomunikasi (JET), Vol. 21, No. 1, August 2021, pp. 19-26
Accredited by RISTEKDIKTI, Decree No: **32a/E/KPT/2017**
doi: 10.14203/jet.v21.19-26

Speech Enhancement Using Deep Learning Methods: A Review

Asri Rizki Yuliani [a, *], M. Faizal Amri [b], Endang Suryawati [a], Ade Ramdan [a], Hilman F. Pardede [a]

[a] *Research Center for Informatics*
Indonesian Institute of Sciences
Bandung, Indonesia
[b] *Technical Implementation Unit for Instrumental Development*
Indonesian Institute of Sciences
Bandung, Indonesia

Abstract

Speech enhancement, which aims to recover the clean speech of the corrupted signal, plays an important role in the digital speech signal processing. According to the type of degradation and noise in the speech signal, approaches to speech enhancement vary. Thus, the research topic remains challenging in practice, specifically when dealing with highly non-stationary noise and reverberation. Recent advance of deep learning technologies has provided great support for the progress in speech enhancement research field. Deep learning has been known to outperform the statistical model used in the conventional speech enhancement. Hence, it deserves a dedicated survey. In this review, we described the advantages and disadvantages of recent deep learning approaches. We also discussed challenges and trends of this field. From the reviewed works, we concluded that the trend of the deep learning architecture has shifted from the standard deep neural network (DNN) to convolutional neural network (CNN), which can efficiently learn temporal information of speech signal, and generative adversarial network (GAN), that utilize two networks training.

Keywords: speech enhancement, deep learning, neural networks, speech signal processing, non-stationary noise

I. INTRODUCTION

Speech is used by humans as an intermediary tool for communication. Speech chain model consists of two stages, speech production and speech perception [1]. Speech production is when a speaker expresses an idea through speech. When it happens, there is a conversion from linguistic structure into a sound wave pressure. While speech perception is the process in the auditory system of a listener. The task is interpreting the sound wave pressure coming from the speaker.

There has been a number of speech applications such as automatic speech recognition (ASR), spoken dialogue systems (e.g. voice dialing and data entry), digital hearing aids, etc. However, in real-world applications, speech signal is easily contaminated by external factors such as interference due to environmental noise, background noise, and reverberation. A listener often has difficulties understanding speech in the presence of these noises, especially when the signal-to-noise ratio (SNR) is at low level. Therefore, speech enhancement (SE) plays an important role in speech signal processing. SE is implemented to improve the intelligibility and quality of speech by removing noise from the corrupted speech signal [2]. The process of reducing the effects of noise is still very challenging in practice.

* Corresponding Author.
Email: asri006@lipi.go.id
Received: December 2, 2020 ; Revised: January 11, 2021
Accepted: January 27, 2021 ; Published: August 31, 2021

Recent advance of deep learning technologies has provided great support for the progress in SE research field. Unlike conventional SE approaches that depend on statistical model, deep learning approaches build on a data-driven paradigm. Now conventional SE such as spectral subtraction [3], Wiener filtering [4], and minimum mean square error [5], have been outperformed by deep learning methods. The development of deep learning is one of the most significant technology nowadays, hence deserves a dedicated survey. Some deep learning models for SE are using mapping-based methods [6]–[27], while some others are using masking-based methods [28]–[34].

Previously, there has been a lot of work presenting a survey in the speech field. The survey article by Zhang *et al.* [35] provides an extensive overview of relevant deep learning approaches, specifically for noise robust speech recognition task. A survey on audio-visual speech enhancement and separation based on deep learning is investigated by Michelsanti *et al.* [36]. The overview article by Das *et al.* [37] covers fundamentals of SE, but it only discusses deep learning based techniques in general. In this paper, we provide in detail a review of recent deep learning approaches that are designed to address SE task. We describe the advantages and disadvantages of these approaches. We also discuss challenges and trends of this field. Moreover, in order to carry this review, we selected papers which were published on the span of 8 years between 2012 and 2020. The query of 'deep learning' and 'speech enhancement' are used for the paper selection.

The remainder of this paper is organized as follows. Section II introduces the basic signal model and problem formulation of SE task. Section III reviews SE based on mapping and masking methods. Section IV presents the standard audio corpora and evaluation metrics that are frequently used. Section V describes deep learning-based SE methods. Finally, we provide conclusion, summarizing the challenges and trends of SE works presented in the paper.

II. SIGNAL MODEL AND PROBLEM FORMULATION

In real-world environments, speech signal is easily corrupted by noise. Noises, including reverberations, can be grouped into stationary noise (unchanging as a function of time) and non-stationary noise (changing when shifted in time). Example of background noise, which belongs to the non-stationary category, are street noise, train noise, babble noise (the voice of other speakers), and instrumental sounds. The relation between speech and noise in the time domain can be written as (1).

$$y(t) = s(t) * h(t) + n(t), \quad (1)$$

where $s(t)$ is clean speech signal, $h(t)$ is convolutional noise or room impulse response (RIR), $n(t)$ is additive noise, and $y(t)$ is noisy speech. Denoting $x(t) = s(t) * h(t)$ as target speech, we can rewrite (1) as (2).

$$y(t) = x(t) + n(t). \quad (2)$$

Let t be the time index. The signal can be represented as $y = [y(1), \dots, y(T)]$ where T is the length of the utterance. When applying short-time Fourier transform (STFT), we can represent the acoustic signal model of (2) in the time-frequency (TF) domain as (3).

$$Y(k,l) = X(k,l) + N(k,l), \quad (3)$$

where k is the frequency bin index, l denotes the time frame index, and $Y(k,l)$, $X(k,l)$, and $N(k,l)$ are the STFT coefficients of the noisy speech signal, the target signal, and the noise signal, respectively. The definitions provided above are valid for single channel microphone. In this case, SE task is aimed at recovering the target speech signal x from the noisy speech signal y. As for multichannel SE, the signal in the time domain is expressed as (4).

$$y_m(t) = x_m(t) + n_m(t), \quad m = 1,2,\dots,M. \quad (4)$$

where M is the total number of microphone array.

III. SPEECH ENHANCEMENT

According to the type of degradation and noise in the speech signal, approaches to SE vary. They are commonly designed in supervised manner. Deep learning models should be trained using representative data that match with the testing condition. Thus, a large training data need to be collected in a wide variety of settings in order to have good performance. Practically, the systems are trained using a large amount of synthetic data. First, the data are generated by artificially adding the noise at several SNRs to the target speech. Pairs of corrupted and clean data are then presented to the model. In this way, supervised deep learning models can automatically learn to perform SE. Previous researches have shown the effectiveness of this method to improve speech quality and intelligibility.

Previous works on SE applied their model either using cepstral (TF domain) or directly using raw signal (time domain). TF domain such as Mel-frequency cepstral coefficients (MFCC) [10], [18], [23], Mel-spectrogram [7], [16], [24] and magnitude spectrogram [8], [12], [21], [22], [27], can be obtained by applying STFT to the raw signal. These features are able to make the learning process easier and more generalized [38]. However, the limitation of utilizing spectral or cepstral representations is that it only uses magnitude spectrum and discards a potentially valuable phase spectrum. Hence, recent work is shifting from TF domain to time domain [14], [15], [17], [19], [20], [25]. This provides an opportunity to map the noisy speech to clean speech directly, which can retain the complete speech information including phase.

Works that focus on removing noise of the derived features are commonly regarded as feature enhancement. On the other hand, works implemented on waveform domain are regarded as speech enhancement. In this review, we treat both of them as enhancement models as they often share similar algorithms. Figure 1 shows an overview of SE system. The boxes with dash line (i.e. STFT and inverse-STFT) are necessary to derive speech features. In the case of SE, the training targets are divided into two groups: mapping, and masking-based methods. These groups came from the perspectives of supervised learning problem. It is referred to regression problem if the target is to directly map the noisy speech to clean speech. While it is referred to as classification problem if the target is to estimate a matrix, known as a mask. The mask is applied as filter to the output to produce the enhanced clean speech signal [39].

A. Mapping-Based Method

When using the mapping-based method, the target is to map a non-linear function F from the noisy speech $y(t)$ into the enhanced clean speech $x(t)$, as written in (5).

$$y(t) \xrightarrow{F} x(t). \quad (5)$$

Due to the fast-variation problems of using raw speech signal, mapping-based method is commonly applied to magnitude spectrogram of the speech signal (TF domain), which is created by applying STFT with time windowing responses of a filterbank. Then, the inverse operation of STFT is performed to reconstruct the

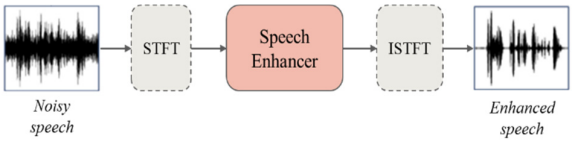

Figure 1. Speech Enhancement System Overview.

spectrogram back to the time domain signal using the phase information from the original noisy speech. However, most recent works have performed this mapping approach on the time domain [14], [17], [19], [20], [25]. The mapping-based method is known to be less sensitive to SNR variations [40], thus it is more useful for application with a wide range of SNRs.

Neural networks with mapping-based method are trained to reconstruct the target output from the observed input. The target output is obtained from the clean speech $x(t)$, while the observed input is extracted from the noisy speech $y(t)$. Specifically, the neural network learns F function by minimizing the mean square error (MSE) loss between the input spectrogram and its reconstructed input, as described in (6).

$$\mathcal{L}_{MSE} = \|Y - F(X)\|^2, \quad (6)$$

or mean absolute error (MAE) loss between the input of raw speech signal and its reconstructed input, as expressed in (7).

$$\mathcal{L}_{MAE} = \|y - F(x)\|, \quad (7)$$

where $\|\,.\,\|^2$ is the square loss, and $\|\,.\,\|$ is the absolute loss.

B. Masking-Based Method

When using the masking-based method, the target is to estimate a mask. This approach is implemented using magnitude spectrogram (TF domain), where the estimated mask is applied as filter to the input spectrogram. There are two most commonly used masks, ideal binary mask (IBM) and ideal ratio mask (IRM). In IBM, the frequency bins of the spectrogram that have high speech amplitude (local SNR is greater than a threshold R) are set to 1, while others with high noise intensity are set to 0, as described in (8).

$$IBM(k,l) = \begin{cases} 1, & if\ SNR(k,l) > R, \\ 0, & otherwise, \end{cases} \quad (8)$$

where k and l denote frequency bin and the time, respectively. R is a threshold that classifies the value into 1 or 0. The threshold value is chosen based on the experiment trials, however in previous work, it is stated that the best value is 5 dB lower than the SNR of noisy signal [41]. On the other hand, IRM, also known as soft masking, uses a probability value between 0 and 1, as written in (9).

$$IRM(k,l) = \left(\frac{X(k,l)^2}{X(k,l)^2 + N(k,l)^2}\right)^{\beta}, \quad (9)$$

where $X(k,l)^2$ and $N(k,l)^2$ denote the speech and noise energy in TF domain, respectively. β is a tunable parameter for mask scaling.

IV. Audio Corpora and Evaluation Metrics

A large training data is one of the key aspects that allow the implementation of deep learning technology. The choice of a dataset is critical to compare the effectiveness of the different approaches. In this section, we introduce a set of existing resources that are listed in Table 1 and Table 2. These datasets are widely used in the field of SE. Some data such as Aurora-2, Voice Bank corpus, and some noise datasets, are publicly available on the website.

Based on our survey, the most commonly used dataset is the noisy version of the Wall Street Journal (WSJ) [42], such as the CHiME series corpus. WSJ is originally a clean speech corpus. However, for the SE experiments, clean speech data is corrupted with several noise types. The CHiME series dataset involves not only additive noise but also reverberation [43]–[45]. Specifically, CHiME-2 consists of reverberant and noisy speech utterances. It was created by first convolving WSJ with a binaural room impulse response, then mixing it with binaural noise recorded in a living room at different SNRs. While CHiME-3 contains simulated and real noisy data recorded on different locations, e.g. on the bus, cafe, pedestrian area, and street junction, using a microphone array on a tablet. CHiME-4 is an extended version of CHiME-3. The CHiME series are commonly used when it is intended for ASR applications. In other works [7], [9], [26], they simply add noise to the WSJ corpus with non-speech environmental sounds or musical noises.

Other frequently used datasets are Voice Bank corpus [46], TIMIT corpus [47], and Aurora series [48]. Voice Bank and TIMIT originally contain clean speech data. These datasets are typically corrupted with some noise datasets such as NOISEX and DEMAND that include noises such as voice babble, factory noise, environmental noises such as in the bus, park, and cafe, and speech-shaped noise (SSN). While Aurora series are the noisy dataset that is developed by the European Telecommunications Standards Institute (ETSI). Aurora-2 is a digit recognition task and Aurora-4 is a large vocabulary continuous speech recognition task (LVCSR) which is based on WSJ corpus. Overall, these datasets were created for scenarios from small vocabularies such as digit recognition to large vocabularies. The datasets vary from simulated to real recordings data, and from additive noise to reverberation. These datasets are frequently adopted probably for two reasons. First, the amount of data is suitable for training deep learning. The datasets contain hours of speech recordings. Second, the datasets have become a benchmark for SE task. Moreover, the CHiME series are continuously updated through annual challenges.

The standard metrics to evaluate the performance of SE systems can be grouped into two, subjective and objective measures. Typical subjective measures are mean opinion score (MOS), the signal (SIG) distortion, and the background (BAK) noise intrusiveness. SIG, BAK, and MOS have a scale from 1 to 5 where the higher value is better. On the other hand, typical objective measures include segmental signal-to-noise ratio (segSNR), distance measures, source-to-distortion ratio (SDR), perceptual evaluation of speech quality (PESQ), and short-time objective intelligibility (STOI). PESQ is measured from -0.5 to 4.5, the higher value is the better speech quality. All of these measures are performed to assess the speech quality and intelligibility. In addition,

word error rate (WER) or word accuracy (WA) is a common metric used to particularly evaluate the performance of ASR systems.

V. DEEP LEARNING METHODS FOR SPEECH ENHANCEMENT

In this section, we review recent deep learning methods that are designed to address SE problem including DNN, DAE, RNN-LSTM, CNN, and GAN. The works using mapping-based methods are summarized in Table 1 and masking-based methods in Table 2. We summarize the advantages and disadvantages of the different methods in Table 3.

A. Based on DNN

Deep neural network (DNN), also known as feed-forward fully connected layer or multilayer perception (MLP) with many hidden layers, is one of the most used architectures for SE [8], [21], [26], [27]. The network is called fully-connected network because each node of the layer shares a connection to every node in the previous layer. Therefore, DNN has very large parameters. The work by Karjol *et al.* [27] proposed an enhancement scheme using multiple DNN based system involving *n* number of DNN, each is contributing to the final enhanced speech, and utilizing a gating network which provides the weights to combine the outputs of the *n* DNN. The model uses *n=4*, each of which is three layers deep. Average PESQ of 2.65 is achieved for seen noise and 2.19 for unseen noise on average SNR -5 to 10 dB on TIMIT corpus. While in [28], DNN masking-based method could achieve higher PESQ of 2.705. In [8], DNN was extended by incorporating a speech intelligibility metric into the loss function. The results showed an average under mismatched SNR with PESQ of 1.99. Another work by Bagchi *et al.* [21] also attempted to extend the model by combining mimic loss with the traditional criterion to train the speech enhancer. The mimic loss is defined as the mean square difference between spectral classifier outputs.

Although DNN has been successfully performed as a regression model for SE, the enhanced speech resulted from the model often deteriorates in low SNR conditions. Gao *et al.* [26] introduced a progressive learning framework for DNN-based SE. The model decomposed a conventional DNN into multiple stages with a different SNR in each stage, instead of mapping directly from noisy speech to clean speech. The model was trained on single-SNR and multi-SNR setting using WSJ corpus and tested on seen and unseen noises such as babble, factory, and destroyer engine. The results achieved PESQ score of 1.93 for single-SNR training and 1.82 for multi-SNR training on average (SNR -5 and 5 dB).

B. Based on DAE

In most of the works [18], [24], deep autoencoder is based on DNN whose outputs have the same dimensionality as the inputs. Deep autoencoder is often used for representation learning. The spectral mapping method with a denoising criterion, namely denoising autoencoder (DAE), was introduced by Lu *et al.* [24]. The model was extended to deep DAE by Feng *et al.* [18]. The model performed mapping from noisy to clean speech in the Mel-spectral domain. DAE is originally trained to convert the corrupted input y into a hidden representation z using the encoder as written in (10).

$$z = \sigma(Wy + b), \tag{10}$$

where σ is a non-linear activation function. W and b are weight matrix and bias vector respectively. The resulting representation z is subsequently converted back to a reconstructed input $\acute{y}$ using the decoder as described in (11).

$$\acute{y} = \sigma(W'z + b'), \tag{11}$$

where W' and b' are appropriately sized parameters of W and b, respectively. DAE is trained by minimizing the MSE loss between the input y and its reconstructed input $\acute{y}$. However, DAE network has a limitation of learning temporal information. Thus, it is typical to train the network using a small context window. Recently, there is a work that has employed DAE using convolutional layers to cope with the temporal problem [12].

C. Based on RNN-LSTM

When dealing with a sequence-based data such as speech signal, there is recurrent neural network (RNN) and long short-term memory (LSTM) which can handle context information. This network does not exploit the information only from the current input, but also from the previous hidden layer. Maas *et al.* [23] performed RNN to denoise corrupted input features, such as MFCC. The enhancement process using RNN has been shown to be more effective than using DNN to enhance noisy speech signal with SNR at various levels.

More recent work by Gao *et al.* [9] has been inspired by curriculum learning. They proposed a progressive learning framework with LSTM network to improve the performance of DNN-based speech in low SNR environment. Each of the target layers is designed to learn the transition speech with higher SNR and clean speech at the last layer. Furthermore, LSTM-RNN has been implemented to overcome the problem of highly non-stationary additive noise [11], reverberation [10], and multichannel noisy speech [34]. In [11], hand-crafted features like MFCC has been outperformed by using bottleneck features resulted from the bidirectional LSTM network (BN-BLSTM). The average WA using MFCC is 38.13%, while using BN-BLSTM is 43.55%. Speech processing systems have been largely improved with the help of LSTM-RNN. However, learning the RNN parameters is known to be difficult and takes a lot of resources.

D. Based on CNN

Convolutional neural network (CNN) has received a lot of attention from researchers in speech field [6], [12]–[17], [22], [33]. CNN has the ability to capture pattern in the neighboring frames by using a set of local connections. Figure 2 shows the architecture of a two-dimensional (2D) fully CNN on spectrogram features. It has also been reported to be more effective than standard feed-forward neural network [15] and more efficient than

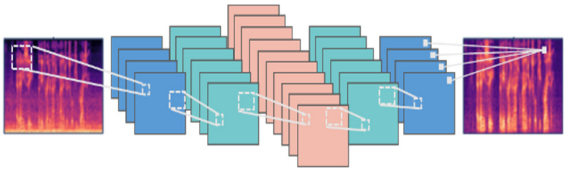

Figure 2. Architecture of A Fully CNN for Speech Enhancement.

RNN [12], [49]. Park & Lee [12] has demonstrated that CNN can achieve better performance with network 12 times smaller than RNN. CNN is capable of dealing with local temporal-spectral structures of speech, thus it is effective for separating the speech and noise elements of the noisy signals. CNN has shown its effectiveness for enhancing speech in both spectral and waveform domain.

Park & Lee [12] introduced spectral mapping using redundant convolutional encoder decoder (R-CED). This network uses spectrogram as the input that is viewed as an image of 2D representations with 1 channel. The encoder and decoder of this network consist of repetitions of convolutional layers. Unlike typical autoencoder network, R-CED encodes features of the spectrum into higher dimension and achieves the compression along the decoder. Skip connections are used to preserve information from the encoding stage to the decoding stage to improve the performance. This SE work is attempted for an embedded device such as the hearing aid. Thus, the objective is to find an efficient denoising algorithm which can be achieved by using CNN. This work showed that CNN achieved best STOI, PESQ, and SDR scores compared to DNN and RNN.

There are also recent works of CNN-based SE for ASR application [6], [16], [30]. Kinoshita *et al.* [30] employed speech denoising based on masking estimation using CNN. This work is motivated by the success of temporal convolution networks for speech separation (Conv-TasNet) [50]. They adapted the network architecture for denoising task, namely Denoising-TasNet, which is performed on both time and TF domain. This work also investigated multi-task loss that predicted two outputs, speech, and noise. The best performance is achieved by network in time domain with multi-task loss. Moreover, an extended version of CNN using residual network (ResNet) has been proposed [22]. An improved result can be achieved since the architecture of ResNet matches the SE task, which is to reconstruct the input signal by removing the residual noisy signal.

In addition, [14], [15], [17] proposed an end-to-end learning method for speech denoising by processing raw waveforms directly. In [17], the model was based on a novel network structure called WaveNet [49]. The network consists of a series of non-causal and dilated convolutional layers that learn in a supervised fashion via minimizing the regression loss. The dilation factors contribute to a receptive field that can significantly reduce the computational complexity. The overall results of this method show that CNN is better compared to conventional Wiener filter with 23% relative improvement of MOS quality.

E. Based on GAN

Generative adversarial network (GAN) has also received increasing attention to improve the model enhancement performance further. GAN consists of a generator network (G) and a discriminator network (D), as shown in Figure 3. GAN training is commonly built on convolutional layers [19], [20], [25] or fully connected layers [32]. Speech enhancement based on GAN training (SEGAN) was first introduced by Pascual *et al.* [25]. The generator network learns to map features of the noisy speech into the clean speech. The discriminator network, which acts as a binary classifier, subsequently assesses whether the samples come from the clean speech (real) or the enhanced speech (fake). In general, the two networks are trained in an adversarial manner and optimized by using (12).

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}[log\, D(x)] + \mathbb{E}_{\tilde{x}\sim P_{data}(\tilde{x})}[log(1 - D(G(y)))]. \quad (12)$$

Based on the discriminator results, the generator attempts to adapt the distribution to produce better outputs until the discriminator cannot distinguish the outputs whether those are real or fake. However, training GAN is difficult and unstable. Many other works tried to improve the performance of SEGAN [19], [20], [32]. Baby and Verhulst [19] implemented a relativistic loss function at the discriminator network with gradient penalty. This work showed that an improved discriminator could produce a cleaner speech. In addition, the work also utilized gradient penalty as stabilizer of the training. Phan *et al.* [20] proposed to use multiple generators instead of a single generator. The purpose was to gradually perform multi-stage enhancement mapping. The proposed method was better than SEGAN in terms of PESQ, CSIG, CBAK, COVL, and SSNR. While [7] attempted to modify SEGAN architecture for feature enhancement. Unlike SEGAN, the work is implementing on log-Mel features instead of waveforms because it is intended for ASR applications.

While GAN is gaining popularity in the mapping-based method, some works ([29], [32]) attempted to use it in the masking-based method. In [29], GAN is utilized to predict the masks. A regularized objective function of MSE is applied to improve the vanilla GAN. The results show an improvement of PESQ and STOI over a recent GAN-based speech enhancement.

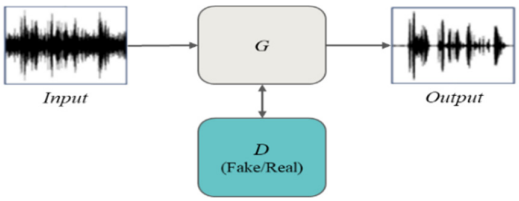

Figure 3. Architecture of A GAN for Speech Enhancement.

TABLE 1
SUMMARY OF DEEP LEARNING WITH MAPPING-BASED METHODS FOR SPEECH ENHANCEMENT

Method	Features	Refs.	Dataset	Evaluation Metrics	Results
DNN	(Log/power) mag.	[8]	IEEE corpus + NOISEX	PESQ, SDR, STOI	Results averaged under mismatched SNR (-3 to 3 dB) PESQ: 1.99, SDR: 11.35, STOI: 90.61%.
	(Log/power) mag.	[21]	CHiME-2	WER	Error rate of 14.7%.
	(Log/power) mag.	[27]	TIMIT + noises from Aurora dataset	PESQ, STOI, SegSNR	Average best PESQ achieves 2.65 for seen noise and 2.19 for unseen noise.
	LPS (log-power spectral)	[26]	WSJ + environmental and musical noises	PESQ, STOI, SSNR	PESQ 1.93 for single-SNR training, PESQ 1.82 for multi-SNR training, tested on unseen noise.
DAE	MFCC	[18]	CHiME-2	WER	Error rate of 34%.
	(Log) Mel	[24]	Japanese corpus + environmental noises	PESQ	Average PESQ on factory noise is 3.13 and on car noise is 4.08.
RNN-LSTM	LPS (log-power spectral)	[9]	WSJ + environmental and musical noises	SDR, STOI	Average results SDR: 9.46 and STOI: 0.86.
	MFCC	[10]	CHiME-2	WA, WER	Average accuracy is 85%.
	MFCC, bottleneck features (BN)	[11]	Buckeye (spontaneous speech) + CHiME noises	WA	Average WA using MFCC: 38.13%, BN-BLSTM: 43.55%.
	MFCC	[23]	Aurora-2	WER, MSE	Average error rate (SNR 0-20 dB) on seen noise is 10.28% and on unseen noise is 12.90%.
	(Log/power) mag., (log) Mel	[6]	CHiME-2 + environmental noises	WER	Best average error rate of 7.8% is achieved using magnitude features (accuracy of 92.2%).
	(Log/power) mag.	[12]	TIMIT + environmental noises	PESQ, STOI, SDR	CNN achieved best accuracy compared to DNN and RNN, PESQ: 2.34, STOI: 0.83, SDR: 8.62.
	(Log/power) mag.	[22]	CHiME-2	WER	Error rate of 9.3% is achieved by using ResNet + mimic loss.
	(Log) Mel	[16]	Aurora-4, AMI	WER	WER of 8.31% on Aurora-4.
	Raw signal, (log/power) mag.	[13]	TIMIT + NOISEX + SSN	PESQ, STOI, SI-SDR	Results show that Autoencoder CNN achieved better performance than SEGAN.
	Raw signal	[14]	Voice Bank + DEMAND	SNR, SIG, BAK, OVL	SNR:19.00, SIG: 3.86, BAK: 3.33, OVL: 3.22.
	Raw signal	[15]	TIMIT + environmental noises	PESQ, STOI	Best STOI is achieved by fully ConvNet, while best PESQ is achieved by DNN.
	Raw signal	[17]	Voice Bank + DEMAND	SIG, BAK, OVL, MOS	MOS of 3.60 is achieved. Overall results are better compared to Wiener filter.
GAN	(Log) Mel	[7]	WSJ + environmental and musical noises	WER	Error rate of 17.6%.
	Raw signal	[19]	Voice Bank + DEMAND	STOI, PESQ, SegSNR	STOI: 0.942, PESQ: 2.62, SegSNR: 17.68.
	Raw signal	[20]	Voice Bank + DEMAND	PESQ, CSIG, CBAK, COVL, SSNR, STOI	PESQ: 2.39, CSIG: 3.55, CBAK: 3.11, COVL: 2.93, SSNR: 8.72.
	Raw signal	[25]	Voice Bank + DEMAND	PESQ, CSIG, CBAK, COVL, SSNR	PESQ: 2.16, CSIG: 3.48, CBAK: 2.94, COVL: 2.80, SSNR: 7.73.

TABLE 2
SUMMARY OF DEEP LEARNING WITH MASKING-BASED METHODS FOR SPEECH ENHANCEMENT

Method	Features	Refs.	Dataset	Evaluation Metrics	Results
DNN	MFCC, LPS (log-power spectral)	[28]	TIMIT + environmental and musical noises	PESQ, STOI, SSNR	PESQ: 2.705, STOI: 0.871, SSNR: 5.194.
RNN-LSTM	(Log/power) mag.	[34]	CHiME-3	PESQ, STOI	PESQ achieves over 2.50 and STOI up to 0.9.
CNN	(Log/power) mag., raw signal	[30]	CHiME-4, Aurora-4	WER, SDR	Chime-4: WER 8.3% (real data), 10.8% (simulated), SDR: 14.24, Aurora-4: 6.3%.
	(Log/power) mag.	[31]	Grid corpus + CHiME-3 noises	PESQ, STOI	PESQ: 2.60 and STOI: 0.70 for seen noises, and 2.63 and 0.74 for unseen noises.
	(Log/power) mag., phase	[33]	Voice Bank + DEMAND	PESQ, CSIG, CBAK, COVL, SSNR	PESQ: 3.24, CSIG: 4.34, CBAK: 4.10, COVL: 3.81, SSNR: 16.85.
GAN	(Log/power) mag.	[29]	Voice Bank + DEMAND	PESQ, CSIG, CBAK, MOS, STOI	PESQ: 2.53, SIG: 3.80, BAK: 3.12, MOS: 3.14, STOI: 0.93.
	(Log/power) mag.	[32]	TIMIT + NOISEX + SSN	PESQ, STOI	GAN gives consistently better STOI score, but not much of an improvement in PESQ.

TABLE 3
SUMMARY OF ADVANTAGES AND DISADVANTAGES OF DEEP LEARNING APPROACHES

Method	Advantages	Disadvantages
DNN	Familiarity of the model architecture since the networks tend to be straightforward	DNN has a very large parameters because each node of the layer shares a connection to every node in the previous layer.
DAE	DAE performs dimensional reduction and the features of the bottleneck layer might be useful.	DNN-based DAE has a limitation of learning temporal information.
RNN-LSTM	- Best for dealing with a sequence-based data such as speech signal. - RNN-LSTM can handle context information.	Learning the RNN parameters is known to be difficult and takes a lot of resources.
CNN	- CNN has the ability to capture pattern in the neighboring frames of speech structures. - CNN is more effective than standard DNN and more efficient than RNN.	Lack of the ability to be invariant to the changes of the input data.
GAN	The combined networks in GAN can be powerful if it is trained properly.	The adversarial training tends to be difficult and unstable.

CONCLUSION

We provided an overview of several recent deep learning-based methods, such as DNN, DAE, RNN-LSTM, CNN, and GAN that are designed to address speech enhancement problem. We evaluated each method to show that there is considerable room for further research in the area. We concluded that, the above reviewed works revealed that a trend for speech enhancement task has gradually shifted from using cepstral or spectral representation (TF domain) to using waveform representation (time domain). This trend is mainly supported by the powerful capability of deep learning. The capability to directly extract representation from raw signal allow us to retain complete information of speech compared with the hand-crafted features like MFCC. Moreover, the availability of resources such as cloud computing and a massive collection of training data have also provided great support in the improvement of deep speech enhancement model.

Related to the network architecture, it starts shifting from the standard DNN to CNN. The main reason is that DNN-based algorithms are not able to efficiently learn the temporal information structure of speech signal. Moreover, CNN can significantly reduce the computational load compared with DNN and RNN-LSTM because it utilizes convolution and pooling operations as well as enforces parameter sharing. Furthermore, the network training strategy also starts shifting from a conventional method with a single network to GAN with an adversarial training that utilizes two networks. As shown in Table 1, acoustic features such as magnitude and Mel spectrogram are preferable than MFCC in recent neural network structure, since they are more suitable for deep learning method. With this rapid development of deep learning training, we are able to achieve further improvement for speech or feature enhancement model. While studies indicated that deep learning-based approaches performed really well, those still had poor adaptability of the system in real-world environment. Thus, exploring this aspect further is an interesting future direction.

REFERENCES

[1] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Wiley-IEEE Press, 2000.

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC press, 2013.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust.*, vol. 27, pp. 113–120, Apr. 1979.

[4] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *J. Acoust. Soc. Am.*, vol. 125, no. 1, 2009.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error-log-spectral amplitude estimator," *IEEE Trans. Acoust.*, vol. 33, Apr. 1985.

[6] P. Wang and D. L. Wang, "Enhanced spectral features for distortion-independent acoustic modeling," in *Proc. Annu. Conf. Speech Communication Association Interspeech 2019*, 2019, pp. 476–480.

[7] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *2018 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc.*, 2018.

[8] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually guided speech enhancement using deep neural networks," in *2018 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc.*, 2018, pp. 5074-5078.

[9] T. Gao, J. Du, L. R. Dai, and C. H. Lee, "Densely connected progressive learning for LSTM-Based speech enhancement," in *2018 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc.*, 2018, pp. 5054-5058.

[10] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich feature enhancement approach to the 2013 CHiME challenge using BLSTM recurrent neural networks," in *2nd Int. Workshop Machine Listening Multisource Environments,* 2013.

[11] M. Wollmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise," in *2013 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc.,* 2013, pp. 6822-6826.

[12] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Annu. Conf. Speech Communication Association Interspeech 2017*, 2017.

[13] A. Pandey and D. Wang, "A new framework for CNN-Based speech enhancement in the time domain," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, July. 2019.

[14] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *Proc. Annu. Conf. Speech Communication Association Interspeech 2019*, 2019.

[15] S. W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. 9th Asia-Pacific Signal and Information Processing Association Annu. Summit and Conf. 2017*, 2018.

[16] J. Rownicka, P. Bell, and S. Renals, "Multi-Scale octave convolutions for robust speech recognition," in *IEEE Int. Conf. Acoustics Speech and Signal Processing Proc.*, 2020.

[17] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech

denoising," in *2018 IEEE Int. Conf. IEEE Int. Conf. Acoustics Speech and Signal Processing Proc.*, 2018.
[18] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," *2014 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc.*, 2014.
[19] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *2019 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc.*, 2019.
[20] H. Phan *et al.*, "Improving GANs for speech enhancement," *IEEE Signal Process. Lett.*, vol. 27, 2020.
[21] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," in *2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc.*, 2018.
[22] P. Plantinga, D. Bagchi, and E. Fosler-Lussier, "An exploration of mimic architectures for residual network based spectral mapping," in *2018 IEEE Workshop Spoken Language Technology Proc.,* 2019.
[23] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *13th Annu. Conf. Speech Communication Association Interspeech 2012*, 2012.
[24] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Annu. Conf. Int. Speech Communication Association Interspeech 2013,* 2013.
[25] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Annu. Conf. Int. Speech Communication Association Interspeech 2017*, 2017.
[26] T. Gao, J. Du, L. R. Dai, and C. H. Lee, "SNR-based progressive learning of deep neural network for speech enhancement," in *Proc. Annu. Conf. Int. Speech Communication Association Interspeech 2016*, 2016.
[27] P. Karjol, M. A. Kumar, and P. K. Ghosh, "Speech enhancement using multiple deep neural networks," in *2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc.*, 2018.
[28] Y. Xu, J. Du, Z. Huang, L. R. Dai, and C. H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in *Proc. Annu. Conf. Int. Speech Communication Association Interspeech 2015*, 2015.
[29] M. H. Soni, N. Shah, and H. A. Patil, "Time-Frequency masking-based speech enhancement using generative adversarial network," in *2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc.*, 2018.
[30] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *2020 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc.*, 2020.
[31] Z. Xu, S. Elshamy, and T. Fingscheidt, "Using separate losses for speech and noise in mask-based speech enhancement," in *2020 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc.*, 2020.
[32] A. Pandey and D. Wang, "On adversarial training and loss functions for speech enhancement," in *2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc.,* 2018.
[33] H. S. Choi, J. H. Kim, J. Huh, A. Kim, J. W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *Proc. Int. Conf. Learning Representations 2019,* 2019.
[34] X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *2019 IEEE Workshop Applications Signal Processing to Audio and Acoustics*, 2019.
[35] Z. Zhang, J. Geiger, J. Pohjalainen, A. E. D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Trans. Intelligent Syst. Technol.*, vol. 9, no. 5, 2018.
[36] D. Michelsanti *et al.*, "An overview of deep-learning-based audio-visual speech enhancement and separation," *arXiv*, 2020.
[37] N. Das, S. Chakraborty, J. Chaki, N. Padhy, and N. Dey, "Fundamentals, present and future perspectives of speech enhancement," *Int. J. Speech Technol.*, 2020.
[38] L. Hertel, H. Phan, and A. Mertins, "Comparing time and frequency domain for audio event recognition using deep learning," in *2016 Int. Joint Conf. Neural Networks Proc.*, 2016.
[39] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "Mapping and masking targets comparison using different deep learning based speech enhancement architectures," in *2020 Int. Joint Conf. Neural Networks Proc.,* 2020, pp. 1–8.
[40] X. L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, May. 2016.
[41] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, Dec. 2014.
[42] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. Workshop Speech and Natural Language,* 1992.
[43] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second "chime" speech separation and recognition challenge: Datasets, tasks and baselines," in *2013 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc.,* 2013.
[44] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Comput. Speech Lang.*, vol. 46, 2017.
[45] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third "CHiME" speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop Automatic Speech Recognition and Understanding Proc.,* 2016.
[46] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 Int. Conf. Oriental COCOSDA 2013 Conf. Asian Spoken Language Research and Evaluation*, 2013.
[47] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon Tech. Rep. N, vol. 93, 1993.
[48] D. Pearce and H. G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Sixth Int. Conf. Spoken Language Processing Interspeech 2000*, Beijing, China, 2000.
[49] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," *arXiv*, 2016.
[50] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, Aug. 2019.