Artificial Intelligence Review (2023) 56:S3651–S3703
https://doi.org/10.1007/s10462-023-10612-2

Deep neural network techniques for monaural speech enhancement and separation: state of the art analysis

Peter Ochieng[1]

Accepted: 18 September 2023 / Published online: 25 October 2023

Abstract

Deep neural networks (DNN) techniques have become pervasive in domains such as natural language processing and computer vision. They have achieved great success in tasks such as machine translation and image generation. Due to their success, these data driven techniques have been applied in audio domain. More specifically, DNN models have been applied in speech enhancement and separation to perform speech denoising, dereverberation, speaker extraction and speaker separation. In this paper, we review the current DNN techniques being employed to achieve speech enhancement and separation. The review looks at the whole pipeline of speech enhancement and separation techniques from feature extraction, how DNN-based tools models both global and local features of speech, model training (supervised and unsupervised) to how they address label ambiguity problem. The review also covers the use of domain adaptation techniques and pre-trained models to boost speech enhancement process. By this, we hope to provide an all inclusive reference of all the state of art DNN based techniques being applied in the domain of speech separation and enhancement. We further discuss future research directions. This survey can be used by both academic researchers and industry practitioners working in speech separation and enhancement domain.

Keywords Dereverberation · Denoising · Speaker separation · Deep neural network · Time-domain · Speech enhancement · Speech extraction

1 Introduction

Techniques for monaural speech intelligibility improvement can be categorised either as speech enhancement or separation. Speech enhancement involves isolating a target speech either from noise (Bando et al. 2018) or a mixed speech (Xiao et al. 2019). Speech enhancement involves tasks such as dereverberation, denoising and speaker extraction. Speaker separation on the other hand seeks to estimate independent speeches composed in a mixed speech (Wang and Wang 2013). Speech enhancement and separation have applications in multiple domains such as automatic speech recognition, mobile

✉ Peter Ochieng
po304@cam.ac.uk

[1] Department, University of Cambridge, 15 JJ Thomson Ave, Cambridge CB3 0FD, UK

speech communication and designing of hearing aids (Wang et al. 2014). Initial research on speech enhancement and separation exploited techniques such as non-negative matrix factorization (NMF) (Schmidt and Olsson 2006; Wang and Sha 2014; Virtanen and Cemgil 2009) probabilistic models (Virtanen 2006) and computational auditory scene analysis (CASA) (Shao and Wang 2006). However, these techniques are tailored for closed-set speakers (i.e., do not work well with mixtures with unknown speakers) which significantly restricts their applicability in real environments. Due to the recent success of deep learning models in different domains such natural language processing and computer vision, these data driven techniques have been introduced to process audio dataset. In particular, DNN models have become popular in speech enhancement and separation and have achieved great performance in terms of boosting speech intelligibility and their ability to enhance speech with unknown speakers (Luo and Mesgarani 2019; Subakan et al. 2021). In order to be effective in speech enhancement and separation, DNN models must extract important features of speech, maintain order of audio frames, exploit both local and global contextual information to achieve coherent separation of speech data. This necessitates that DNN models should include techniques tailored to meet these requirements. Discussion of these techniques is the core subject of this review. Further, in computer vision and text domain, large pre-trained models are used to extract universal representations that are beneficial to downstream tasks. The review discusses the impact of pre-trained models to the speech enhancement and separation domain. It also discusses DNN techniques being adopted by speech enhancement and separation tools to reduce computation complexity to enable them work in low latency and resource constrained environments. The review therefore focuses on the whole pipeline of DNN application to speech enhancement and separation, i.e., from feature extraction, model implementation, training and evaluation. Our goal is to uncover the dominant techniques at each level of DNN implementation. In each section, we highlight key emerging features and challenges that exist. A recent review (Wang and Chen 2018) only looked at supervised techniques of performing speech separation and in this review, we discuss both supervised and unsupervised methods. Moreover, with the fast-growing field of deep learning, new techniques have emerged that necessitates a new look into how these techniques have been implemented in speech enhancement and separation. The review is constrained to discussing how DNN techniques are being applied to monaural speech enhancement and so we do not focus on multi-channel speech separation (which has been covered in Gannot et al. (2017)).

The paper first explains the types of speech enhancement and separation (Sect. 2) by highlighting their key elements and the tools that focus on each type. It discusses the key speech features that are being used by speech enhancement and separation tools in Sect. 3. This section looks at how the features are derived and how they are used to train the DNN models in supervised learning technique. Section 5 discusses the techniques the tools use to model long dependencies that exist in speech. The paper discusses model size compression techniques in Sect. 6. In Sect. 7, the paper discusses some of the popular objective functions used in speech enhancement and separation. Section 8 discusses how some tools are implementing unsupervised techniques to achieve speech enhancement and separation. Section 9 discusses how the speech separation and enhancement tools are being adapted to the target environment. In Sect. 10 the paper looks at how pre-trained models are being utilized in the speech enhancement and separation pipeline. Finally, Sect. 11 looks at future direction. Figure 1 gives an overall organization and topics covered by the paper.

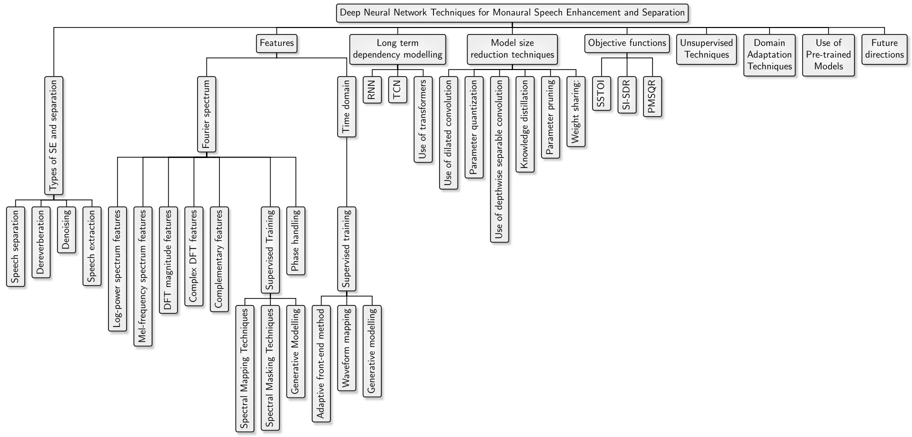


Fig. 1 Overall structure of the topics covered by this review

2 Types of speech separation and enhancement

2.1 Speech separation

Scenarios arise where more than one target speech signals are composed in a given speech mixture and the goal is to isolate each independent speech composed in a mixture. This problem is known as speech separation. For a mixture that is composed of C independent speech signals $x_c(n)$ with $c = 1, \dots, C$, a recording $y(n)$ composed of the C speech signals can be represented as:

$$y(t) = \sum_{c=1}^{C} x_c(t) \tag{1}$$

Here, t indexes time. The goal of speech separation is to estimate each independent x_c speech signal composed in $y(n)$. Separating speech from another speech is a daunting task by the virtue that all speakers belong to the same class and share similar characteristics (Hershey et al. 2016). Some models such as Wang et al. (2016, 2017) lessen this by performing speech separation on a mixed speech signal based on gender voices present. They exploit the fact that there is large discrepancy between male and female voices in terms of vocal track, fundamental frequency contour, timing, rhythm, dynamic range etc. This results in a large spectral distance between male and female speakers in most cases to facilitate a good gender segregation. For speech separation that the mixture involves speakers of the same gender, the separation task is much difficult since the pitch of the voice is in the same range (Hershey et al. 2016). Most speech separation tools that solve this task such as Zeghidour and Grangier (2021), Huang et al. (2011), Weng et al. (2015), Isik et al. (2016), Hershey et al. (2016) and Luo and Mesgarani (2019) cast the problem as a multi-class regression. In that case, training a DNN model involves comparing its output to a source speaker. DNN models always output a dimension for each target class and when multiple sources of the same type exist, the system needs to select arbitrarily which output dimension to map to each output and this raises a permutation problem (permutation ambiguity) (Hershey et al. 2016). Taking a case of a two speaker separation, if the model estimates $\hat{a}_1$ and $\hat{a}_2$ as the magnitude of the reference speech magnitudes a_1 and a_2 respectively, it is unclear the order in which the model will output the estimates i.e. the order of output can either be $\{\hat{a}_1, \hat{a}_2\}$ or $\{\hat{a}_2, \hat{a}_1\}$. A naive approach shown in Fig. 2 (Kolbæk et al. 2017a) is to present the reference speech magnitudes in a fixed order and hope that it is the same order in which the system will output its estimation.

In case of a mismatch, the loss computation will be based on the wrong comparison resulting in low quality of separated speeches. Systems that perform speaker separation have an extra burden of designing mechanisms that are geared towards handling the permutation problem. Several strategies are being implemented by speech separation tools to tackle permutation problem. In Weng et al. (2015), a number of DNN techniques are implemented that estimates two clean speeches contained in a two-talker mixed speech. They employ supervised training to train DNN models to discriminate the two speeches based on average energy, pitch and instantaneous energy of a frame. Work in Yu et al. (2017) and Kolbæk et al. (2017a) introduce permutation invariant training (PIT) technique of computing permutation loss such that permutations of reference labels are presented as a set to be compared with the output of the system. The permutation with the lowest loss is adopted as the correct order. For a a two-speaker separation system introduced earlier, the reference sources permutation will be $\{a_1, a_2\}$ and $\{a_2, a_1\}$ such that the possible permutation losses are computed as:

$$\begin{aligned} loss_1 &= D([a_1, a_2], [\hat{a}_1, \hat{a}_2]) = D[a_1, \hat{a}_1] + D[a_2, \hat{a}_2] \\ loss_2 &= D([a_2, a_1], [\hat{a}_1, \hat{a}_2]) = D[a_2, \hat{a}_1] + D[a_1, \hat{a}_2] \end{aligned}$$

The one that returns the lowest loss between the two is selected as the permutation loss to be minimized (see Fig. 3). For an S speaker separation system a total of $S!$ permutations are generated.

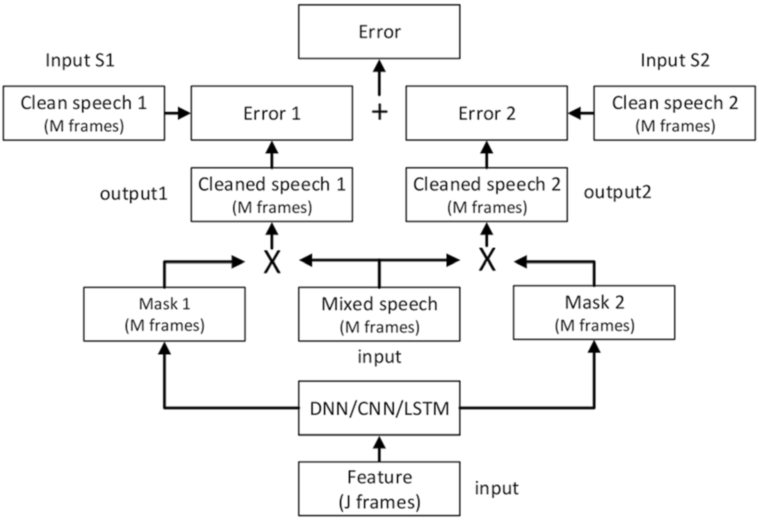


Fig. 2 Naive approach of solving label matching problem for a two-talker speech separation model. Here, the reference speech S1 and S2 are presented in a fixed order with expectation that the separated speeches will also be output in the same order. In case of mismatch the training will be optimizing wrong error values

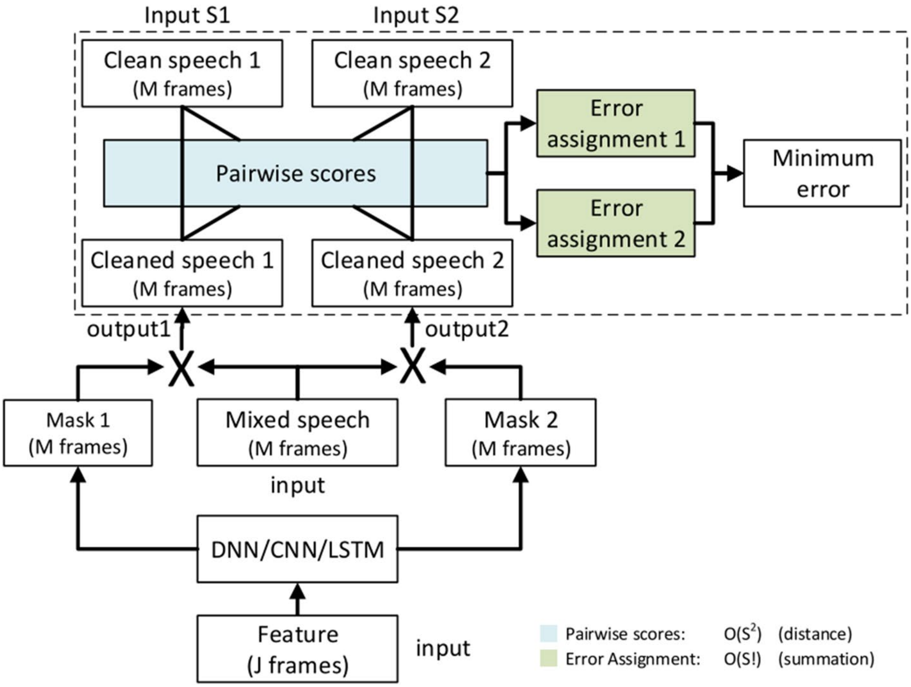


Fig. 3 Permutation invariant training implementation of label matching for a two-talker speech separation model. Here, a given reference speech is compared to all possible outputs. For instance, reference S1 is compared to both estimated clean speech 1 and 2

For a system that performs S speaker separation and S is high (e.g. 10), implementation of PIT which has a computation complexity of $O(S!)$ is computationally expensive (Tachibana 2021; Dovrat et al. 2021). Due to this, Dovrat et al. (2021) casts the permutation problem as a linear sum problem where Hungarian algorithm is exploited to find the permutation which minimizes the loss at computation complexity of $O(S^3)$. Work in Tachibana (2021) proposes SinkPIT loss which is based on Sinkhorn's matrix balancing algorithm. They utilize the loss to reduce the complexity of PIT loss from $O(C!)$ to $O(kC^2)$. Work in Zeghidour and Grangier (2021) employs minimum loss permutation computation at each time step t. The best permutation (argmin) at each time-step is exploited to re-order the embedding vectors to be consistent with the training labels. To evade the permutation problem, they train two separate DNN models for each of the two speakers to be identified. Another prominent technique of handling permutation problem is to employ a DNN clustering technique (Hershey et al. 2016; Byun and Shin 2021; Isik et al. 2016; Qin et al. 2020; Lee et al. 2022) to identify the multiple speakers present in a mixed speech signal. The DNN f_θ accepts as its input the whole spectrogram X and generates a D dimension embedding vector V i.e., $V = f_\theta(X) \in R^{N\times D}$. Here, the embedding V learns the features of the spectrogram X and is considered a permutation- and cardinality-independent encoding of the network's estimate of the signal partition. For the network f_θ to be learn how to generate an embedding vector V given the input X, it is trained to minimize the cost function.

$$C_Y(V) = ||VV^T - YY^T||_F^2 = \sum_{ij}(< v_i, v_j > - < y_i, y_j >)^2 \tag{2}$$

Here, $Y = \{y_{i,c}\}$ represents the target partition that maps the spectrogram S_i to each of the C clusters such that $y_{i,c=1}$ if element i is in cluster c . YY^T is taken here as a binary affinity matrix that represents the cluster assignment in a partition-independent way. The goal in Eq. (2) is to minimise the distance between the network estimated affinity matrix VV^T and the true affinity matrix YY^T. The minimization is done over the training examples. $||A||_F^2$ is the squared Frobenius norm. Once V has been established, its rows are clustered into partitions that will represent the binary masks. To cluster the rows v_i of V, K-means clustering algorithm is used. The resulting clusters of V are then used as binary masks to separate the sources by applying the masks on mixed spectrogram X. The separated sources are then reconstructed using inverse Short-time Fourier transform (STFT). Even though PIT is popular in speech separation models, it is unable to handle the output dimension mismatch problem where there is a mismatch on the number of speakers between training and inference (Jiang and Duan 2020). For example, training a speech separation model on n speaker mixtures but testing it on $t \neq n$ speaker mixtures. The PIT-based methods cannot directly deal with this problem due to their fixed output dimension. Most speech separation models such as Nachmani et al. (2020), Kolbæk et al. (2017a), Liu and Wang (2019), Luo and Mesgarani (2020) deal with the problem by setting a maximum number of sources C that the model should output from any given mixture. If an inference mixture has K sources, where $C > K$, $C - K$ outputs are invalid, and the model needs to have techniques to handle the invalid sources. In case of invalid sources, some models such as Liu and Wang (2019), Nachmani et al. (2020), Kolbæk et al. (2017a) design the model to output silences for invalid sources while (Luo and Mesgarani 2020) outputs the mixture itself which are then discarded by comparing the energy level of the outputs relative to the mixture. The challenge with models that output silences for invalid sources is that they rely on a pre-defined energy threshold, which may be problematic if the mixture also has a very low energy (Luo and Mesgarani 2020). Some models handle the output dimension mismatch problem by generating a single speech in each iteration and subtracting it from the mixture until no speech is left (Shi et al. 2018; Kinoshita et al. 2018; Takahashi et al. 2019; Neumann et al. 2019; von Neumann et al. 2020). The iterative technique despite being trained with a mixture with low number of sources can generalize to mixtures with a higher number of sources (Takahashi et al. 2019). It however faces criticism that setting iteration termination criteria is difficult and the separation performance decreases in later iterations due degradations introduced in prior iterations (Takahashi et al. 2019). Other speech separation models include Luo et al. (2018), Yul et al. (2017), Chang et al. (2020), Liu and Wang (2019), Weng et al. (2015), Isik et al. (2016), Wang et al. (2018a).

2.2 Speaker extraction

Some speech enhancement DNN models have been developed where in a mixed speech such as an Eq. (1), they design methods to extract a single target speech. These models focus only on a single target speech x_{target} and treat all other speeches as interfering signals, therefore they modify Eq. (1) as shown in (3).

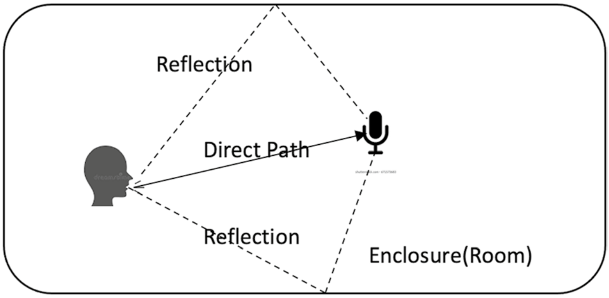
$$y(t) = \sum_{c=1}^{C} x_c(t) = x_{target}(t) + \sum_{c \neq target}^{C} x_c(t) \tag{3}$$

where $x_{target}(t)$ is the target speech at time t and $x_c(t)$ is the interfering signal. By focusing on only a single target speech, the permutation ambiguity problem is avoided. They formulate the speech extraction task into a binary classification problem, where the positive class is the target speech, and the negative class is formed by the combination of all other speakers. A popular technique of speaker extraction is to give as input to the DNN models additional speaker dependent information that can be used to isolate a target speaker (Veselỳ et al. 2016). Speaker dependent information can be injected into the DNN models by either concatenating speaker dependent auxiliary clues with the input features or adapting part of the DNN model parameters for each speaker (Wang et al. 2018b). This addition information about a speaker injects a bias that is necessary to differentiate the target speaker from the rest in the mixture (Xiao et al. 2019). Several auxiliary clues have been exploited by DNN models which include pre-recorded enrolment utterances of the target speaker (Wang et al. 2018b, c; Xiao et al. 2019; Ji et al. 2020; Zhang et al. 2021a; Delcroix et al. 2018), electroglottographs (EGGs) of the target speaker (Chen et al. 2023a) and i-vectors extracted at speaker level (Miao et al. 2015; Senior and Lopez-Moreno 2014). Tool in Ochiai et al. (2014) adapt parameters for each speaker by allocating a speaker dependent module to a selected intermediate layer of DNN. Speech extraction tool in Chen et al. (2017) does not use auxiliary clues of the target speaker but design attractor points that are compared with the mixed speech embeddings to generate the mask used to extract the target speech.

2.3 Dereverberation

This is a speech enhancement technique that seeks to eliminate the effect of reverberation contained in speech. When speech is captured in an enclosed space by a microphone that is at distance d from the talker, the observed signal consists of a superposition of many delayed and attenuated copies of the speech resulting from reflections of the enclosed space walls and existing objects within the space (see Fig. 4) (Naylor 2010). The signal received by the microphone consists of direct sound, reflections that arrive shortly after direct sound (within approximately 50 ms) i.e., early reverberation and reflections that arrive after early reverberation i.e., late reverberation (Williamson and Wang 2017a). Normally, early reverberation does not affect speech intelligibility much (Arweiler and Buchholz 2011) and much of perceptual degradation of speech is attributed to late reverberation. Speech

Fig. 4 How Reverberation happens. It shows how reverberation is composed direct speech and reflected speech which arrive late



degradation due to reverberation can be attributed to two types of masking (Nábělek et al. 1989), overlap masking- where the energy of a preceding phoneme overlaps with the one following or self-masking-where internal temporal which refers to the time and frequency alterations of an individual phoneme. Reverberation therefore can be viewed as the convolution of the direct sound and the room impulse response (RIR). A reverberant speech can be formally represented according to Eq. (4):

$$y(t) = h(t) * s(t) \tag{4}$$

Here, $*$ represents convolution, $s(t)$ is the clean anechoic speech. $h(t)$ represents room impulse response i.e., direct speech $h_d(t)$, early reverberation $h_e(t)$and late reverberation $h_l(t)$. Hence $h(t)$ can be represented as

$$h(t) = h_d(t) + h_e(t) + h_l(t) \tag{5}$$

Using the distributive property of convolution (Oppenheim 1999) Eq. (4) becomes:

$$y(t) = h_d(t) * s(t) + h_e(t) * s(t) + h_l(t) * s(t) \tag{6}$$

The goal of dereverberation is therefore to establish $s(t)$ from $y(t)$. Hence it can be viewed as a deconvolution between the speech signal and RIR (Zhou et al. 2022). Dereverberation is considered a more challenging task than denoising for a number of reasons. First, it is difficult to pinpoint direct speech from its copies especially when the reverberation is strong. Secondly, the key underlying assumption of sparsity and orthogonality of speech representations in the feature domain that is commonly used in monaural mask-based speech separation does not hold for speech under reverberation (Cord-Landwehr et al. 2021). Due to these unique features of reverberation, most tools designed for denoising, or speaker separation are ill poised to perform dereverberation (Cord-Landwehr et al. 2021). The DNN tools for speaker separation and denoising mostly make assumption that they are working on reverberation free speech hence do not make special consideration for eliminating reverberation (with exception of a few such as Su et al. (2020), Choi et al. (2020)). For instance, in Cord-Landwehr et al. (2021) they demonstrate that SepFormer (Subakan et al. 2021) performance can significantly improve by making adjustments to include techniques that handle reverberation. Several deep learning models have been designed with a goal to estimate clean speech from a reverberant one. Similar to speech denoising and speech separation, DNN models performing dereverberation exploit these models to fit a nonlinear function to map features of a reverberant speech to features of clean anechoic speech either directly (Wang et al. 2019; Han et al. 2015; Jiang et al. 2014; Gamper and Tashev 2018; Zhao et al. 2020; Ueda et al. 2016) or by use of mask (Huang et al. 2011; Williamson and Wang 2017a, b; Jin and Wang 2009; Jiang et al. 2014; Williamson and Wang 2017a; Jin and Wang 2009). Therefore, one way of categorising the existing dereverberation DNN tools is based on the type of target (spectrogram or ratio mask) they employ. Another way in which dereverberation tools can be categorised is based on whether a tool performs general dereverberation (i.e suppress $h(t)$ see Eq. (4)) or focus only on eliminating late reverberation ($h_l * s(t)$ see equation 6). Tools such as León and Tobar (2021), Zhou et al. (2022), Défossez et al. (2020), Isik et al. (2020), Li et al. (2021a), Valin et al. (2022) explore elimination of late reverberation. This is because early reverberation does not affect speech intelligibility much. Finally, the DNN dereverberation tools can be categorised based on the type of training technique used (supervised or unsupervised). Tools such as Han et al. (2015) and Huang et al. (2011) perform speech dereverberation by implementing supervised training where the DNN model is trained to directly estimate features clean speech when given features from a reverberant speech.

$$D(k,f) = M(k,f) \times Y(k,f) \tag{7}$$

Here $D(k, f)$, $M(k, f)$and $Y(k, f)$ are the Short-time Fourier transform(STFT) of the clean speech, the ideal ratio mask, and the reverberant speech at time frame k and frequency channel f respectively. Work in Fu et al. (2022) exploits conditional GAN to perform unsupervised dereverberation of a reverberant speech.

Dereverberation in discrete Fourier transform (DTF) magnitude domain When dereverberation is to be performed in DFT magnitude domain (see Sect. 3.1), a DFT has to be applied to Eq. (4) such that,

$$DFT(y(t)) = Y(t,f) = H(t,f) \times S(t,f) \tag{8}$$

the assumption in Eq. (8) is that the convolution of the clean signal $s(t)$ with RIR $h(t)$ corresponds to the multiplications of their Fourier transform in the T-F domain. However, this is only true if the extent of $H(t, f)$ is smaller than the analysis window (Cord-Landwehr et al. 2021). Therefore, when performing dereverberation in the TF domain the selection of the window is crucial on the performance of the DNN model (Cord-Landwehr et al. 2021).

Target selection in dereverberation In dereverberation training, most tools use direct speech as the target. This therefore means that the estimated speech will have to be compared with the direct path speech via a selected loss function. This has the potential of resulting in large prediction errors which can cause speech distortion (Zhou et al. 2022). Due to this, recent work Valin et al. (2022) proposes the use of a target that has early reverberation. By doing this, they suggest it will improve the quality of enhanced speech. In fact, experiments in Valin et al. (2022) demonstrate that allowing early reverberation in the target speech improves the quality of enhanced speech.

2.4 Speech denoising

This is a speech enhancement technique of separating background noise from the target speech. Formally, the noisy speech is represented as:

$$y(t) = s(t) + n(t) \tag{9}$$

where $y(t)$ is the noisy speech signal at time t, $s(t)$ is the target speech signal and $n(t)$ is the noise signal at time t. Speech denoising seeks to isolate a single target speech from noise. Hence data driven DNN models are optimized to predict s_t from y_t. Since speech denoising has only a single target speech it does not suffer from global permutation ambiguity problem. Some DNN tools that perform speech denoising include Leglaive et al. (2018, 2019, 2020), Bando et al. (2018), Kolbæk et al. (2017b), Lu et al. (2021, 2022), Fu et al. (2016, 2018a); Gao et al. (2016).

3 Speech separation and enhancement features

Speech enhancement and separation tools' input features can be categorised into two:

1. Fourier spectrum features.
2. Time domain features.

3.1 Fourier spectrum features

Speech enhancement and separation tools that use these features do not work directly on the raw signal (i.e., signal in the time domain) rather they incorporate the discrete Fourier transform (DFT) in their signal processing pipeline mostly as the first step to transform a time domain signal into frequency domain. These models recognise that speech signals are highly non-stationary, and their features vary in both time and frequency. Therefore, extracting their time-frequency features using DFT will better capture the representation of speech signal (Portnoff 1980). To demonstrate the DFT process we exploit a noisy speech signal shown in Eq. (10). The same process can be applied in speech separation.

$$y(t) = x(t) + n(t) \tag{10}$$

where $x(t)$ and $n(t)$ represent discrete clean speech and noise respectively at time t. Since speech is assumed to be statistically static for a short period of time, it is analysed frame-wise using DFT as shown in Eq. (11) (Portnoff 1980; Allen 1982; Allen and Rabiner 1977).

$$Y[t,k] = \sum_{m=\infty}^{-\infty} y(m)w(t-m)\exp^{-j2\pi km/L} \tag{11}$$

Here, k represents the index (frequency bin) of the discrete frequency, L is the length of the frequency analysis and $w(n)$ is the analysis window. In speech analysis, the Hamming window is mostly used as $w(n)$ (Paliwal et al. 2011). Once the DFT has been applied to the signal $y(t)$, it is transformed into time-frequency domain represented as:

$$Y[t,k] = X[t,k] + N[t,k] \tag{12}$$

$Y[t, k]$, $X[t, k]$ and $N[t, k]$ are the DFT representations of the noisy speech, clean speech and noise respectively. Each term in Eq. (12) can be expressed in terms of DFT magnitude and phase spectrum. For example, the polar form (including magnitude and phase) of the noisy signal $Y[t, k]$ is:

$$Y[t,k] = |Y[t,k]| \exp^{j\angle Y[t,k]} \tag{13}$$

$|Y[t, k]|$ and $\angle Y[t,k]$ are the magnitude and phase spectra of $Y[t, k]$ respectively. Equation (13) can be written in Cartesian coordinates as shown in Eq. (14).

$$Y[t,k] = |Y[t,k]|(\cos\theta + i\sin\theta) = |Y[t,k]|\cos\theta + i|Y[t,k]|\sin\theta \tag{14}$$

Both phase and the magnitude are computed from the real and the imaginary part of $Y[t, k]$ i.e.

$$|Y[t,k]| = \sqrt{\mathbb{R}(Y[t,k])^2 + \mathfrak{I}(Y[t,k]^2)} \tag{15}$$

$$\angle Y[t,k] = \tan^{-1}\frac{\mathfrak{I}(Y[t,k])}{\mathbb{R}(Y[t,k])} \tag{16}$$

All models that work with the Fourier spectrum features either use the DFT representations directly as the input of the model or further modify the DFT features. The features based on Fourier spectrum include:

1. Log-power spectrum features.
2. Mel-frequency spectrum features.
3. DFT magnitude features.
4. Complex DFT features.
5. Complementary features.

DFT magnitude features These are features where the mixed raw waveform $y(t)$ is first converted into time-frequency (TF) representation (spectrogram) using DFT (specifically, short-time Fourier transform (STFT) (Eq. 11) (Natsiou and O'Leary 2021). The magnitude of the time-frequency representation (Eq. 13) acts as the input to a deep neural network (DNN) model for speech separation. The DNN model is then trained to learn how to separate the TF-bins such that those that comprise each source speech are grouped together. DNN speech enhancements and separation models that exploit DFT features include systems such as Nossier et al. (2020a) Fu et al. (2018a), Grais and Plumbley (2018), Fu et al. (2019), Jansson et al. (2017), Kim and Smaragdis (2015). The use of DFT magnitude as features work with high frequency resolution hence necessitating the use of larger time window which is typically more that 32 ms (Isik et al. 2016; Kolbæk et al. 2017a) for speech and more than 90 ms for music separation (Luo et al. 2017). Due to this, these models must handle increased computational complexity (Baby et al. 2014). This has motivated other speech separation models to work with lower dimensional features as compared to those of DFT magnitude.

DFT complex features Unlike the DFT magnitude features that only use the magnitude of T-F representations, tools that use DFT complex features include both the magnitude and the phase of the noisy (mixed) speech signal in the estimation of the enhanced or separated speech. Therefore, each T-F unit of a complex features is a complex number with a real and imaginary component (see Eq. 13). The magnitude and phase of a signal is computed according to Eqs. (15) and (16) respectively. Tools that use DFT complex features include Fu et al. (2017), Williamson and Wang (2017a), Kothapally and Hansen (2022a, b).

Mel-frequency cepstral coefficients (MFCC) features Given the mixed speech signal such as in Eq. (10), to extract Mel frequency cepstral features, the following steps are executed:

1. Perform DFT of the input noisy signal $DFT(y(t)) = Y[t,k] = X[t,k] + N[t,k]$
2. Given the DFT features $Y[n, k]$ of the input signal, a filterbank with M filters i.e. a $1 \leq m \leq M$ is defined where m is a triangular filter given by:

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \frac{(2(k-f[k-m])}{f[m+1]-f[m-1])(f[m]-f[m-1])} & f[m-1] \leq k \leq f[m] \\ \frac{2(f[m+1]-1)}{(f[m+1]-f[m-1])(f[m+1]-f[m])} & f[m] \leq k \leq f[m+1] \\ 0 & k > f[m+1] \end{cases} \tag{17}$$

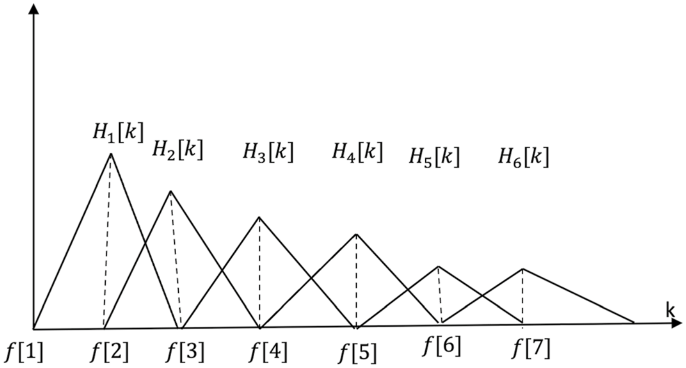


Fig. 5 Triangular filters used in the computation of the Mel-cepstrum using Eq. (18)

The filters are used to compute the average spectrum around centre frequencies with increasing bandwidths as shown in Fig. 5. Here, $f[m]$ are uniformly spaced boundary points in the Mel-scale which is computed according to Eq. (18). The Mel-scale B is given by Eq. (19) and B^{-1} which is its inverse is computed as shown in Eq. (20).

$$f[m] = \frac{N}{F_s} B^{-1}\left(B(f) + m\frac{B(f_h) - B(f_l)}{M+1} \right) \tag{18}$$

F_s is the sampling frequency, f_l and f_h represent the lowest and the highest frequencies of the filter bank in Hz. N is the size of DFT and M is the number of filters.

$$B(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \tag{19}$$

$$B^{-1}(b) = 700\left(\exp\left(\frac{b}{1125}\right) - 1\right) \tag{20}$$

3. Scale the magnitude spectrum $|Y[t, k]|$ of the noisy signal in both frequency and magnitude using mel-filter bank $H(k, m)$ and then take the logarithm of the scaled frequency.

$$X'(m) = \log(\sum_{k=0}^{N-1} |Y[t, k]|^2 H_m[k]) \tag{21}$$

for $m = 0, \dots, M$where M is the number of filter banks.

4. Compute the Mel frequency by computing the discrete cosine transform of the m filter outputs as shown in Eq. (22).

$$c[n] = \sum_{m=0}^{m-1} X'(m) \cos\left(\pi n(m + 1/2)/M\right) \tag{22}$$

where $0 \le n < M$

The motivation for working with MFCC is that it results in reduced resolution space as compared to DFT features. Fewer parameters are easier to learn and may generalise better to unseen speakers and noise (Baby et al. 2014). The challenge however with working on a reduced resolution such as MFCC is that the DNN estimated features must be extrapolated to the DFT feature space. Due to working on a reduced resolution, MFCC degree-of-freedom will be restricted by the dimensionality of the reduced resolution feature space which is much less than that of the DFT space. The low-rank approximation generates a sub-optimal Wiener filter which cannot account for all the added noise content and yields reduced SDR (Baby et al. 2014). MFCC features have been exploited in tools such as Liu et al. (2022), Ueda et al. (2016), Du et al. (2020), Fu et al. (2018a), Weninger et al. (2014), Donahue et al. (2018).

Log-power spectra features To compute these features, a short-time Fourier analysis is applied to the raw signal computing the DFT of each overlapping waveform (see Eq. 11). The log-power spectra are then computed from the output of the DFT. Consider a noisy speech signal in the time-frequency domain i.e., where DFT has been applied to the signal (see Eq. 12). From Eq. (14), the power spectrum of the noisy signal can be represented as in Eq. (23).

$$|Y[k]|^2 = |X[k]|^2 + N[k]|^2 = |X[k]|^2 + |N[k]|^2 + 2|X[k]||N[k]|\cos\theta \tag{23}$$

Here, θ represents the angle between the two complex variables $|X[k]|$ and $|N[k]|$. Most models that exploit log-power spectra features ignore the last term (assume the value to be zero) and employ equation 24.

$$Y^l = log(|Y[k]|^2) \tag{24}$$

Figure 6 Du and Huo (2008) summarises the process of log-power feature extraction.

Examples of models that use Log-power spectra features include (Fu et al. (2017), Du and Huo (2008), Xu et al. (2015), Du et al. (2014)).

Complementary features Since different features strongly capture different acoustic features which characterise different properties of the speech signal, some DNN models exploit a combination of the features to perform speech separation. This is based on works such as Garau and Renals (2008) and Zolnay et al. (2007) which demonstrated that complementary features significantly improve performance in speech recognition. The complementary features used in Zolnay et al. (2007), Wang et al. (2013), Williamson et al. (2016) include perceptual linear prediction, amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear product (RASTA-PLP), Gammatone frequency cepstral coefficient, MFCC, pitch-based features. The complementary features are combined by concatenation. Research in Williamson et al. (2016) reports that the use of complementary features registered better results as compared to those of DFT magnitude. The challenge with using complementary features is how to

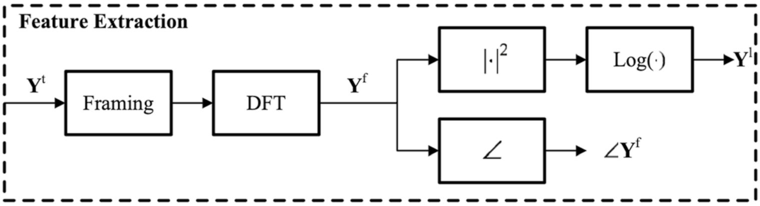


Fig. 6 Demonstrating feature extraction. Here, Y^t represents the noisy signal in time domain, Y^f represents the transformed signal in the frequency domain. Y^l is the log power features of the input signal

effectively combine the different features, such that those complementing each other are retained while redundant ones are eliminated (Wang et al. 2013).

3.1.1 Supervised speech enhancement and separation training with Fourier spectrum features

DNN models that are trained via supervised learning using Fourier spectrum features employ several strategies to learn how to generate estimated clean signal from a noisy (mixed) signal. These strategies can be classified into three categories based on the target of the model.

1. Spectral mapping techniques.
2. Spectral masking techniques.
3. Generative modelling.

3.1.1.1 Spectral mapping techniques These models fit a nonlinear function to learn a mapping from a mixed signal feature to an estimated clean signal feature (see Fig. 7).

The training dataset of these models consist of a noisy speech signal (source) and clean speech (target) features. The process of training these models can be generalised in the following steps:

1. Given N raw waveforms of mixed (noisy) speech, convert the N raw waveform of noisy speech to the desired representation (such as spectrogram).
2. Convert the respective N clean speech waveform in time domain to the same representation as that of the noisy speech.
3. Create an annotated dataset consisting of a pair of noisy speech features and that of clean speech i.e., $< noisy\_speech\_features_i, clean\_speech\_features_i >$ with $1 \leq i \leq N$
4. Train a deep learning model g_θ to learn how to estimate clean features $clean\_speech\_features_i$ given a noisy speech feature as input $noisy\_speech\_features_i$ by minimizing an objective function.
5. Given new a noisy speech features x_j the trained model g_θ should estimate a clean speech feature y_j.
6. Using the estimated clean speech features y_j, reconstruct its raw waveform by performing the inverse of the feature generation process (such as using the inverse short-time Fourier transform if the features are in time-frequency domain).

The above generalisation has been exploited in Fu et al. (2018a), Grais and Plumbley (2018), Kim and Smaragdis (2015), Xu et al. (2014a, 2015), Lu et al. (2013), Fu et al. (2016), Gao et al. (2016) to achieve speech enhancement and in Jansson et al. (2017)and Weninger et al. (2014) to perform speech separation and enhancement. Figure 8 gives a

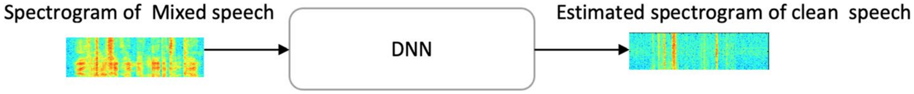

Spectrogram of Mixed speech
DNN
Estimated spectrogram of clean speech

Fig. 7 Supervised training of speech enhancement model using spectrogram as input and spectrogram as output. The DNN maps noisy(mixed) spectrogram to a clean one directly

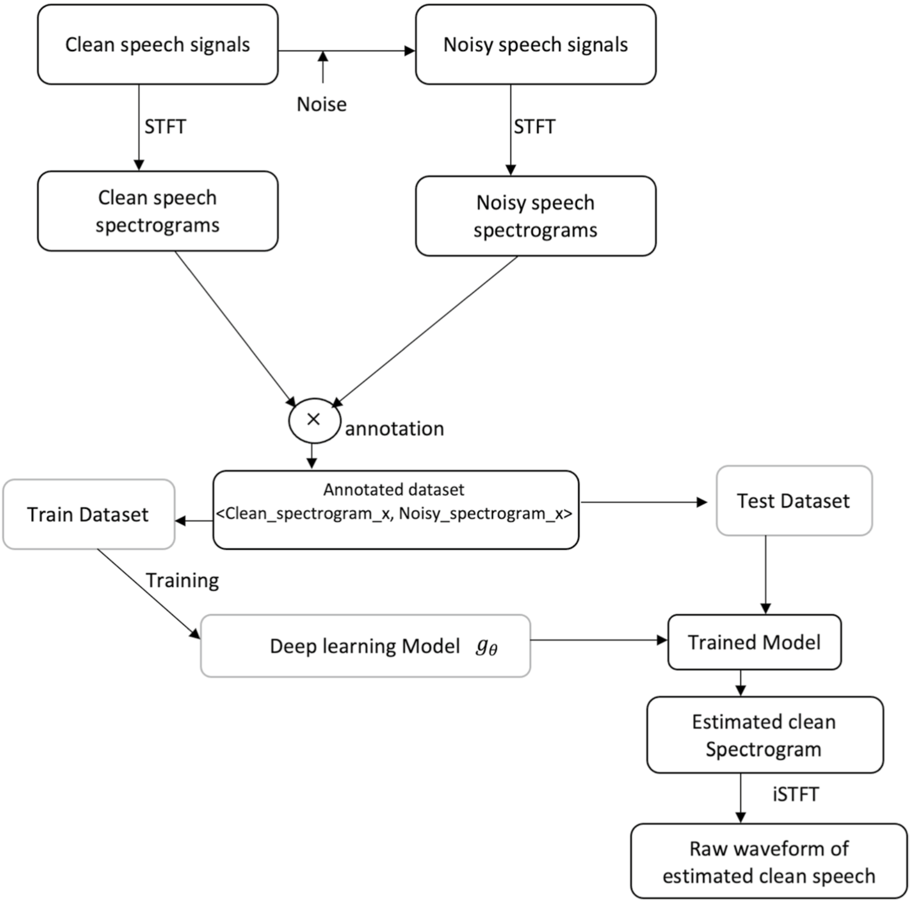

Clean speech signals
Noise
Noisy speech signals
STFT
STFT
Clean speech spectrograms
Noisy speech spectrograms
×
annotation
Train Dataset
Annotated dataset <Clean_spectrogram_x, Noisy_spectrogram_x>
Test Dataset
Training
Deep learning Model g_θ
Trained Model
Estimated clean Spectrogram
iSTFT
Raw waveform of estimated clean speech

Fig. 8 Showing steps involves to train speech enhancement model to fit a regression function from noisy spectrogram to an estimated clean speech spectrogram

summary of the steps when time-frequency (spectrogram) of a noisy speech is exploited as the input of the speech enhancement model.

3.1.1.2 Spectral masking techniques Here, the task of estimating clean speech features from a noisy (mixed) speech input features is formulated as that of predicting real-valued or complex-valued masks (Williamson and Wang 2017a). The mask function is usually constrained to be in range the [0,1] even though different types of soft masks have been proposed (see Kolbæk et al. 2017a; Narayanan and Wang 2013; Erdogan et al. 2015; Nossier et al. 2020b). Source separation based on masks is predicated on the assumption of sparsity and orthogonality of the sources in the domain in which the masks are computed (Cord-Landwehr et al. 2021). Due to the sparsity assumption, the dominant signal at a given range (such as time-frequency bin) is taken to be the only signal at that range (i.e. all other signals are ignored except the dominant signal). In that case, the role of DNN estimated mask is to estimate the dominant source at a given range. To do this, the mask is applied on the input features such that it eliminates portion of the signal(where the mask has a value of 0) while allowing others (mask value of 1) (Kjems et al. 2009; Wang 2008). The masks are always established by computing the signal-to-noise (SNR) within each TF bin against a threshold or a local criterion (Kjems et al. 2009). It has been demonstrated experimentally that the use of masks significantly improves speech intelligibility when an original speech is composed of noise or a masker speech signal (Brungart et al. 2006; Nossier et al. 2020a). For deep learning models working on the time-frequency domain, a model g_θ is designed such that given a noisy or mixed speech spectrogram $Y[t, n]$ at time frame t, it estimates the mask m_t at that time frame. The established mask m_t is then applied to the input spectrogram to estimate target or denoised spectrogram i.e., $\hat{S}_t = m_t \otimes Y[t, n]$ (see Fig. 9). Here, $\hat{S}_t$ is the spectrogram estimate of the clean speech at time frame t and $\otimes$ denotes element wise multiplication. To train the model g_θ, there are two key objective variants. The first type minimizes an objective function D such as mean squared error (MSE) between the model estimated mask $\hat{m}_m$ and the target mask (tm).

$$\mathcal{L} = \sum_{u,t,f} D|tm_{u,t,f}, \hat{m}_{u,t,f}|$$

This approach however cannot effectively handle silences where $|Y[t, n]| = 0$ and $|X[t, n]| = 0$, because the target masks tm will be undefined at the silence bins. Note that target masks such ideal amplitude mask (IRM) that is defined as $IRM(t,f) = \frac{|X_s(t,f)|}{\sum_{i=1}^{s} |Y_s(t,f)|}$ involves division of $|X[t, n]|$ by $|Y[t, n]|$ hence silence regions will make the target mask undefined (Kolbæk et al. 2017a). This cost function also focuses on minimizing the disparity between the masks instead of the features of estimated signal and the target clean signal (Kolbæk et al. 2017a). The second type of cost function seeks to minimize the features of estimated signal $\hat{S}_t = m_t \otimes Y[t, n]$ and those of target clean signal S directly as shown Eq. (25).

$$\mathcal{L} = \sum_{u,t,f} D|\hat{m}_{u,t} Y_{u,t,f}, S_{u,t,f}| \tag{25}$$

The sum is over all the speech u and time-frequency bin (t, f). Here, Y and S represents noisy (mixed) and clean (target) speech respectively. So, for DNN tools using indirect estimation of clean signal features, instead of them estimating the clean features directly from the noisy features input, the models first estimate binary masks. The binary masks are then applied to the noisy features to separate the sources (see Fig. 9, here, the features are the TF spectrogram). This technique has been applied in Wang and Wang (2013), Isik et al. (2016), Weninger et al. (2014), Fu et al. (2016), Narayanan and Wang (2013), Chen et al. (2015), Huang et al. (2015), Hershey et al. (2016), Grais et al. (2014), Zhang and Wang

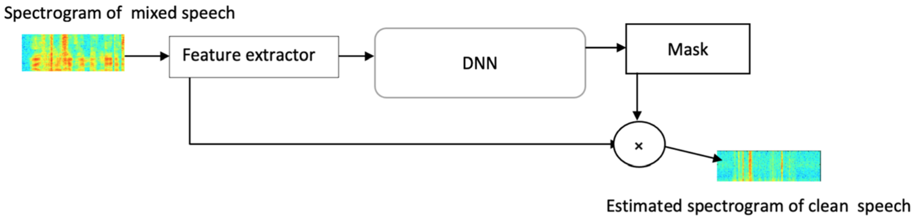

Spectrogram of mixed speech
Feature extractor
DNN
Mask
×
Estimated spectrogram of clean speech

Fig. 9 DNN model for mask estimation from a noisy spectrogram. Here, DNN model is used to estimate a mask which is applied on the noisy input features to estimate clean speech features

(2016), Narayanan and Wang (2015), Weninger et al. (2015), Huang et al. (2011), Zhang and Wang (2016), Liu and Wang (2019).

3.1.1.3 Generative modelling Given an observed sample x, the goal of a generative model is to model its true distribution $p(x)$. The established model can then be used to generate new samples that are similar to the observed samples x. In speech separation and enhancement, these models have been exploited almost exclusively to perform speech denoising. These generative models are required to generate clean speech from a noisy one and maintains the vocal fingerprint of the noisy input speech i.e the models should not collapse to generating the same output even though the input is changing. Several generative models have been employed in a supervised manner to generate clean speech from a noisy one. Generative adversarial network (GAN) (Goodfellow 2016) is an unsupervised model that constitutes two key parts: the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$, where $\mathcal{G}$ generates samples which are then judged by the $\mathcal{D}$. The generator $\mathcal{G}$ generates the synthetic data by sampling from a simple prior $z \sim p(z)$ and the outputs a final sample $g_\theta(z)$ where g_θ is non-linear function more specifically a DNN. The discriminator $\mathcal{D}$ on the other hand must be able to catch synthetic data as fake and real data from $p(x)$ as real. The training objective is shown in Eq. (26).

$$\min_{\mathcal{G}} \max_{\mathcal{D}} = \mathbb{E}_{x \sim p(x)}[\log D(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))] \tag{26}$$

The objective function in Eq. (26) is maximised w.r.t to $D(.)$ and minimised w.r.t $G(.)$. Therefore the objective of GAN is to learn a mapping from a noise vector z to an output x i.e. $G : z \rightarrow x$. In case of speech enhancement, the models seek to map a noisy speech x_c to a clean speech x. Equally, the models are required to preserve the vocal fingerprints of the input noisy speech. Based on this, most GAN based speech enhancement models use conditioned GAN (Isola et al. 2017) i.e. $G : (x_c, z) \rightarrow x$ as shown in the objective below.

$$\min_{\mathcal{G}} \max_{\mathcal{D}} = \mathbb{E}_{x \sim p(x, x_c)}[\log D(x, x_c)] + \mathbb{E}_{z \sim p(z), x_c \sim p(x_c)}[\log(1 - D(G(x_c, z), x_c))]$$

The random noise z is added to avoid the DNN producing deterministic outputs and therefore fail to learn any distribution other than a delta function. To further improve the quality of generated enhanced speech, the GAN based denoising models use least square GAN (LSGAN) (Mao et al. 2017). In the original GAN objective (Eq. 26), as long as a generated sample lies in the correct side of the decision boundary, the objective causes almost no loss. However, for LSGAN it penalizes samples based on how far they are from the decision boundary even if they lie on the correct side. Due to this, LSGAN generates speeches that maintain the vocal fingerprints of the noisy input speech. CGAN (Donahue et al. 2018) uses conditioned GAN that accepts input in T-F domain to generate a denoised speech. Since most automatic speech recognition (ASR) tools work in T-F domain, CGAN hypothesises that the generative model working in T-F domain will be more robust for ASR as compared to those working in raw waveform. Therefore, CGAN can be seen as a speech enhancement tool targeting ASR. To address the problem of mismatch between the training objective used in CGAN and the evaluation metrics, MetricGAN (Fu et al. 2019) proposes to integrate evaluation metric in the discriminator. By doing this, instead of the generator giving a false (0) or true (1) discrete values, it will generate continuous values based on the evaluation metric. MetricGAN can therefore be trained to generate data according to the the selected metric score. Through this modification, MetricGAN produces more robust enhanced speech. Both these two models use conditioned GAN to preserve the vocal fingerprints of the input speech. Another common generative group of generative models is variational auto-encoder (VAE) technique (Kingma and Welling 2014). Like GAN, VAE is mainly used for denoising i.e. where the mixture is modelled as:

$$x_{fn} = \sqrt{g_n} s_{fn} + b_{fn} \tag{27}$$

Here, x_{fn} denotes the mixture at the frequency index f and the time-frame index n, $g_n \in R_+$ is a frequency independent but frame dependent gain while s_{fn} and b_{fn} represent the clean speech and the noise respectively at the frequency index f and the time-frame index n. We first give a brief overview of VAE before we discuss how it is adapted for speech enhancement. Mathematically, given an observable sample s, the goal of a generative VAE model is to model true data distribution $p(s)$. To do this, VAE assumes that the observed sample s are generated by associated latent variable z and their joint distribution is $p(s, z)$. The model therefore seeks to learn how to maximize the likelihood $p(s)$ over all observed data.

$$p(s) = \int p(s, z) dz$$

Integrating out all the latent variables z in the above equation is intractable. However, using evidence lower bound (ELBO) which quantifies the log-likelihood of observed data, $p(s)$ can be estimated. ELBO is given in Eq. (28) (refer to Lu et al. (2022) to see derivation of relationship between $p(s)$ and ELBO).

$$\log p(s) \geq \mathbb{E}_{q_\phi(z|s)}\left[\log \frac{p(s, z)}{q_\phi(z \mid s)}\right] \tag{28}$$

Here, $q_\phi(z \mid s)$ is a flexible variational distribution with parameters ϕ that the model seeks to maximize. Equation (28) can be written as Eq. (29) using Bayes theorem.

$$\log p(s) \geq \mathbb{E}_{q_\phi(z|s)}\left[\log \frac{p_\theta(s \mid z) p(z)}{q_\phi(z \mid s)}\right] \tag{29}$$

Equation (29) can be expanded as:

$$\log p(s) \geq \mathbb{E}_{q_\phi(z|s)}\left[\log p_\theta(s \mid z)\right] + \mathbb{E}_{q_\phi(z|s)}\left[\frac{\log p(z)}{q_\phi(z \mid s)}\right] \tag{30}$$

Equation (30) can be expanded as:

$$\log p(s) \geq \mathbb{E}_{q_\phi(z|s)}[\log p_\theta(s \mid z)] + D_{KL}(q_\phi(z \mid s) \mid\mid p(z)) \tag{31}$$

The second term on the right of Eq. (31) seeks to learn the posterior $q_\phi(z \mid s)$ via prior $p(z)$ while the first term reconstructs data based on the learned latent variable z. $q_\phi(z \mid s)$ is always modelled by a DNN and referred to as encoder and the reconstruction term is another DNN referred to as decoder. Both the encoder and decoder are trained simultaneously. The encoder is normally chosen to model a multivariate Gaussian with diagonal covariance and the prior is often selected to be a standard multivariate Gaussian:

$$q_\phi(z \mid x) = \mathcal{N}(z; \mu_\theta(x), \delta_\theta^2(x) I)$$
$$p(z) = \mathcal{N}(z; 0, I)$$

To estimate clean speech based on variational-autoencoder pre-training, the tools execute several techniques that can be generalised into the following steps:

1. Train a model such that it can maximise the likelihood $p_\theta(s \mid z)$. Here, s denotes the clean speech dataset that is composed of F-dimensional samples i.e $s_t \in R^F, 1 \leq t \leq T$. The variational autoencoder assumes a D-dimensional latent variable $z_t \in R^D$. The latent variable z_t and the clean speech s_t have the following distribution:

$$z_t \sim \mathcal{N}(0, I_D)$$
$$s_t \sim p(s_t|z_t)$$

   Here, $\mathcal{N}(\mu, \delta)$ denotes a Gaussian distribution with mean μ and variance δ. Basically, a decoder $p_\theta(s_t \mid z_t)$ is trained to generate clean speech s_t when given the latent variable z_t, the decoder is parameterized by θ. The decoder $p_\theta(s_t \mid z_t)$ is learned by deep learning model during training. The encoder is trained to estimate the posterior $q_\phi(z_t|s_t)$ using a DNN. The overall objective of the variational auto-encoder training is to maximise Eq. (32).

$$p(s) = *argmin_{\theta,\phi} \sum_{i=1}^{L} \log p_\theta(s \mid z^i) + D_{KL}(q_\phi(z^i \mid s) \mid\mid p(z)) \tag{32}$$

   The posterior estimator $q_\phi(z \mid s)$ is a Gaussian distribution with parameters μ_d and δ_d. These parameters are to be established by the encoder deep neural network such that $\mu_d : R^F \rightarrow R$ and $\delta_d : R^F \rightarrow R_+$.
2. Set up a noise model using unsupervised techniques such as NMF (Hien et al. 2015). For example, in case of NMF the noise b_{fn} in equation 27 can be modelled as

$$b_{bf}; w_{b,f}, h_{b,n} \sim \mathcal{N}(0, (W_b, H_b)_{f,n}) \tag{33}$$

   where $\mathcal{N}(0, \delta)$ is a Gaussian distribution with zero mean and variance of δ.
3. Set up a mixture model such that $p(x \mid z, \theta_s, \theta_u)$ is maximised. Here x is the noisy speech signal, θ_s are parameters from the pre-trained model in step 1 i.e ϕ and θ. $\theta_u = g_n, (W_b, H_b)_{f,n}$ represents the parameters to be optimised. The parameters are θ_u are optimised by appropriate Bayesian inference technique.
4. Reconstruct the estimated clean speech $\hat{s}$ such that $p(\hat{s}|\theta_u, \theta_s, x)$ is maximised based on the parameters θ_u, θ_s from step 1 and 3 respectively and the observed mixed speech x.

Works that exploit different versions of variational auto-encoder technique include Leglaive et al. (2018, 2019, 2020), Bando et al. (2018)). Another generative modelling technique that has been used in speech enhancement is the variational diffusion model (VDM) Sohl-Dickstein et al. 2015. VDM is composed of two processes i.e., diffusion and reverse process. The diffusion process perturbs data to noise and the reverse process seeks to recover data from noise. The goal of diffusion therefore is to transform a given data distribution into a simple prior distribution mostly standard Gaussian while the reverse process recovers data by learning a decoder parameterised by DNN. Formally, representing true data samples and latent variables as x_t where $t = 0$ represents true data and $1 \leq t \leq T$ represents a sequence of latent variables, the VDM posterior is represented as:

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}) \tag{34}$$

The VDM encoder $q(x_t \mid x_{t-1})$ unlike that of VAE, is not learned rather it is a predefined linear Gaussian model. The Gaussian encoder is parameterized with mean $u_t(x_t) = \sqrt{\alpha_t}x_{t-1}$ and variance $\varepsilon_t = (1-\alpha_t)I$. Therefore, the encoder $q(x_t \mid x_{t-1})$ can mathematically be represented as

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I). \tag{35}$$

α_t evolves over time such that the final distribution of the latent $p(x_T)$ is a standard Gaussian. The reverse process seeks to train a decoder that starts from the standard Gaussian distribution $p(x_T)$. Formally the reverse process can be represented as:

$$p(x_{0:T}) = p(x_T) \prod_{i=1}^{T} p_\theta(x_{t-1} \mid x_t) \tag{36}$$

Here $p(x_T) = \mathcal{N}(x_T; 0, I)$. The reverse process seeks to set up a decoder $p_\theta(x_{t-1} \mid x_t)$ that optimizes the parameter θ such that: the conditionals $p_\theta(x_{t-1} \mid x_t)$ are established. Once the VDM is optimized, a sample from the Gaussian noise $p(x_T)$ can iteratively be denoised through transitions $p_\theta(x_{t-1} \mid x_t)$ for T steps to generate a simulated x_0. Using reparameterization trick, x_t in Eq. (34) can be rewritten as:

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon \tag{37}$$

where $\epsilon \sim \mathcal{N}(\epsilon, O, I)$ Similarly,

$$x_{t-1} = \sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_{t-1}}\epsilon \tag{38}$$

Based on this and through iterative derivation of Eq. (34), it can be shown that:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_0 \tag{39}$$

In the reverse process in Eq. (36), the transition probability $p_\theta(x_{t-1} \mid x_t)$ can be represented by two parameters μ_θ and δ_θ as $\mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \delta_\theta(x_t, t)^2 I)$ with θ being the learnable parameters. It has been shown in Luo (2022) that $\mu_\theta(x_t, t)$ can be established as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon_\theta(x_t, t) \tag{40}$$

Based on Eq. (40), to estimate $\mu_\theta(x_t, t)$ the DNN $\epsilon_\theta(x_t, t)$ needs to estimate the Gaussian noise ϵ in x_t which was injected during the diffusion process. Like VAE, VDM uses ELBO objective for optimization. Please see Luo (2022) for a thorough discussion on VDM. In speech denoising, work in Lu et al. (2022) uses conditional diffusion process to model the encoder $q(x_t \mid x_{t-1})$. In conditional encoder, instead of $q(x_t \mid x_{t-1})$, they define it as $q(x_t \mid x_0, y)$ i.e., $q(x_t \mid x_0, y) = \mathcal{N}(x_t; (1-m_t)\sqrt{\bar{\alpha}}x_0 + m_t\sqrt{\bar{\alpha}}y, \delta_t I)$. Here x_0,y represents the clean speech and noisy speech respectively. The encoder is modeled as a linear interpolation between clean speech x_0 and the noise speech y with interpolation ratio m_t. The reverse process $p_\theta(x_{t-1} \mid x_t)$ is also modified to $p_\theta(x_{t-1} \mid x_t, y) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, y, t), \delta I)$. Here,

$\mu_\theta(x_t, y, t)$ is the mean of the conditional reverse process. similar to Eq. (40), $\mu_\theta(x_t, y, t)$ is estimated as

$$\mu_\theta(x_t, y, t) = c_{xt}x_t + c_{yt}y - c_{\epsilon t}\epsilon_\theta(x_t, y, t) \tag{41}$$

where $\epsilon_\theta(x_t, y, t)$ is a DNN model to estimate the combination of Gaussian and non-Gaussian noise. The coefficients c_{xt}, c_{yt} and $c_{\epsilon t}$ are established via the ELBO optimization.

3.1.2 Highlights on Fourier spectrum features

1. When performing a DFT on the input signal, an optimum window length must be selected. The choice of the window has a direct impact on the frequency resolution and the latency of the system. To achieve good performance, most systems use 32ms. This may limit the use of the DFT based models in environments which require short latency (Luo and Mesgarani 2018).
2. DFT is a generic method for signal transformation that may not be optimised for waveform transformation in speech separation and enhancement. It is therefore important to know to what extent does it place an upper bound on the performance level of speech enhancement techniques.
3. Accurate reconstruction of estimated clean speech from the estimated features is not easy and the erroneous reconstruction of clean speech places an upper bound on the accuracy of the reconstructed audio.
4. Perhaps the biggest challenge when working in the frequency domain is how to handle the phase. Most DNN models only use the magnitude spectrum of the noisy signal to train the DNN then factor in the phase of the noisy signal during reconstruction. Recent works such as Paliwal et al. (2011) have shown that this technique does not generate optimum results.
5. While working in the frequency domain, experimental research has demonstrated that spectral masking generates better results in terms of enhanced speech quality as compared to the spectral mapping method (Nossier et al. 2020a).

3.1.3 Handling of phase in frequency domain

The assumption made by most DNN models that use Fourier spectrum features is that phase information is not crucial for human auditory. Therefore, they exploit only the magnitude or power of the input speech to train the DNN models to learn the magnitude spectrum of the clean signal and factor in the phase during the reconstruction of the signal (see Fig. 10) (Xu et al. 2014a; Kumar and Florencio 2016; Du and Huo 2008; Tu et al. 2014; Li et al. 2017). The use of the phase from the noisy signal to estimate the clean signal is based on works such as Ephraim and Malah (1984) that demonstrated that the optimal estimator of the clean signal is the phase of the noisy signal. Further, most speech separation models work on frames that are of size between 20 and 40 ms and believe that the short-time phase contain low information (Lim and Oppenheim 1979; Oppenheim and Lim 1981; Vary and Eurasip 1985; Wang and Lim 1982) and therefore not crucial when estimating clean speech. However, recent research (Paliwal et al. 2011) have demonstrated through experiments that further improvements in quality of estimated clean speech can be attained by processing both the short-time phase and magnitude spectra. Further, the factoring in of the noisy input phase during reconstruction has been noted to be a problem since the phase

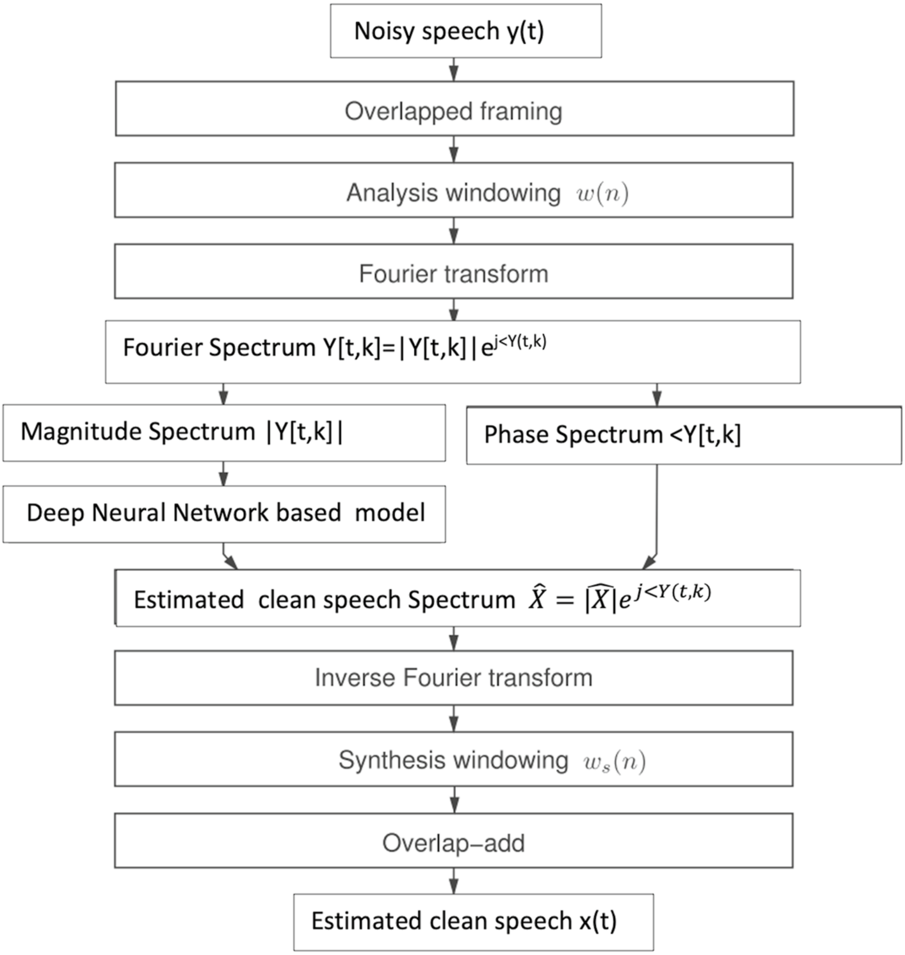

Noisy speech y(t)
Overlapped framing
Analysis windowing $w(n)$
Fourier transform
Fourier Spectrum Y[t,k]=|Y[t,k]|$e^{j<Y(t,k)}$
Magnitude Spectrum |Y[t,k]|
Phase Spectrum <Y[t,k]
Deep Neural Network based model
Estimated clean speech Spectrum $\hat{X} = |\widehat{X}|e^{j<Y(t,k)}$
Inverse Fourier transform
Synthesis windowing $w_s(n)$
Overlap–add
Estimated clean speech x(t)

Fig. 10 Showing how DNN models exploit the phase of the noisy signal during reconstruction of the estimated signal

errors in the input interact with the amplitude of the estimated clean signal hence causing the amplitude of the estimated clean signal to differ with the amplitude of the actual clean signal being estimated (Erdogan et al. 2015; Han et al. 2015). Based on the realisation of the importance of phase, some studies have avoided factoring in the phase of the noisy signal but rather exploit a modified phase to estimate the clean signal. The existing techniques of modifying the phase by the DNN models working on Fourier spectrum features can be categorised into two:

Phase learning These models make the phase part of the objective function i.e., they learn the phase of the estimated clean signal during training. To integrate the phase in the learning process works such as Erdogan et al. (2015), Kolbæk et al. (2017a) use a phase sensitive objective by replacing Eqs. (42) with (43). It essentially exploits a phase sensitive spectrum approximation objective by minimising the distance between the raw waveform of the estimated speech and that of the target clean speech.

```
**Require:** Target clean magnitude $X^0$, noisy phase $\phi^0$, iteration N.
  $X \leftarrow X^0, \phi \leftarrow \phi^0, n \leftarrow 1$
  **while** $n \leq N$ **do**
      $s^n \leftarrow iSTFT(X, \phi)$
      $(X^n, \phi^n) \leftarrow STFT(s^n)$
      $X \leftarrow X^0$
      $\phi \leftarrow \phi^n$
      $n \leftarrow n + 1$
  **end while**
  $s \leftarrow s^n$
```

Algorithm 1 Iteratively updating the phase of a noisy signal

$$\mathcal{L} = \sum_{u,t,f} D|\hat{m}_{u,t} Y_{u,t,f}, S_{u,t,f}| \tag{42}$$

$$\mathcal{L} = \sum_{u,t,f} D|\hat{m}_{u,t} Y_{u,t,f}, S_{u,t,f} \cos(\theta_y - \theta_s)| \tag{43}$$

Here D, is a selected objective function such as MSE, θ_y and θ_s represent the phase of the noisy and clean (target) speech respectively. The sum is over all the speech u and time-frequency bin (t, f). Experiments conducted based on the objective function in Eq. (43) show superior results in terms of signal-to-distortion ratio (SDR) (Erdogan et al. 2015). Work in Williamson et al. (2016) trains a DNN model to generate masks that are composed of both the real and imaginary part (see Eq. 44). The complex mask will then be applied to a complex representation of the noisy signal to generate the estimated clean signal. By learning a mask that has both the real and imaginary part, they integrate the phase as part of the learning.

$$L = \frac{1}{2N} \sum_{t} \sum_{f} [(O_r(t,f) - M_r(t,f))^2 + (O_i(t,f) - M_i(t,f))^2] \tag{44}$$

O_r is the real part of the mask estimated by the DNN model while O_i is the imaginary part. M_r is the real part of the target mask while M_i is the imaginary part. N is the number of frames and (t,f) is a given TF bin. The complex mask implementation has been exploited in Williamson and Wang (2017a), Erdogan et al. (2015), Lee et al. (2017) where the targets are formulated in the complex coordinate system i.e. the magnitude and phase are composed as part of the learning process. Work in Wang et al. (2018a) proposes a model that learns the phase during training via input spectrogram inversion (MISI) algorithm (Gunawan and Sen 2010). Work in Ai et al. (2021) proposes a generative adversarial network (GAN) (Goodfellow 2016) based technique of learning the phase during training. Other works that learn the phase during training include Wang et al. (2019) and Le Roux et al. (2019). Techniques that include phase as part of the training face the difficulty of processing a phase spectrogram which is randomly distributed and highly unstructured (Zheng and Zhang 2019). To mitigate this problem and derive a highly structured phase-aware target masks, Zheng and Zhang (2019) employs instantaneous frequency (IF) (Friedman 1985) to extract structured patterns from phase spectrograms.

Post-processing phase update The models that use this technique, train the DNN models using only the magnitude spectrum. Once the model has been trained to estimate

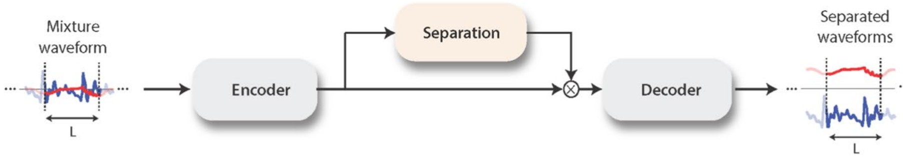


Fig. 11 Showing adaptive front-end implementation. The model replaces STFT with a differentiable transform in the encoder that is jointly trained with the separation model

the magnitude spectrum of the clean signal, they iteratively update the phase of the noisy signal to be as close as possible to that of the target clean signal. The algorithm being exploited by the models performing post-processing phase update is based on the Griffin-Lim algorithm proposed in Griffin and Lim (1984). For example, in Han et al. (2015), they exploit the magnitude X^0 of the target clean signal to iteratively obtain an optimal phase ϕ from the phase of the noisy signal (see Algorithm 1). The obtained phase is then used in the reconstruction of the estimated clean signal together with the magnitude $\hat{X}$ estimated by the DNN. The technique is also used in Zhao et al. (2019). Techniques that implement Griffin-Lim algorithm such as in algorithm 1 perform iterative phase reconstruction of each source independently and may not be effective for multiple source separation where the sources must sum up to the mixture (Wang et al. 2018a). Work in Wang et al. (2018a) proposes to jointly reconstruct the phase of all sources in a given mixture by exploiting their estimated magnitudes and the noisy phase using the multiple input spectrogram inversion (MISI) algorithm (Gunawan and Sen 2010). They ensure that the sum of the reconstructed time-domain signals after each iteration must sum to the mixture signal. Work (Li et al. 2016) and (Choi et al. 2020) also uses post-processing to update the phase of the noisy signal.

3.2 Time-domain features

Due to the challenges highlighted in Sect. 3.1.2 of working in the time-frequency domain, different models such as Luo and Mesgarani (2018), Luo et al. (2020), Luo and Mesgarani (2019), Venkataramani et al. (2018), Zhang et al. (2020a), Subakan et al. (2021), Tzinis et al. (2020a), Tzinis et al. (2020b), Kong et al. (2022), Su et al. (2020), Lam et al. (2021a), Lam et al. (2021b) explore the idea of designing a deep learning model for speech separation that accepts speech signal in the time-domain. The fundamental concept for these models is to replace the DFT based input with a data-driven representation that is jointly learned during model training. The models therefore accept as their input the mixed raw waveform sound and then generates either the estimated clean sources or masks that are applied on the noisy waveform to generate clean sources. By working on the raw waveform, these models address two key limitations of DFT based models. First, the models are designed to fully learn the magnitude and phase information of the input signal during training (Luo et al. 2020). Secondly, they avoid reconstruction challenges faced when working with DFT features. The time domain methods can broadly be classified into two categories (Luo et al. 2020).

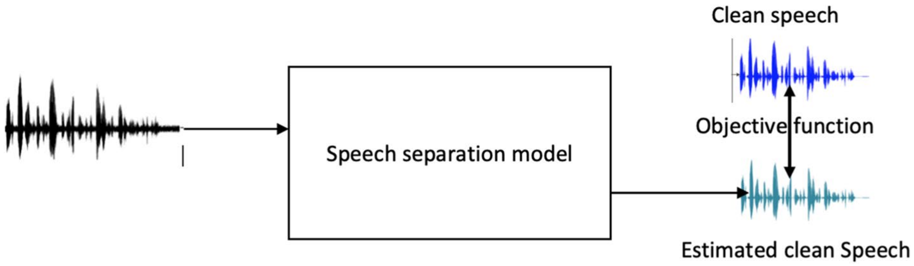


Fig. 12 Direct approach training of DNN models using raw waveform

3.2.1 Adaptive front-end based method

The models in this category can roughly be discussed as composed of three key modules i.e., the encoder, separation and decoder modules (see Fig. 11).

1. *Encoder* The encoder can be regarded as an adaptive front-end which seeks to replace STFT with a differentiable transform that is jointly trained (learned) with the separation model. It accepts as its input a time-domain mixture signal then learns STFT-like representation (Subakan et al. 2021; Kong et al. 2022). By working directly with the time-domain signal, these models avoid the decoupling of the magnitude and phase of the input signal (Luo and Mesgarani 2019). Most systems employ 1-dimensional convolution as the encoder to learn the features of the input signal. The transform generated by the encoder is then passed to the separation module. Work in Kavalerov et al. (2019) demonstrates that learned bases from raw data produce better results for speech/non-speech separation.
2. *Separation module* This module is fed by the output of the encoder. It implements techniques to identify the different sources present in the input signal.
3. *Decoder* It accepts input from the separation module and sometimes from the encoder(for residual implementation). It is mostly implemented as an inverse of the encoder in order to reconstruct the separated signals (Luo and Mesgarani 2018, 2019; Subakan et al. 2021).

3.2.2 Waveform mapping

The second category of systems implement end-to-end systems where they utilise deep learning models to fit a regression function that maps an input mixed signal to its constituent estimated clean signal without an explicit front-end encoder (see Fig. 12). The models are trained using a pair of mixed(noisy) and clean speech. The model is fed with features of mixed signal for it to estimate clean speech. The training involves minimising an objective function such as minimum mean square error(MMSE) between the features of the clean signal and the estimated clean signal generated by the model. This approach has been implemented in Stoller et al. (2018), Fu et al. (2018b), Lluís et al. (2019).

3.2.3 Generative modelling

SEGAN (Pascual et al. 2017) is GAN based model for speech denoising that conditions both G and D of Eq. (26) on extra information z representing latent representation of the input. To solve the problem of vanishing gradient associated with optimizing objective in Eq. (26), they replace the cross-entropy loss by a least square function in Eq. (45).

$$\min_{\mathcal{G}} = \mathbb{E}_{z \sim p(z), \bar{x} \sim p(\bar{x})}[(D(G(z,\bar{x}),\bar{x}) - 1)^2)] + \lambda ||G(z,\bar{x}) - x||_1] \tag{45}$$

Here, $\bar{x}$ is the noisy speech, x is the clean speech, z is the extra input latent representation and $||.||_1$ is the l_1 norm distance between the clean sample x and the generated sample $|G(z,\bar{x})$ to encourage the generator G to generate more realistic audio. Work in Pascual et al. (2019) improves SEGAN to handle a more generalised speech signal distortion case which involves distortions such as chunk removal, band reduction, clipping and whispered speech. Work (Phan et al. 2020) improves SEGAN by implementing multiple generators as opposed to one and demonstrates that by doing so the speech quality of the enhanced speech is better than when a single generator is used. Work in Adiga et al. (2019) proposes a variation of SEGAN that is more tailored towards speech synthesis and not ASR. They replace the original loss function used in SEGAN with Wasserstein distance with gradient penalty(WGAN) (Gulrajani et al. 2017). They also exploit gated linear unit as activation function which has been shown in van den Oord et al. (2016) to be more robust in generating realistic speech. Other GAN based models for speech enhancement in the time domain include Xiao et al. (2021). Other tools that implement supervised conditional GAN include Donahue et al. (2018), Li et al. (2018a), Qin and Jiang (2018), Fu et al. (2019). In Qian et al. (2017), Bayesian network is exploited to generate estimated clean speech from a noisy one.

3.3 Challenges of working with time-domain features

1. Time domain features lack direct frequency representation; this hinders the features from capturing speech phonetics that are present in the frequency domain. Due to this, artefacts are always introduced in the reconstructed speech in the time domain (Cao et al. 2022).
2. The time domain waveform has a large input space. Based on this, models working with raw waveforms are often deep and complex in order to effectively model the dependencies in the waveform. This is computationally expensive (Défossez et al. 2020; Pascual et al. 2017; Subakan et al. 2021; Wang et al. 2021).

.

4 Which feature produces superior quality of enhanced speech?

We performed analysis of 500 papers that exploit DNN to perform speech enhancement(i.e., multi-talker speech separation or denoising or dereverberation). We selected papers published from 2018 to 2022. We were interested to answer the question, which features are more popular with these tools? The summary is presented in Fig. 13. Based on the analysis, time-domain features popularity has grown rapidly from 2018 to 2022. The use of DFT

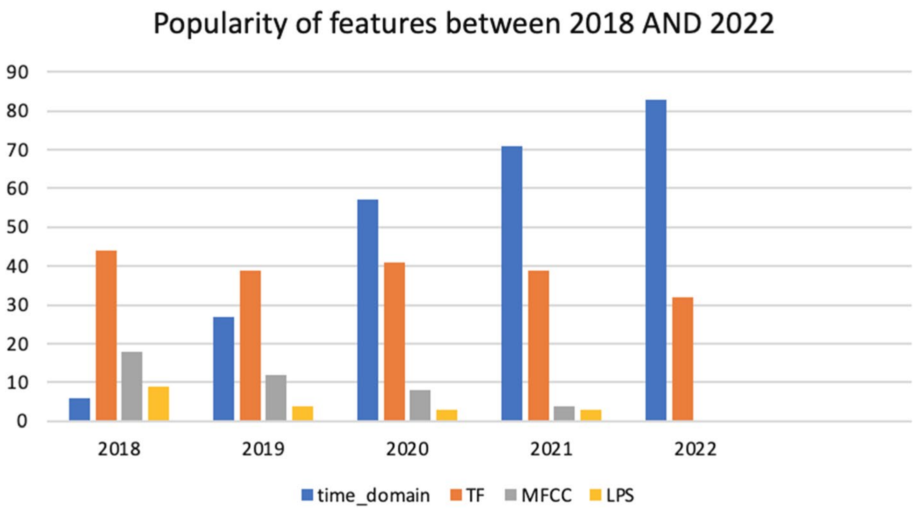


Fig. 13 Feature popularity between 2018 and 2022

Table 1 Comparison of different encoder and decoder combination using Lw = 4 ms and Ls = 2ms on the test set of the WSJ0-2mix database

Loss	Encoder	Decoder	si_SDR dB	SDB dB	WER %
L^{SI_SDR}	Learned	Learned	14.4	14.7	21.71
L^{SI_SDR}	STFT	Learned	13.9	14.3	21.92
L^{SI_SDR}	Learned	iSTFT	14.1	14.5	21.87
L^{SI_SDR}	STFT	iSTFT	12.4	12.8	24.69

features has slightly dropped, however remains popular over the five years. The popularity of MFCC and LPS has diminished. The popularity of features that are computationally expensive such as time-domain and DFT features may be attributed to the improved computation power of computers and efficient sequence modelling techniques such as transformers and temporal convolutional networks (see Sect. 5 for discussion). Features such as MFCC are becoming less popular due to their reduced resolution, which must be extrapolated during reconstruction hence placing an upper bound on the quality of enhanced speech.

We also investigated whether DFT or time-domain features produced the highest quality enhanced speech. Several works have conducted experiments with the goal to answer this question. Notable works include Heitkaemper et al. (2020) and Bahmaninezhad et al. (2019). For example, Heitkaemper et al. (2020) investigates Conv-TasNet's Luo and Mesgarani (2019) performance under different input types in the encoder and decoder. Conv-TasNet uses a frame length of 4ms, stride of 2 ms and overlap of 2 ms. Sample results presented in Heitkaemper et al. (2020) are presented in Table 1 where evaluation parameters include scale-invariant signal-to-distortion (si_SDR), signal-to-distortion (SDR), word error rate (WER). The results in Table 1 show that the Conv-TasNet model gives marginally better results in terms of *si_SDR*, *SDBdB* and *WER* when the input is in time domain where the signal representation is learned by the encoder and output is learned by the decoder. The results are significantly reduced in all the three parameters if STFT is used as the input and its inverse used in the decoder. For instance, Conv-TasNet model achieves a SDR of 14.7 when time-domain features are used. This drops to 12.8 when DFT features are used. This shows that working in the time domain may be better for this setting as compared to the frequency domain. Work in Bahmaninezhad et al. (2019) also shows the same trend where working in time-domain provides better results as compared to frequency domain. However, for mixed speech with reverberation, the use of a time domain signal does not improve the same results as compared to the frequency domain and further investigation on behaviour of both time and frequency features in the presence of reverberation is needed (Bahmaninezhad et al. 2019).

5 Long term dependencies modelling

To effectively perform speech separation, the speech separation tools need to model both long and short sequences within the audio signal. To do this, existing tools have employed several techniques:

5.1 Use of RNN

The initial speech separation models such as Gao et al. (2016), Xu et al. (2015), Xu et al. (2014b) relied on a feedforward DNN to estimate clean speech from a noisy one. However, feedforward DNN models are ill poised for speech data since they are unable to effectively model long dependencies across time that are present in the speech data. Due to this, researchers progressively introduced recurrent neural networks (RNN) which have a feedback structure such that the representations at given time step t is a function of the data at time t, the hidden state and memory at time $t-1$. One such RNN that has been exploited in speech separation is long-short-term memory (LSTM) (Hochreiter and Schmidhuber 1997). LSTM has memory blocks that are composed of a memory cell to remember the temporary state and several gates to control the information and gradient flow. LSTM structures can be used to model sequential prediction networks which can exploit long-term contextual information (Hochreiter and Schmidhuber 1997). Works in Weninger et al. (2014), Chen et al. (2015), Han et al. (2019) exploit LSTM to perform speech separation while (Erdogan et al. 2015) uses bidirectional long short-term memory (BLSTM) networks to make use of contextual information from both sides in the sequence. Due to their inherently sequential nature, RNN models are unable to support parallelization of computation. This limits their use when working with large datasets with long sequences due to slow training (Subakan et al. 2021). Moreover, in speech separation, a typical frame(input features) is usually 25ms which corresponds to 400 samples at a 16kHz sampling rate, for LSTM to work directly on the raw waveform, it would require unrolling the LSTM for an unrealistic large number of time steps to cover an audio of modest length (Sainath et al. 2015). Other models that use different versions of RNN include Parveen and Green (2004). Models such as Wichern and Lukin (2017) use the gated recurrent unit (GRU) (Cho et al. 2014) to perform speech denoising.

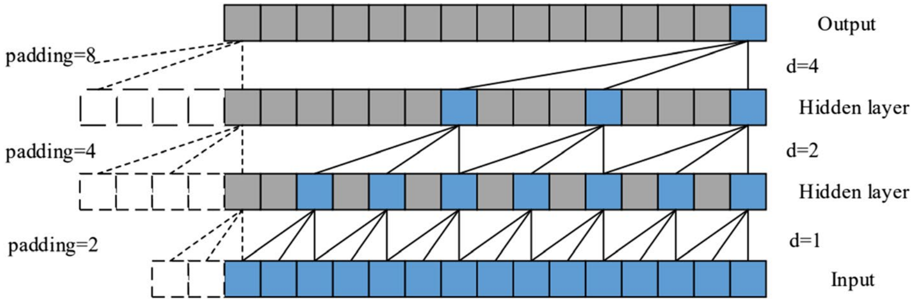


Fig. 14 Dilated TCN with four layers

5.2 Use of temporal convolution network

Conventional convolution neural networks(CNN)have been used to design speech separation models (Jansson et al. 2017; Chandna et al. 2017). However, CNNs are limited in their ability to model long-range dependencies due to limited receptive fields (Chen et al. 2020). They are therefore mainly tailored to learn local features. They exploit local window which maintain translation equivariance to learn a shared position-based kernel (Gulati et al. 2020). For CNN to capture long range dependencies (i.e., to enlarge the receptive field), there is a need to stack many layers. This increases computation cost due to the large number of parameters. These shortcomings of the CNN and RNN, have motivated the use of dilated temporal convolution network (TCN) in speech separation to encode long-range dependencies using hierarchical convolutional layers (Rethage et al. 2018; Li et al. 2018b; Lea et al. 2017; Zeghidour and Grangier 2021; Zhang et al. 2020b). TCN is composed of two key distinguishing characteristics: the convolution in the model must be causal i.e., a given activation of a certain layer l at time t is only influenced by activations of the previous layer $l-1$ from time steps that are less that t, 2) the model takes the sequence of any length and maps it into an output sequence of the same length. To achieve the second characteristic, TCN models are implemented using a 1-dimensional convolutional network such that each hidden layer is the same length as the input layer. To ensure same length, a zero padding of length $filtersize - 1$ is added to keep subsequent layers the same length as previous ones (Bai et al. 2018) (see Fig. 14). The first property is achieved through the use of causal convolutions i.e. where an output at time t is convolved only with elements from time t and earlier in the previous layer. To increase the receptive fields, models implement dilated TCN. Dilated convolution is where the filter is applied to a region larger than its size (He and Zhao 2019). This is achieved by skipping input with certain specified steps (see Fig. 10). More formally, for 1D sequence such as speech signal, the input $x \in R^n$ and the kernel $f : \{0, \dots, k-1\} \rightarrow R$, the dilated convolution operation F on an element s of a given sequence is defined according to Eq. (46) (Bai et al. 2018).

$$F(s) = (x *_d f)(s) = \sum_{i=0}^{k-1} f(i)x_{s-di} \tag{46}$$

where x is the $1D$ input signal, k is the kernel and d is the dilation factor. The effect of this is to expand the receptive field without loss of resolution and drastically increase the number of parameters. Stacked dilated convolution expands the receptive field with only a few layers. The expanded receptive field allows the network to capture temporal dependence of various resolutions with the input sequences (Zhang et al. 2020a). In effect, TCN introduces the idea of time-hierarchy where the upper layers of the network model longer input sequences on larger timescales while local information are modelled by lower layers and are mainly maintained in the network through residuals and skip connections (Zhang et al. 2020a). TCN also uses causal convolution where a given output at layer l in time step t is computed only based on time steps up to time step $t-1$ in the previous layer. The dilated TCN is exploited by Luo and Mesgarani (2019) to model sequences that exist within the input speech signal. They implement TCN such that each layer is composed of 1-D dilated convolution blocks. The layers have 1-D CNN blocks with increasing dilation factors. This is to uncover long range dependencies that exist in the audio input. The dilation factors increase exponentially over the layers in order to cover a large temporal context window to exploit the long-range dependencies that exist within a speech signal.

$$y(m,n) = \sum_{i=1}^{M} \sum_{j=1}^{N} x(m + r \times i, n + r \times j) w(i,j) \tag{47}$$

Here, $y(m, n)$ is the output of a given layer of dilated convolution, $x(m, n)$ is the input and $w(i, j)$ is the filter with the length and the width of M and N respectively. The parameter r is the dilation rate. Note that if $r = 1$, the dilated convolution becomes the normal convolution convolution.

5.3 Use of transformers

A transformer (Vaswani et al. 2017) is an attention-based deep learning technique that has been successful in modelling sequences and allows uncovering of dependencies that exist within an input without regard to the distance between any two values of the input. Transformers consist only of feed-forward layers which allows them to exploit the parallel processing capabilities of GPUs leading to fast training (Vaswani et al. 2017). In speech separation, Subakan et al. (2021) introduces a speech separation system that fully relies on transformers to model the dependencies that exist in the mixed audio signal. This is used to extract a mask for each of the speakers in the audio mixture. The transformer is used to uncover both the short-term dependencies (within a frame) and long-term dependencies (between frames). Work in Chen et al. (2020) also exploits transformers in the encoder to model the dependencies that exist in the mixed audio while (Zhao et al. 2020) uses transformers to perform speech dereverberation. Despite their ability to model long-range dependencies and ability to work well with parallelization, the attention mechanism of transformers, has $O(N^2)$ complexity that brings a major memory bottleneck (Subakan et al. 2022a). For a sequence of length N, the transformer needs to compare N^2 elements which results in a computational bottleneck especially for long signals such as speech. Transformers also use many parameters aggravating the memory problem further. Several versions of transformers such as Longformer (Beltagy et al. 2020), LinFormer (Wang et al. 2020) and Reformer (Kitaev et al. 2020) have been proposed with a goal to reduce the computation complexity of the transformers. Work in Subakan et al. (2022b) investigates the performance of the three versions of transformers in speech separation and concludes that they are suitable for speech separation applications since they achieve a highly favourable trade-off between performance and computational requirements. Work in Luo et al. (2022) proposes a technique of parameter sharing to reduce the computation complexity of the transformer while (Subakan et al. 2022a) reduces complexity by avoiding frame overlap. In Chen et al. (2021), a teacher-student speech separation model based on transformer is proposed. The student model which is much smaller than the teacher model is used to reduce computation complexity. Other transformer-based speech enhancement tools include Wang et al. (2021), de Oliveira et al. (2022). Another key limitation of a transformer is that while it can model long-range global context, they do not extract fine-grained local features patterns well. Based on this, transformer-based speech separation tools apply attention within a frame (chunk) to capture local features and between frames(chunks) to capture global features (Subakan et al. 2021; Qiu and Hu 2022).

6 Model size reduction techniques

To achieve high performance i.e., generate speech with high intelligibility, DNN models for speech enhancements are becoming large by exploiting millions of parameters (Lutati et al. 2022). High number of parameters increase the memory requirements, computation complexity and latency. To reduce these parameters significantly without compromising quality and make speech enhancement tools to work in resource constrained platform, several techniques are being exploited. The techniques include:

Use of dilated convolution To increase the receptive field of 1D CNN and subsequently increase the temporal window and model long range dependencies within a speech, speech separation such as Luo and Mesgarani (2018) and Rethage et al. (2018) implement dilated CNN. Dilated convolution initially introduced by van den Oord et al. (2016) involves a convolution where a kernel is applied to an area that is larger that it. This is achieved by skipping input values by a defined step. It is like implementing a sparse kernel (i.e., dilating the kernel with zeros). When dilated convolution is applied in a stacked network, it enables the network to increase its receptive field with few layers hence minimizing parameters and reducing computation (Lea et al. 2017) (see Fig. 14). This ensures that the models can capture long range dependencies while keeping the number of parameters at minimum. The dilating factors are made to increase exponentially per layer (see Fig. 10).

Parameter quantization To reduce computation, inference complexity of DNN models and to scale down the number of parameters, models such as Wu et al. (2019), Sun and Li (2017), Lin et al. (2019), Fedorov et al. (2020), Hsu et al. (2019) use parameter quantization. In quantization, the objective is to reduce the precision of model parameters and activation values to a low precision with minimal effects on the generalization capability of the DNN model. To achieve this, a quantization operator Q is defined that maps a floating value to a quantized one (Gholami et al. 2022).

Use of depthwise separable convolution This type of convolution, decouples the convolution process into two i.e. depthwise convolution where a single filter is applied to each input channel and pointwise convolution which is applied to the output of depthwise convolution to achieve a linear convolution of the depthwise layer. Depthwise separable convolution has been shown to reduce the number of parameters as compared to the convectional one (Avery et al. 2014; Howard et al. 2017). Speech enhancement tools that exploit depthwise separable convolution include Luo and Mesgarani (2019), Byun and Shin (2021), Zhao et al. (2020).

Knowledge distillation Knowledge distillation involves training a large teacher model which can easily extract the structure of data then the knowledge learned by the teacher is distilled down to a smaller model called the student. Here, the student is trained under the supervision of the teacher (Hinton et al. 2015; Gou et al. 2021; Wang and Yoon 2022). The student model must mimic the teacher and by doing so achieve superior or similar performance but at reduced computation cost due to reduced parameters. Knowledge distillation technique has been exploited to reduce latency in speech enhancement tool (Chen et al. 2021; Aihara et al. 2019).

Parameter pruning In order to reduce the number of parameters and hence speed up computation, some speech enhancement tools use parameter pruning (Fedorov et al. 2020; Tan and Wang 2021). Pruning involves converting a dense DNN model into a sparse one by significantly scaling down the number of parameters without compromising model's output's quality. In Ye et al. (2019) they train a speech enhancement DNN model to obtain an initial parameter set Θ, they then prune the parameters by dropping the weights whose absolute values are below a set pruning threshold. The sparse network is again re-trained to obtain final parameters. Work in Wu et al. (2019) estimates the sparsity $S(k)$ of a given channel F_{jk}. If the sparsity $S(k) > \theta$ where θ is a predefined threshold, the weights within F_{jk} is set to zero and the model is retrained. After several iterations, the channel F_{jk} is dropped.

Weight sharing This involves identifying clusters of weights that have a common value. The clusters are normally identified using K-means algorithm. So instead of storing each weight value, only the indexes of the shared values are stored. Through this memory requirements of the model is reduced (Dupuis et al. 2020). Speech enhancement tools that use weight sharing include Sun and Li (2017), Hu et al. (2021).

7 Objective functions for speech enhancement and separation

Most DNN monaural speech enhancement and separation models especially those working on features in the frequency domain exploit mean-square-error (MSE) as the training objective (Kolbæk et al. 2017a; Luo et al. 2018; Ephraim and Malah 1984; Venkataramani et al. 2018). The DNN models that have the mask as the target use the training objective to minimise the MSE between the estimated mask and the ideal mask target. For models that predict estimated features (such as T-F spectrogram) of the clean source speech, MSE is used to minimise the difference between target features and the estimated features by the model. Despite the dominance of MSE as an objective function in the speech enhancement tools, it has been criticised since it is not closely related to human auditory perception (Fu et al. 2019). Its major weakness is that it treats estimation elements independently and equally. For instance, it treats each time-frequency unit separately rather than whole spectral (Xu et al. 2015). This leads to muffled sound quality and compromises intelligibility (Xu et al. 2015). MSE also treats every estimation element with equal importance which is not the case (Zhang et al. 2018). It also does not discriminate between the positive or negative differences between the clean and estimated spectra. A positive difference between the clean and estimated spectra represents attenuation distortion, while a negative spectral difference represents amplification distortion. MSE treats the effects of these two distortions on speech intelligibility as equivalent which is problematic (Loizou and Kim 2011; Loizou 2013). Moreover, the MSE is usually defined in the linear frequency scale while the human auditory perception is on the Mel-frequency scale. To avoid the problem of treating every estimation element with equal importance (Xia and Bao 2014; Shivakumar and Georgiou 2016) propose a weighted MSE. Due to the shortcomings of MSE, objective functions that are closely related to the human auditory perception have been introduced to train the DNN (Zhang et al. 2018; Koizumi et al. 2017; Kolbcek et al. 2018; Yan et al. 2018; Chen et al. 2015). Some of the human auditory perception training objectives being used by speech enhancement tools are also used as metrics for perceptual evaluation. They include:

1. Short-time objective intelligibility (STOI) (Taal et al. 2011).
2. Scale invariant signal-to-distortion ratio (SI-SDR) (Roux et al. 2019).
3. Perceptual metric for speech quality evaluation (PMSQE).

Scale invariant signal-to-distortion ratio Work in Roux et al. (2019) proposes an intelligibility measure such that given the target signal s and the model estimated signal $\hat{s}$, they re-scale either s or $\hat{s}$ such that the residual $(s - \beta\hat{s})$ after scaling $\hat{s}$ or $(\alpha s - \hat{s})$ after scaling the target s is orthogonal to the target as: $(s - \beta\hat{s}).s = 0$ or $(\hat{s} - \alpha s).s = 0$ based on this,α can be computed as:

$$\begin{aligned}(\hat{s} - \alpha s).s &= 0\\ \hat{s}.s - \alpha s.s &= 0\\ \alpha &= \frac{\hat{s}^T s}{s.s}\end{aligned}$$

based on scaling of the target s, the signal-to-noise ratio(SNR) equation

$$SDR = 10\log_{10}\frac{||s||^2}{||s - \hat{s}||^2} \tag{48}$$

is transformed to:

$$SDR = 10\log_{10}\frac{||\alpha s||^2}{||s - \hat{\alpha} s||^2} \tag{49}$$

replacing the α we get the SI-SDR:

$$SI - SDR = 10\log_{10}\frac{||\frac{\hat{s}^T s}{||s^2||}s||^2}{||\frac{\hat{s}^T s}{||s^2||}s - \hat{s}||^2} \tag{50}$$

This objective function has been used in Subakan et al. (2021), Luo and Mesgarani (2019), Bahmaninezhad et al. (2019), Nachmani et al. (2020), Subakan et al. (2022a), Li et al. (2022), Fan et al. (2020), Byun and Shin (2021), Lee et al. (2022), Lutati et al. (2022).

Short-time objective intelligibility This objective has been used in Kolbcek et al. (2018), Yan et al. (2018), Zhang et al. (2018), Fu et al. (2019). STOI Taal et al. (2011), Taal et al. (2010) is a speech intelligibility measure that is achieved by executing the following steps:

1. Given discrete time signals of clean speech signal $x(n)$ and enhanced speech $y(n)$, perform a DFT on both $x(n)$ and $y(n)$ i.e $X(n, k) = DFT(y(n))$ and $Y(n, k) = DFT(y(n))$. Here, k refers to the index of the discrete frequency.
2. Remove silences in both the clean signal and the enhanced signals. Silences are removed by first identifying the frame with maximum energy(max_{energy}) in the clean signal. All frames with energy of 40 dB less than max_{energy} are dropped.
3. Reconstruct both the clean and enhanced speech signals.
4. Perform a one-third band octave analysis on both clean and enhanced speech by grouping DFT bins i.e the complex-valued STFT coefficients, $X(n, k)$, are combined into J third-octave bands by computing the TF units.

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} |X(n,k)|^2 j = 1, \dots, J} \tag{51}$$

   Here, k_1 and k_2 represent the one-third octave band edges. The same octave analysis is performed on the enhanced speech. The one-third octave of the enhanced speech is defined in a similar manner.
5. Define a short temporal envelope of both enhanced and clean speech as: $Y_{j,m} = [Y_j(m-N+1), Y_j(m-N+2), \dots, X_j(m)]^T$ and $X_{j,m} = [X_j(m-N+1), X_j(m-N+2), \dots, X_j(m)]^T$ respectively. STOI exploits correlation coefficients to compare the temporal envelopes of clean and enhanced speech for a short time region. Note that N=30.
6. Normalise the short temporal envelopes of the enhanced speech. Let $y_{j,m}(n)$ denote the n^{th} envelope of enhanced speech. The normalised enhanced speech $y'_{j,m}(n)$ of $Y_{j,m}(n)$ is given by $y'_{j,m}(n) = \frac{|X_{j,m}|}{||Y_{j,m}||} Y_{j,m}(n)$. ||.|| is the l_2 norm. The intuition behind normalisation of the enhanced speech is to reduce global level differences between clean and enhanced speech. These global level differences should not have a strong effect on speech intelligibility.
7. Clip the normalised enhanced speech as $\bar{y}_{j,m}(n) = \min(y'_{j,m}, (1 + 10^{\frac{\beta}{20}})x_{j,m}(n))$. Clipping is done to ensure the effects of severely degraded frames of the enhanced speech on the model is upper bounded. Here, $\beta = -15dB$ is the lower signal-to-distortion(SDR) bound.
8. Compute intermediate intelligibility measure as

$$d_{j,m} = \frac{(x_{j,m} - \mu_{x_{j,m}})^T (y_{j,m} - \mu_{y_{j,m}})}{||x_{j,m} - \mu_{x_{j,m}}|| y_{j,m} - \mu_{y_{j,m}}||} \tag{52}$$

   Here,$\mu_{(.)}$ refers to the sample mean of the corresponding vector.
9. Compute the average intermediate intelligibility of all frames as

$$d = \frac{1}{JM} \sum_{j,m} d_{j,m} \tag{53}$$

   where M represents the total number of frames and J the number of one-third octave band.

Short-time spectral amplitude mean square error. Let $X[n, k]$ with $1 \leq n \leq N$ and $1 \leq k \leq K$ be an N point DFT of x and K is the number of frames. Let $A[n, k] = |X[n, k]|$ with $k, \dots, \frac{N}{2} + 1$ and $k = 1, \dots, K$ denote the single sided amplitude spectra of $X[n, k]$.

Let $\hat{A}[n, k]$ be an estimate of $A[n,\ k]$, the short-time spectral amplitude mean square error(STSA-MSE) is given by

$$\mathcal{L}_{STSA-MSE} = \frac{1}{(N/2+1)K} \sum_{n=1}^{N=K/2+1} \sum_{k=1}^{k=K} (\hat{A}[n,k] - A[n,k])^2 \tag{54}$$

Equation (54) represents the mean square error between single-sided amplitude spectra of the clean speech x and the DNN estimated speech $\hat{x}$. Equation (54) is not sensitive to the phase spectrum of the two signals. The objective function has been utilized in Kolbaek et al. (2018), Kolbaek et al. (2020).

Perceptual metric for speech quality evaluation(PMSQE). This is an objective function that is based on the adaptation of perceptual evaluation of speech quality (PESQ) algorithm (Rix et al. 2001). Given the MSE loss in the log-power spectrum with mean and variance normalisation i.e.

$$\begin{aligned} MSE_t =& \frac{1}{k} \sum_{n=1}^{K} \left(\frac{\log |x[n,k]|^2 - \mu_k}{\delta_k} - \frac{\log |\hat{X}[n,k]|^2 - \mu_k}{\delta_k} \right)^2 \\ =& \frac{1}{K} \sum_{k=1}^{k=K} \frac{1}{\delta_k^2} \left(\log \frac{|X[n,k]|^2}{|\hat{X}[n,k]|^2} \right)^2 \end{aligned} \tag{55}$$

Here, $X[n,k]|^2$ and $\hat{X}[n,k]|^2$ represent the power spectra of the clean and enhanced speech respectively. μ_k is the mean log-power spectrum and δ_k is its standard deviation. The indices n and k represent the frame and frequency, while K is the number of frequency bins. From Eq. (55), the MSE is entirely dependent on power spectra across frequency bands hence not factoring in the perceptual factors such as loudness difference, masking and threshold effects (Martin-Donas et al. 2018). To factor in the perceptual factors in the MSE, PMSQE modifies the MSE loss by incorporating two disturbance terms (symmetrical disturbance and asymmetrical disturbance) which are based on the PESQ algorithm both computed in a frame-by-frame basis (Martin-Donas et al. 2018).

$$MSE_t = \sum_t MSE_t + \alpha D_t^s + \beta D_t^a \tag{56}$$

Here D_t^s and D_t^a represent symmetrical and asymmetrical disturbances respectively. The parameters α and β are weighting factors which are determined experimentally. Work in Martin-Donas et al. (2018) describes how to arrive at the values of D_t^s and D_t^a. Since PESQ is non-differentiable, the PMSQE objective function provides a way of estimating it. PMSQE objective function is designed to be inversely proportional to PESQ, such that a low PMSQE value corresponds to a high PESQ value and vice versa. The key question here is: Which objective function is superior? work in Kolbaek et al. (2020) tries to answer this question where they evaluate six objective functions. Their conclusion is that the evaluation metric should be a major factor in deciding on the objective function to use in the speech enhancement model. In case a given model targets to improve a specific evaluation metric, then selection of an objective function related to that metric is advantageous.

8 Unsupervised techniques for speech enhancement

Although supervised techniques of speech enhancement and separation have achieved great success towards improving speech intelligibility, the inherent problems associated with supervised learning still prohibits their applications in all scenarios. First, collecting parallel data of clean and noisy (mixed) data remains costly and time consuming. This limits the amount of data that can be used to train these models. Consequently, the models are not exposed to enough variations of the recording during training hence affecting their generalizability to noise types and acoustic conditions that were not seen during training (Bie et al. 2022; Fujimura et al. 2021). Collecting clean audio is always difficult and requires a well-controlled studio exacerbating the already high cost of data annotation Fujimura et al. (2021). Unsupervised learning offers an alternative to solving these problems. The existing unsupervised techniques for speech enhancement and separation can roughly be categorised into three: MixIT based techniques, generative modelling technique and teacher-student based techniques. Few novel techniques have also been proposed that fall outside these three dominant categories. Work in Wisdom et al. (2020) proposes mixture invariant training (MixIT) to perform unsupervised speech separation. Given a set of X that is composed of mixed speech i.e. $X = \{x_1, x_2, \dots, x_n\}$ where each mixture x_i is composed of up to N sources, mixtures are drawn at random from the set X without replacement and a mixture of mixture (MoM) created by adding the drawn mixtures, for example if two mixtures x_1 and x_2 are drawn from the set X, MoM $\bar{x}$ is created by adding x_1 and x_2 i.e $\bar{x} = x_1 + x_2$. The MoM $\bar{x}$ is the input to a DNN model which is trained to estimate sources $\hat{s}$ composed in x_1 and x_2. The DNN model is trained to minimize the loss function in Eq. (57).

$$L_{MixIT} = \min_{A} \sum_{i=1}^{2} L(x_i, [A\hat{s}]_i) \tag{57}$$

For a case where MoM is composed of only two mixtures, $A \in B^{2\times M}$ is a set of binary matrices where each column sums to 1. The loss function is trained to minimize the loss between mixtures x_i and the remixed separated sources $A\hat{s}$. MixIT has been criticised for over-separation where it outputs estimated sources greater than the actual number of underlying sources in the mixtures x_i (Zhang et al. 2021b). Further, MixIT does not work well for speech enhancement (i.e., denoising) (Saito et al. 2021). MixIT teacher-student unsupervised model has been proposed in Zhang et al. (2021b) to tackle the problem of over-separation in MixIT. It trains a student model such that its output matches the number of sources in the mixed speech x. Another MixIT based technique for solving over-separation problem is discussed in MixCycle (Karamatlı and Kırbız 2022). Work in Trinh and Braun (2022) proposes to improve MixIT to make it more tailored for denoising by exploiting an ASR pre-trained model to modify MixIT's loss function. Work in Saito et al. (2021) also seeks to improve MixIT for denoising by improving loss function and noise augmentation scheme. RemixIT (Tzinis et al. 2022) is an unsupervised speech denoising tool that exploits teacher-student DNN model. Given a batch of noisy speeches of size b, the teacher estimates the clean speech sources $\hat{s}_i$ and noises $\hat{n}_i$ where $1 \leq i \leq b$. The teacher estimated noises $\hat{n}_i$ are mixed at random to generate n^p. The mixed noise n^p together with the teacher estimated sources are used to generate new mixtures $\hat{m}_i = \hat{s}_i + n^p$. The new mixtures $\hat{m}_i$ are used as input to the student. The student is optimised to generate $\hat{s}_i$ and noise n^p i.e., $\hat{s}_i$ and n^p are the targets. Through this the teacher-student model is trained to denoise the speech. In RemixIT, a pre-trained speech enhancement model is used as the teacher model. Motivated by RemixIT, Chen et al. (2023b) also proposes a speech denoising unsupervised tool using teacher-student DNN model. They propose various techniques of student training. MetricGAN-U (Fu et al. 2022) is an unsupervised GAN based speech enhancement tool that trains a conditioned GAN discriminator without a reference clean speech, MetricGAN-U employs objective in Eq. (58) to train the speech enhancement model.

$$L = \mathbb{E}_x[(D(G(x)) - Q'(G(x)))^2 + (D(x) - Q'(G(x))^2] \tag{58}$$

In Eq. (58), $Q\prime$ is a non-intrusive metric (i.e does not require reference clean speech) that is used to score the enhanced speech from the generator. The scores obtained by $Q\prime$ are used to optimize the model. In MetricGAN-U, DNSMOS (Reddy et al. 2021) is used as $Q\prime$. Another GAN based technique for unsupervised learning is Xiang and Bao (2020) which exploits CycleGAN (Zhu et al. 2017) multi-objective learning to perform parallel-data-free speech enhancement. Tools in Bie et al. (2022) and Li et al. (2021b) propose unsupervised speech denoising technique based on variations of VAE. Work in Fujimura et al. (2021) proposes a speech denoising technique that uses only the noisy speech. It exploits the idea that was first proposed in Lehtinen et al. (2018) where they demonstrated that it is possible to recover signals under corruptions without observing clean signals. Predicating their work on thesefindings, given a noisy speech signal x, and noise n, Fujimura et al. (2021) creates a more noisy speech $y = x + n$. They then train a DNN model to predict an enhanced speech $\hat{s}$ by having the more noisy input y as the input and noisy speech x as the target. Consequently, the DNN is trained by minimizing the loss in Eq. (59).

$$L = \frac{1}{M}\sum_{i=1}^{M} D(\hat{s}_m, x_m) \tag{59}$$

Here, D is the objective function and M is the sample size. This technique works on the basis that DNN cannot predict random noise hence the noise component in the training data is mapped to their expected values. Therefore, by assuming the noise as zero mean random variable, the objective function eliminates the noise (Fujimura et al. 2021). Work in Wang et al. (2016) proposes unsupervised techniques to perform speech separation based on gender. They exploit i-vectors to model large discrepancy in vocal tract, fundamental frequency contour, timing, rhythm, dynamic range, etc between speakers of different genders. In this case DNN model can be viewed as gender separator.

9 Domain adaptive techniques for speech enhancement and separation

Training data used to train speech enhancement and separation tools mostly have acoustic features that are significantly different from the acoustic features of the speech signals where the tools are deployed. This mismatch between the training data and target data leads to degradation in the tool's performance in their deployed environment (Pan et al. 2010).The target environment dataset's acoustic features may vary from the training data in noise type, speaker and signal-noise-ratio (Li et al. 2021b). One potential way of tackling this problem is to collect massive training data that covers different variation of deployment environment. However, this is mostly not possible due to prohibitive cost. Due to this, some tools are proposing DNN based techniques for domain adaptation. Domain adaptation seeks to exploit either labelled or unlabelled target domain data to transfer a given tool from training data domain to the target data domain. Basically, domain adaptation seeks to reduce the covariance shift between the source and target domains data. The domain adaptation techniques in literature for speech separation and enhancement tools can be categorised into two: Unsupervised domain adaptation techniques such as Liao et al. (2018), Wang et al. (2018d) which use unlabelled target domain dataset to adapt a DNN model and supervised domain adaptation techniques such as Xu et al. (2014c), Li et al. (2021b), Pascual et al. (2018) which exploit limited labelled target domain dataset to perform domain adaptation of a DNN model for speech enhancement or separation. To make speech enhancement tools portable to new a new language, Xu et al. (2014c) proposes to use transfer learning. Transfer learning entails tailoring trained DNN models to apply knowledge acquired during training to a new domain where there is some commonality in type of task. The tool fine-tunes the top layers of a trained DNN model for speech enhancement by using labelled data of a new language while freezing the lower layer which are composed of parameters acquired during training of the original language. Work in Pascual et al. (2018) also uses transfer learning to show that pre-trained SEGAN can achieve high performance in new languages with unseen speakers and noise with just short training time. To make it more adaptable to different types of noise, tool in Li et al. (2021b) proposes to employ multiple encoders where each encoder is trained in a supervised manner to focus only on given acoustic feature. The features are categorized into two i.e., utterance-level features such as gender, age, ascent of the speaker, signal-to-noise ratio and noise type and the signal-level features such high and low frequency of the speech parts. Feature focused encoders are trained to learn how to extract a given feature representation such as gender representation composed in the speech. Through the feature focused encoders, the experimental results show that the tool can adapt more to unseen noise types as compared to using a single global encoder. To adapt the DNN speech enhancement model to unseen noise type, work in Liao et al. (2018) utilizes domain adversarial training (DAT) (Ganin et al. 2016) to train an encoder to extract noise-invariant features. To do this, it utilizes the labelled source data and unlabelled target data. Through the feedback from the discriminator which gives the probability distribution over multiple noise types, the encoder is trained to produce noise-invariant features, hence reducing the mismatch problem. Work in Wang et al. (2018d) also exploits unsupervised DAT for speaker mismatch resolution. Work in Li et al. (2021b) exploits importance-weighting (IW) using the classifiers of the networks to classify the source domain samples from the outlier weights and hence reduces the shift between the source and target domain.

10 Use of pre-trained models in speech separation and enhancement

Pre-trained models have become popular especially in Natural language processing(NLP) and Computer vision. In NLP, for example, large corpus of text can be used to learn universal language representations which are beneficial for downstream NLP tasks. Due to their success in domains such as NLP and computer vision, pre-trained models based on unsupervised learning have been introduced in audio data (Chung et al. 2019, 2020; Liu et al. 2020, 2021; Baevski et al. 2020; Hsu et al. 2021). Such pre-trained models are beneficial in several ways:

1. Pre-trained models are trained in large speech dataset hence can learn universal speech representations which can be beneficial to speech separation by boosting the quality of enhanced speech generated by these models.
2. Pre-trained models provide models with better initialization which can result in better generalization and speed up convergence during training of speech enhancement models.
3. Pre-trained speech models can act as regularizers to help speech enhancements models to avoid over fitting (Erhan et al. 2010).

Work in Huang et al. (2022) seeks to establish if pre-trained speech models will help generate more robust features for downstream speech denoising and separation task as compared to features established without pre-trained models. To do this they use 13 speech pre-trained models to generate features of a noisy speech which are then passed through a three-layer BLSTM network which generates speech denoising or separation mask. They compare the performance of these features with those of baseline STFT and mel filerbank (FBANK) features. Their experiments establish that the 13 pre-trained models used do not significantly improve feature representations as compared to those of baselines. Hence the quality of enhanced and separated speech generated by features of pre-trained models are only slightly better or worse in some cases as compared to those generated based on the baseline features. They attribute this to domain mismatch and information loss. Since most of the pre-trained models were trained with clean speech, they are not portable to a noisy speech domain. Pre-trained models are usually trained to model global features and long-term dependencies hence some local features of the noisy or mixed speech signal may be lost due to this during feature extraction. Using HuBERT Large model (Hsu et al. 2021), they demonstrate that the last layer of the model does not produce the best feature representation for speech enhancement and separation. In fact, for speech separation, the higher layers features are of low quality as compared to lower layers. They show that the weighted-sum representations of the representations from the different layers of pre-trained models where lower layers are given more weight generate better speech the enhancement and separation results as compared to isolated layers representations. They hypothesise that this could be due loss of some local signal information necessary for speech reconstruction tasks in deeper layers. To address the problem of information loss in pre-trained model, Hung et al. (2022) proposes two solutions, first they utilize cross-domain features as model inputs to compensate lost information and secondly, they fine-tune a pre-trained model by using a speech enhancement model such that the extracted features are more tailored towards speech enhancement. Research in Irvin et al. (2023) seeks to synthesise clean speech directly from a noisy one using pre-trained model and HiFiGAN (Kong et al. 2020) speech synthesis model. It exploits the pre-trained model to extract features of the noisy model. The features are then used as input of HiFiGAN which generates estimated clean speech from these features. Based on the results reported in Huang et al. (2022) that demonstrated that the final layer of pre-trained model does not give optimized representation for speech enhancement, they exploit weighted average of the representations of all the layers of the pre-trained model to generate representations of the noisy speech. The novelty of this work is that they do not use a model dedicated for speech denoising rather show that given features of a noisy speech, speech synthesis model can perform denoising. In Germain et al. (2018), a pre-trained model has been exploited to design the loss function. Given a pre-trained model Φ with m layers, the tool uses weighted L^1 loss to compute the difference between the feature activations of clean and noisy speech generated by different layers of the pre-trained model according to Eq. (60).

$$L = \sum_{i=1}^{m} \lambda_w ||\Phi_m(s) - \Phi_m(g_\theta(x))|| \tag{60}$$

Here, s and x are the clean and noisy speech respectively, g_θ is the denoising DNN model and λ_m are the weights of contribution of each pre-trained model layer features to the loss function. Work in Hao et al. (2023) proposes a two-stage speech enhancement framework where they first pre-train a model using unpaired noisy and clean data and utilize the pre-train model to perform speech enhancement. Unlike the previous works that use general pre-trained models for audio, the pre-trained model in Hao et al. (2023) is trained using speech enhancement dataset. They report state of the art results in speech enhancement and ability of the tool to generalize to unseen noise.

11 Future direction for research in speech enhancement

Unsupervised techniques for speech separation Majority of speech enhancements tools use supervised learning technique. For those that use unsupervised learning discussed in Sect. 8, they almost entirely focus on speech denoising and not speaker separation and dereverberation. There is therefore a gap of extending the unsupervised DNN techniques to perform multi-speaker speech separation and dereverberation.

Dimension mismatch problem Most speech separation tools set a fixed number C of speakers therefore cannot deal with an inference mixture with K sources, where $C \neq K$. Currents tools deal with the dimension mismatch problem by either outputting silences if $C > K$ or performing speech separation through iteration. However, both techniques have been found to be inefficient (see discussion in Sect. 2.1). Therefore, there need to explore on developing DNN techniques for speech separation which are dynamic to the number of speakers present in inference mixture and adapt appropriately.

Focus on data Most model compression techniques for speech enhancement speed up the model performance by reducing or optimizing the model parameters. They don't focus on the data-side impact on the model performance. For instance, what is the ideal sequence length and chunk overlap when working in time-domain that can speed up the speech enhancement process without compromising the quality of enhancement. More focus needs to turn towards exploring the data modifications that can speech up the speech enhancement process.

Dataset modification In dereverberation, tools are beginning to explore the use of speech with early reverberation (Valin et al. 2022) as the target as opposed to using anechoic target. Experiments in Valin et al. (2022) demonstrate that allowing early reverberation in the target speech improves the quality of enhanced speech. There is need to develop a standardized dataset where the target is composed of early reverberation to allow for standardized evaluation of the tools on this dataset. *Pre-trained model* The pre-trained models that have been utilized for speech enhancement or separation have been trained on clean dataset hence failing potability test when used to generate features of a noisy speech signal (Huang et al. 2022). There is need for development of a pre-trained model tailored for speech separation and enhancement.

12 Conclusion

This review gives a discussion on how DNN techniques are being exploited by monaural speech enhancement tools. The objective was to uncover the key trends and dominant techniques being used by DNN tools at each stage of speech enhancement process. The review therefore discusses the type of features being exploited by these tools, the modelling of speech contextual information, how the models are trained (both supervised and unsupervised), key objective functions, how pre-trained speech models are being utilized and dominant dataset for evaluating the performance of the speech enhancement tools. In each section we highlight the standout challenges and how tools are dealing with these challenges. Our target is to give an entry into the speech enhancement domain by getting a thorough overview of the concepts and the trends of the domain. The hope is that the review gives a snapshot of the current research on DNN application to speech enhancement.

Author Contributions All the contents of the paper was written by PO.

Declarations

Competing interests The authors declare no competing interests.

References

Adiga N, Pantazis Y, Tsiaras V, Stylianou Y (2019) Speech enhancement for noise-robust speech synthesis using wasserstein gan. In: INTERSPEECH, pp 1821–1825

Aihara R, Hanazawa T, Okato Y, Wichern G, Roux JL (2019) Teacher-student deep clustering for low-delay single channel speech separation. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2019-May, pp 690–694

Ai Y, Li H, Wang X, Yamagishi J, Ling Z (2021) Denoising-and-dereverberation hierarchical neural vocoder for robust waveform generation. In: 2021 IEEE spoken language technology workshop, SLT 2021—proceedings, pp 477–484

Allen JB (1982) Applications of the short time Fourier transform to speech processing and spectral analysis. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 1982-May, pp 1012–1015

Allen JB, Rabiner LR (1977) A unified approach to short-time fourier analysis and synthesis. Proc IEEE 65(11):1558–1564

Arweiler I, Buchholz JM (2011) The influence of spectral characteristics of early reflections on speech intelligibility. J Acoust Soc Am 130(2):996–1005

Avery KR, Pan J, Engler-Pinto CC, Wei Z, Yang F, Lin S, Luo L, Konson D (2014) Fatigue behavior of stainless steel sheet specimens at extremely high temperatures. SAE Int J Mater Manuf 7(3):560–566

Baby D, Virtanen T, Barker T, Van Hamme H (2014) Coupled dictionary training for exemplar-based speech enhancement. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, pp 2883–2887

Baevski A, Zhou H, Mohamed A, Auli M (2020) wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv Neural Inf Process Syst 1:1–19

Bahmaninezhad F, Wu J, Gu R, Zhang SX, Xu Y, Yu M, Yu D (2019) A comprehensive study of speech separation: spectrogram vs waveform separation. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2019-September, pp 4574–4578

Bai S, Kolter JZ, Koltun V (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint arXiv:1803.01271

Bando Y, Mimura M, Itoyama K, Yoshii K, Kawahara T (2018) Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization, ICASSP, In: IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2018-April, no. Mcmc, pp 716–720

Beltagy I, Peters ME, Cohan A (2020) Longformer: the long-document transformer, arXiv preprint http://arxiv.org/abs/2004.05150

Bie X, Leglaive S, Alameda-Pineda X, Girin L (2022) Unsupervised speech enhancement using dynamical variational autoencoders. IEEE/ACM Trans Audio Speech Lang Process 30:2993–3007

Brungart DS, Chang PS, Simpson BD, Wang D (2006) Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation, pp 4007–4018

Byun J, Shin JW (2021) Monaural speech separation using speaker embedding from preliminary separation. IEEE/ACM Trans Audio Speech Lang Process 29:2753–2763

Cao R, Abdulatif S, Yang B (2022) CMGAN: conformer-based metric GAN for speech enhancement, arXiv preprint arXiv:2209.11112, pp 936–940

Chandna P, Miron M, Janer J, Gómez E (2017) Monoaural audio source separation using deep convolutional neural networks, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 10169 LNCS, pp 258–266

Chang X, Zhang W, Qian Y, Le Roux J, Watanabe S (2020) End-to-end multi-speaker speech recognition with transformer. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2020, pp 6134–6138

Chen Z, Watanabe S, Erdogan H, Hershey JR (2015) Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2015-January, 2015, pp 3274–3278

Chen Z, Luo Y, Mesgarani N (2017) Deep attractor network for single-microphone speaker separation. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 246–250. IEEE

Chen J, Mao Q, Liu D (2020) Dual-path transformer network: direct context-aware modeling for end-to-end monaural speech separation. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2020-October, pp 2642–2646

Chen S, Wu Y, Chen Z, Wu J, Yoshioka T, Liu S, Li J, Yu X (2021) Ultra fast speech separation model with teacher student learning. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 3, pp 2298–2302

Chen L-W, Cheng Y-F, Lee H-S, Tsao Y, Wang H-M (2023a) A training and inference strategy using noisy and enhanced speech as target for speech enhancement without clean speech. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, pp 5315–5319

Chen L, Mo Z, Ren J, Cui C, Zhao Q (2023b) An electroglottograph auxiliary neural network for target speaker extraction. Appl Sci 13(1):469

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP 2014-2014 conference on empirical methods in natural language processing, proceedings of the conference, pp 1724–1734

Choi H-S, Heo H, Lee JH, Lee K (2020) Phase-aware single-stage speech denoising and dereverberation with U-Net. arXiv preprint arXiv:2006.00687

Chung YA, Hsu WN, Tang H, Glass J (2019) An unsupervised autoregressive model for speech representation learning. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2019-September, pp 146–150

Chung YA, Tang H, Glass J (2020) Vector-quantized autoregressive predictive coding. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2020-October, no. 1, pp 3760–3764

Cord-Landwehr T, Boeddeker C, von Neumann T, Zorila C, Doddipatla R, Haeb-Umbach R (2021) Monaural source separation: from anechoic to reverberant environments. In: 2022 international workshop on acoustic signal enhancement (IWAENC), pp 1–5. arXiv:org/abs/2111.07578

de Oliveira D, Peer T, Gerkmann T (2022) Efficient transformer-based speech enhancement using long frames and STFT magnitudes, arXiv preprint arXiv:2206.11703., no. 1, pp 2948–2952

Défossez A, Synnaeve G, Adi Y (2020) Real time speech enhancement in the waveform domain. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2020-October, pp 3291–3295

Delcroix M, Zmolikova K, Kinoshita K, Ogawa A, Nakatani T (2018) Single channel target speaker extraction and recognition with speaker beam. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5554–5558. IEEE

Donahue C, Li B, Prabhavalkar R (2018) Exploring speech enhancement with generative adversarial networks for robust speech recognition. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2018-April, no. Figure 1, pp 5024–5028

Dovrat S, Nachmani E, Wolf L (2021) Many-speakers single channel speech separation with optimal permutation training. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 4, pp 2408–2412

Du J, Huo Q (2008) A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, pp 569–572

Du J, Tu Y, Xu Y, Dai L, Lee CH (2014) Speech separation of a target speaker based on deep neural networks. In: International conference on signal processing proceedings, ICSP, vol 2015-January, no. October, pp 473–477

Du Z, Zhang X, Han J (2020) A joint framework of denoising autoencoder and generative vocoder for monaural speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 28:1493–1505

Dupuis E, Novo D, O'Connor I, Bosio A (2020) Sensitivity analysis and compression opportunities in DNNs using weight sharing. In: Proceedings—2020 23rd international symposium on design and diagnostics of electronic circuits and systems, DDECS 2020

Ephraim Y, Malah D (1984) Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process 32(6):1109–1121

Erdogan H, Hershey JR, Watanabe S, Le Roux J (2015) Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2015-August, pp 708–712

Erhan D, Courville A, Bengio Y, Vincent P (2010) Why does unsupervised pre-training help deep learning? In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, pp 201–208

Fan C, Tao J, Liu B, Yi J, Wen Z, Liu X (2020) End-to-end post-filter for speech separation with deep attention fusion features. IEEE/ACM Trans Audio Speech Lang Process 28:1303–1314

Fedorov I, Stamenovic M, Jensen C, Yang LC, Mandell A, Gan Y, Mattina M, Whatmough PN (2020) TinyLSTMs: efficient neural speech enhancement for hearing aids. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2020-October, pp 4054–4058

Friedman DH (1985) Instantaneous-frequency distribution vs. time: an interpretation of the phase structure of speech. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, pp 1121–1124

Fu SW, Tsao Y, Lu X (2016) SNR-aware convolutional neural network modeling for speech enhancement. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 08-12-Sept, pp 3768–3772

Fu SW, Hu TY, Tsao Y, Lu X (2017) Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In: IEEE international workshop on machine learning for signal processing, MLSP, vol 2017-September, pp 1–6

Fu SW, Wang TW, Tsao Y, Lu X, Kawai H, Stoller D, Ewert S, Dixon S, Lu X, Tsao Y, Matsuda S, Hori C, Xu Y, Du J, Dai LR, Lee CH, Gao T, Du J, Dai LR, Lee CH, Fu SW, Tsao Y, Lu X, Weninger F, Hershey JR, Le Roux J, Schuller B, Xu Y, Du J, Dai LR, Lee CH, Lluís F, Pons J, Serra X (2018a) Speech enhancement based on deep denoising autoencoder. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 08-12-Sept, no. 1, pp 7–19

Fu SW, Wang TW, Tsao Y, Lu X, Kawai H (2018b) End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. IEEE/ACM Trans Audio Speech Lang Process 26(9):1570–1584

Fu SW, Liao CF, Tsao Y, Lin SD (2019) MetricGAN: generative adversarial networks based black-box metric scores optimization for speech enhancement. In: 36th international conference on machine learning, ICML 2019, vol 2019-June, pp 3566–3576

Fu S-W, Yu C, Hung K-H, Ravanelli M, Tsao Y (2022) Metricgan-u: unsupervised speech enhancement/dereverberation based only on noisy/reverberated speech. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 7412–7416. IEEE

Fujimura T, Koizumi Y, Yatabe K, Miyazaki R (2021) Noisy-target training: a training strategy for DNN-based speech enhancement without clean speech. In: 2021 29th european signal processing conference (EUSIPCO), pp 436–440. IEEE

Gamper H, Tashev IJ (2018) Blind reverberation time estimation using a convolutional neural network. In: 16th international workshop on acoustic signal enhancement, IWAENC 2018—proceedings, pp 136–140

Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M, Lempitsky V (2016) Domain-adversarial training of neural networks. J Mach Learn Res 17(1):2030–2096

Gannot S, Vincent E, Markovich-Golan S, Ozerov A (2017) A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Trans Audio Speech Lang Process 25(4):692–730

Gao T, Du J, Dai LR, Lee CH (2016) SNR-based progressive learning of deep neural network for speech enhancement. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 08-12-Sept, pp 3713–3717

Garau G, Renals S (2008) Combining spectral representations for large-vocabulary continuous speech recognition. IEEE Trans Audio Speech Lang Process 16(3):508–518

Germain FG, Chen Q, Koltun V (2018) Speech denoising with deep feature losses, arXiv preprint arXiv:1806.10522

Gholami A, Kim S, Dong Z, Yao Z, Mahoney MW, Keutzer K (2022) A survey of quantization methods for efficient neural network inference, low-power computer vision, pp 291–326

Goodfellow I (2016) NIPS 2016 Tutorial: generative adversarial networks, arXiv preprint arXiv. arXiv:org/abs/1701.00160

Gou J, Yu B, Maybank SJ, Tao D (2021) Knowledge distillation: a survey. Int J Comput Vis 129:1789–1819

Grais EM, Plumbley MD (2018) Single channel audio source separation using convolutional denoising autoencoders. In: 2017 IEEE global conference on signal and information processing, GlobalSIP 2017—proceedings, vol 2018-Janua, pp 1265–1269

Grais EM, Sen MU, Erdogan H (2014) Deep neural networks for single channel source separation. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, pp 3734–3738

Griffin DW, Lim JS (1984) Signal estimation from modified short-time fourier transform. IEEE Trans Acoust Speech Signal Process 32(2):236–243

Gulati A, Qin J, Chiu CC, Parmar N, Zhang Y, Yu J, Han W, Wang S, Zhang Z, Wu Y, Pang R (2020) Conformer: convolution-augmented transformer for speech recognition. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2020-October, pp 5036–5040

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC (2017) Improved training of wasserstein gans. Adv Neural Inf Process Syst 30

Gunawan D, Sen D (2010) Iterative phase estimation for the synthesis of separated sources from single-channel mixtures. IEEE Signal Process Lett 17(5):421–424

Han K, Wang Y, Wang DL, Woods WS, Merks I, Zhang T (2015) Learning spectral mapping for speech dereverberation and denoising. IEEE Trans Audio Speech Lang Process 23(6):982–992

Han C, O'Sullivan J, Luo Y, Herrero J, Mehta AD, Mesgarani N (2019) Speaker-independent auditory attention decoding without access to clean speech sources. Sci Adv 5(5):1–12

Hao X, Xu C, Xie L (2023) Neural speech enhancement with unsupervised pre-training and mixture training. Neural Netw 158:216–227

He Y, Zhao J (2019) Temporal convolutional networks for anomaly detection in time series. J Phys 1213(4):042050

Heitkaemper J, Jakobeit D, Boeddeker C, Drude L, Haeb-Umbach R (2020) Demystifying TasNet: a dissecting approach. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2020-May, pp 6359–6363

Hershey JR, Chen Z, Le Roux J, Watanabe S (2016) Deep clustering: discriminative embeddings for segmentation and separation. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 31–35. IEEE

Hien TD, Tuan DV, At PV, Son LH (2015) Novel algorithm for non-negative matrix factorization. New Math Nat Comput 11(2):121–133

Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 2.7, pp 1–9

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) MobileNets: efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861

Hsu YT, Lin YC, Fu SW, Tsao Y, Kuo TW (2019) A study on speech enhancement using exponent-only floating point quantized neural network (EOFP-QNN). In: 2018 IEEE spoken language technology workshop, SLT 2018—proceedings, pp 566–573

Hsu WN, Bolte B, Tsai YHH, Lakhotia K, Salakhutdinov R, Mohamed A (2021) HuBERT: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans Audio Speech Lang Process 29:3451–3460

Hu X, Li K, Zhang W, Luo Y, Lemercier JM, Gerkmann T (2021) Speech separation using an asynchronous fully recurrent convolutional neural network. Adv Neural Inf Process Syst 27:22509–22522

Huang P-S, Kim M, Hasegawa-Johnson M, Smaragdis P (2011) Deep learning for monaural speech separation. Acta Phys Pol B 42(1):33–44

Huang PS, Kim M, Hasegawa-Johnson M, Smaragdis P (2015) Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Trans Audio Speech Lang Process 23(12):2136–2147

Huang Z, Watanabe S, Yang SW, García P, Khudanpur S (2022) Investigating self-supervised learning for speech enhancement and separation. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2022-May, pp 6837–6841

Hung K-H, Fu S-w, Tseng H-H, Chiang H-T, Tsao Y, Lin C-W, (2022) Boosting self-supervised embeddings for speech enhancement, arXiv preprint arXiv:2204.03339

Irvin B, Stamenovic M, Kegler M, Yang L-C (2023) Self-supervised learning for speech enhancement through synthesis. In: ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1–5. IEEE

Isik Y, Le Roux J, Chen Z, Watanabe S, Hershey JR (2016) Single-channel multi-speaker separation using deep clustering. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 08-12-Sept, pp 545–549

Isik U, Giri R, Phansalkar N, Valin JM, Helwani K, Krishnaswamy A (2020) PoCoNet: better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2020-October, pp 2487–2491

Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134

Jansson A, Humphrey E, Montecchio N, Bittner R, Kumar A, Weyde T (2017) Singing voice separation with deep U-Net convolutional networks. In: Proceedings of the 18th international society for music information retrieval conference, ISMIR 2017, pp 745–751

Ji X, Yu M, Zhang C, Su D, Yu T, Liu X, Yu D (2020) Speaker-aware target speaker enhancement by jointly learning with speaker embedding extraction. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 7294–7298. IEEE

Jiang F, Duan Z (2020) Speaker attractor network: generalizing speech separation to unseen numbers of sources. IEEE Signal Process Lett 27:1859–1863

Jiang Y, Wang DL, Liu RS, Feng ZM (2014) Binaural classification for reverberant speech segregation using deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 22(12):2112–2121

Jin Z, Wang D (2009) A supervised learning approach to monaural segregation of reverberant speech. IEEE Trans Audio Speech Lang Process 17(4):625–638

Karamatlı E, Kırbız S (2022) Mixcycle: unsupervised speech separation via cyclic mixture permutation invariant training. IEEE Signal Process Lett

Kavalerov I, Wisdom S, Erdogan H, Patton B, Wilson K, Le Roux J, Hershey JR (2019) Universal sound separation. In: IEEE workshop on applications of signal processing to audio and acoustics, vol 2019-October, pp 175–179

Kim M, Smaragdis P (2015) Adaptive denoising autoencoders: a fine-tuning scheme to learn from test mixtures. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 9237, pp 100–107

Kingma DP, Welling M (2014) Auto-encoding variational bayes. In: 2nd international conference on learning representations, ICLR 2014—conference track proceedings, no. Ml, pp 1–14

Kinoshita K, Drude L, Delcroix M, Nakatani T (2018) Listening to each speaker one by one with recurrent selective hearing networks. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2018-April, pp 5064–5068

Kitaev N, Kaiser Ł, Levskaya A (2020) Reformer: the efficient transformer. In: International conference on learning representations, pp 1–12 arXiv:org/abs/2001.04451

Kjems U, Boldt JB, Pedersen MS, Lunner T, Wang D (2009) Role of mask pattern in intelligibility of ideal binary-masked noisy speech. J Acoust Soc Am 126(3):1415–1426

Koizumi Y, Niwa K, Hioka Y, Kobayashi K, Haneda Y (2017) DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, pp 81–85

Kolbæk M, Yu D, Tan Z-H, Jensen J (2017a) Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Trans Audio Speech Lang Process 25(10):1901–1913

Kolbæk M, Tan ZH, Jensen J (2017b) Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. IEEE/ACM Trans Audio Speech Lang Process 25(1):149–163

Kolbaek M, Tan Z-H, Jensen J (2018a) On the relationship between short-time objective intelligibility and short-time spectral-amplitude mean-square error for speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 27(2):283–295

Kolbcek M, Tan ZH, Jensen J (2018b) Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2018-April, pp 5059–5063

Kolbaek M, Tan ZH, Jensen SH, Jensen J (2020) On loss functions for supervised monaural time-domain speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 28:825–838

Kong J, Kim J, Bae J (2020) Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis. Adv Neural Inf Process Syst 33:17Ã,Â 022-17Ã,Â 033

Kong Z, Ping W, Dantrey A, Catanzaro B (2022) Speech denoising in the waveform domain with self-attention. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2022-May, pp 7867–7871

Kothapally V, Hansen JH (2022a) Skipconvgan: monaural speech dereverberation using generative adversarial networks via complex time-frequency masking. IEEE/ACM Trans Audio Speech Lang Process 30:1600–1613

Kothapally V, Hansen JH (2022b) Complex-valued time-frequency self-attention for speech dereverberation, arXiv preprint arXiv:2211.12632

Kumar A, Florencio D (2016) Speech enhancement in multiple-noise conditions using deep neural networks. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 08-12-September-2016, pp 3738–3742

Lam MW, Wang J, Su D, Yu D (2021a) Effective low-cost time-domain audio separation using globally attentive locally recurrent networks. In: 2021 IEEE spoken language technology workshop, SLT 2021–proceedings, pp 801–808

Lam MW, Wang J, Su D, Yuy D (2021b) Sandglasset: a light multi-granularity self-attentive network for time-domain speech separation. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2021-June, pp 5759–5763

Le Roux J, Wichern G, Watanabe S, Sarroff A, Hershey JR (2019) Phasebook and friends: leveraging discrete representations for source separation. IEEE J Sel Top Sign Process 13(2):370–382

Lea C, Flynn MD, Vidal R, Reiter A, Hager GD (2017) Temporal convolutional networks for action segmentation and detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 156–165

Lee Y-S, Wang C-Y, Wang S-F, Wang J-C, Wu C-H (2017) Fully complex deep neural network for phase-incorporating monaural source separation. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 281–285. IEEE

Lee JH, Chang JH, Yang JM, Moon HG (2022) NAS-TasNet: neural architecture search for time-domain speech separation. IEEE Access 10:56Ã,Â 031-56Ã,Â 043

Leglaive S, Girin L, Horaud R (2018) A variance modeling framework based on variational autoencoders for speech enhancement. In: IEEE international workshop on machine learning for signal processing, MLSP, vol. 2018-September

Leglaive S, Simsekli U, Liutkus A, Girin L, Horaud R (2019) Speech enhancement with variational autoencoders and alpha-stable distributions. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2019-May, pp 541–545

Leglaive S, Alameda-Pineda X, Girin L, Horaud R (2020) A recurrent variational autoencoder for speech enhancement. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2020-May, pp 371–375

Lehtinen J, Munkberg J, Hasselgren J, Laine S, Karras T, Aittala M, Aila T (2018) Noise2noise: learning image restoration without clean data, arXiv preprint arXiv:1803.04189

León D, Tobar F (2021) Late reverberation suppression using U-nets, arXiv preprint arXiv:2110.02144., no. 1

Li K, Wu B, Lee CH (2016) An iterative phase recovery framework with phase mask for spectral mapping with an application to speech enhancement. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 08-12-Sept, pp 3773–3777

Li H, Zhang X, Zhang H, Gao G (2017) Integrated speech enhancement method based on weighted prediction error and DNN for dereverberation and denoising, arXiv preprint arXiv:1708.08251

Li ZX, Dai LR, Song Y, McLoughlin I (2018a) A conditional generative model for speech enhancement. Circ Syst Signal Process 37(11):5005–5022. https://doi.org/10.1007/s00034-018-0798-4

Li Y, Zhang X, Chen D (2018b) CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 1091–1100

Li Y, Sun Y, Horoshenkov K, Naqvi SM (2021a) Domain adaptation and autoencoder-based unsupervised speech enhancement. IEEE Trans Artif Intell 3(1):43–52

Li A, Liu W, Luo X, Yu G, Zheng C, Li X (2021b) A simultaneous denoising and dereverberation framework with target decoupling. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2, pp 796–800

Li H, Chen K, Wang L, Liu J, Wan B, Zhou B (2022) Sound source separation mechanisms of different deep networks explained from the perspective of auditory perception. Appl Sci 12(2):832

Liao C-F, Tsao Y, Lee H-Y, Wang H-M (2018) Noise adaptive speech enhancement using domain adversarial training, arXiv preprint arXiv:1807.07501

Lim JS, Oppenheim AV (1979) Enhancement and bandwidth compression of noisy speech. Proc IEEE 67(12):1586–1604

Lin YC, Hsu YT, Fu SW, Tsao Y, Kuo TW (2019) IA-Net: acceleration and compression of speech enhancement using integer-adder deep neural network. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2019-September, pp 1801–1805

Liu Y, Wang D (2019) Divide and conquer: a deep CASA approach to talker-independent monaural speaker separation. IEEE/ACM Trans Audio Speech Lang Process 27(12):2092–2102

Liu AT, Yang SW, Chi PH, Hsu PC, Lee HY (2020) Mockingjay: unsupervised speech representation learning with deep bidirectional transformer encoders. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2020-May, pp 6419–6423

Liu AT, Li SW, Lee HY (2021) TERA: self-supervised learning of transformer encoder representation for speech. IEEE/ACM Trans Audio Speech Lang Process 29:2351–2366

Liu H, Liu X, Kong Q, Tian Q, Zhao Y, Wang D, Huang C, Wang Y (2022) VoiceFixer: a unified framework for high-fidelity speech restoration, arXiv preprint arXiv:2204.05841, no. September, pp 4232–4236

Lluís F, Pons J, Serra X (2019) End-to-end music source separation: is it possible in the waveform domain? In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2019-September, pp 4619–4623

Loizou PC (2013) Speech enhancement: theory and practice. CRC Press, BOca Raton

Loizou PC, Kim G (2011) Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans Audio Speech Lang Process 19(1):47–56

Lu X, Tsao Y, Matsuda S, Hori C (2013) Speech enhancement based on deep denoising autoencoder. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, no. August, pp 436–440

Lu Y-J, Tsao Y, Watanabe S (2021) A study on speech enhancement based on diffusion probabilistic model. In: 2021 Asia-pacific signal and information processing association annual summit and conference (APSIPA ASC), 2021, pp 659–666. IEEE

Lu Y-J, Wang Z-Q, Watanabe S, Richard A, Yu C, Tsao Y (2022) Conditional diffusion probabilistic model for speech enhancement. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 7402–7406. IEEE

Luo C (2022) Understanding diffusion models: a unified perspective, arXiv preprint arXiv:2208.11970

Luo Y, Mesgarani N (2018) TaSNet: time-domain audio separation network for real-time, single-channel speech separation. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2018-April, pp 696–700

Luo Y, Mesgarani N (2019) Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Trans Audio Speech Lang Process 27(8):1256–1266

Luo Y, Mesgarani N (2020) Separating varying numbers of sources with auxiliary autoencoding loss. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2020-October, pp 2622–2626

Luo Y, Chen Z, Hershey JR, Le Roux J, Mesgarani N (2017) Deep clustering and conventional networks for music separation: stronger together. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, pp 61–65

Luo Y, Chen Z, Mesgarani N (2018) Speaker-independent speech separation with deep attractor network. IEEE/ACM Trans Audio Speech Lang Process 26(4):787–796

Luo Y, Chen Z, Yoshioka T (2020) Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol. 2020-May, pp 46–50

Luo J, Wang J, Cheng N, Xiao E, Zhang X, Xiao J (2022) Tiny-sepformer: a tiny time-domain transformer network for speech separation, arXiv preprint arXiv:2206.13689, no. 1, pp 5313–5317

Lutati S, Nachmani E, Wolf L (2022) SepIt: approaching a single channel speech separation bound, arXiv preprint arXiv:2205.11801, pp 5323–5327

Mao X, Li Q, Xie H, Lau RY, Wang Z, Paul Smolley S (2017) Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2794–2802

Martin-Donas JM, Gomez AM, Gonzalez JA, Peinado AM (2018) A deep learning loss function based on the perceptual evaluation of the speech quality. IEEE Signal Process Lett 25(11):1680–1684

Miao Y, Zhang H, Metze F (2015) Speaker adaptive training of deep neural network acoustic models using i-vectors. IEEE/ACM Trans Audio Speech Lang Process 23(11):1938–1949

Nábělek AK, Letowski TR, Tucker FM (1989) Reverberant overlap- and self-masking in consonant identification. J Acoust Soc Am 86(4):1259–1265

Nachmani E, Adi Y, Wolf L (2020) Voice separation with an unknown number of multiple speakers. In: 37th international conference on machine learning, ICML 2020, vol PartF16814, pp 7121–7132

Narayanan A, Wang D (2013) Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: 2013 IEEE international conference on acoustics, speech and signal processing, pp 7092–7096

Narayanan A, Wang D (2015) Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training. IEEE/ACM Trans Audio Speech Lang Process 23(1):92–101

Natsiou A, O'Leary S (2021) Audio representations for deep learning in sound synthesis: a review. In: Proceedings of IEEE/ACS international conference on computer systems and applications, AICCSA, vol 2021-Decem

Naylor NDG, Patrick A (2010) Speech dereverberation. In: Naylor NDG Patrick A (ed) vol. 53, no. 1. Springer, London (2010)

Neumann TV, Kinoshita K, Delcroix M, Araki S, Nakatani T, Haeb-Umbach R (2019) All-neural online source separation, counting, and diarization for meeting analysis. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2019-May, pp 91–95

Nossier SA, Wall J, Moniri M, Glackin C, Cannings N (2020a) A comparative study of time and frequency domain approaches to deep learning based speech enhancement. In: Proceedings of the international joint conference on neural networks

Nossier SA, Wall J, Moniri M, Glackin C, Cannings N (2020b) Mapping and masking targets comparison using different deep learning based speech enhancement architectures. In: 2020 international joint conference on neural networks (IJCNN). IEEE, pp 1–8

Ochiai T, Matsuda S, Lu X, Hori C, Katagiri S (2014) Speaker adaptive training using deep neural networks. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 6349–6353. IEEE

Oppenheim AV (1999) Discrete-time signal processing, 2ndÃ,Â ed. Prentice-Hall, Upper Saddle River

Oppenheim AV, Lim JS (1981) The importance of phase in signals. Proc IEEE 69(5):529–541

Paliwal K, Wójcicki K, Shannon B (2011) The importance of phase in speech enhancement. Speech Commun 53(4):465–494

Pan SJ, Tsang IW, Kwok JT, Yang Q (2010) Domain adaptation via transfer component analysis. IEEE Trans Neural Netw 22(2):199–210

Parveen S, Green P (2004) Speech enhancement with missing data techniques using recurrent neural networks. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 1, no. Figure 1, pp 13–16

Pascual S, Bonafonte A, Serra J (2017) SEGAN: speech enhancement generative adversarial network. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2017-August, no. D, pp 3642–3646

Pascual S, Park M, Serrà J, Bonafonte A, Ahn K-H (2018) Language and noise transfer in speech enhancement generative adversarial network. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5019–5023. IEEE

Pascual S, Serrà J, Bonafonte A (2019) Towards generalized speech enhancement with generative adversarial networks. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2019-September, pp 1791–1795

Phan H, McLoughlin IV, Pham L, Chen OY, Koch P, De Vos M, Mertins A (2020) Improving GANs for speech enhancement. IEEE Signal Process Lett 27:1700–1704

Portnoff MR (1980) Time-frequency representation of. digital signals. IEEE Trans Acoust Speech Signal Process 28(1):55–69

Qian K, Zhang Y, Chang S, Yang X, Florêncio D, Hasegawa-Johnson M (2017) Speech enhancement using bayesian wavenet. In: Interspeech, pp 2013–2017

Qin S, Jiang T (2018) Improved Wasserstein conditional generative adversarial network speech enhancement. EURASIP J Wirel Commun Netw 1:2018

Qin S, Jiang T, Wu S, Wang N, Zhao X (2020) Graph convolution-based deep clustering for speech separation. IEEE Access 8:82 571-82 580

Qiu W, Hu Y (2022) Dual-path hybrid attention network for monaural speech separation. IEEE Access 10:78Ã‚Â 754-78Ã‚Â 763

Reddy CK, Gopal V, Cutler R (2021) Dnsmos: a non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 6493–6497. IEEE

Rethage D, Pons J, Serra X (2018) A wavenet for speech denoising. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2018-April, pp 5069–5073

Rix AW, Beerends JG, Hollier MP, Hekstra AP (2001) Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2, pp 749–752

Roux JL, Wisdom S, Erdogan H, Hershey JR (2019) SDR—half-baked or well done? In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2019-May, pp 626–630

Sainath TN, Weiss RJ, Senior A, Wilson KW, Vinyals O (2015) Learning the speech front-end with raw waveform CLDNNs. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2015-January, pp 1–5

Saito K, Uhlich S, Fabbro G, Mitsufuji Y (2021) Training speech enhancement systems with noisy speech datasets, arXiv preprint arXiv:2105.12315

Schmidt MN, Olsson RK (2006) Single-channel speech separation using sparse non-negative matrix factorization. In: INTERSPEECH 2006 and 9th international conference on spoken language processing, INTERSPEECH 2006—ICSLP, vol 5, pp 2614–2617

Senior A, Lopez-Moreno I (2014) Improving DNN speaker independence with i-vector inputs. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 225–229. IEEE

Shao Y, Wang D (2006) Model-based sequential organization in cochannel speech. IEEE Trans Audio Speech Lang Process 14(1):289–298

Shi J, Xu J, Liu G, Xu B (2018) Listen, think and listen again: capturing top-down auditory attention for speaker-independent speech separation. In: IJCAI international joint conference on artificial intelligence, vol 2018-July, pp 4353–4360

Shivakumar PG, Georgiou P (2016) Perception optimized deep denoising autoencoders for speech enhancement. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 08-12-September-2016, pp 3743–3747

Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S (2015) Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. PMLR, pp 2256–2265

Stoller D, Ewert S, Dixon S (2018) Wave-u-net: a multi-scale neural network for end-to-end audio source separation, arXiv preprint arXiv:1806.03185

Subakan C, Ravanelli M, Cornell S, Bronzi M, Zhong J (2021) Attention is all you need in speech separation. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2021-June, pp 21–25

Subakan C, Ravanelli M, Cornell S, Grondin F, Bronzi M (2022a) On using transformers for speech-separation. In: International workshop on acoustic signal enhancement, vol 14, no. 8, pp 1–10. arXiv:org/abs/2202.02884

Subakan C, Ravanelli M, Cornell S, Lepoutre F, Grondin F (2022b) Resource-efficient separation transformer, arXiv preprint arXiv:2206.09507, pp 1–5

Su J, Jin Z, Finkelstein A (2020) HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2020-October, no. 3, pp 4506–4510

Sun H, Li S (2017) An optimization method for speech enhancement based on deep neural network. In: IOP conference series: earth and environmental science, vol 69, no 1

Taal CH, Hendriks RC, Heusdens R, Jensen J (2010) A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: IEEE international conference on acoustics, speech, and signal processing, pp 4214–4217

Taal CH, Hendriks RC, Heusdens R, Jensen J (2011) An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans Audio Speech Lang Process 19(7):2125–2136

Tachibana H (2021) Towards listening to 10 people simultaneously: an efficient permutation invariant training of audio source separation using Sinkhorn's algorithm. In: ICASSP, IEEE international conference on acoustics, speech and signal processing— proceedings, vol 2021-June, pp 491–495

Takahashi N, Parthasaarathy S, Goswami N, Mitsufuji Y (2019) Recursive speech separation for unknown number of speakers. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2019-September, pp 1348–1352

Tan K, Wang D (2021) Towards model compression for deep learning based speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 29:1785–1794

Trinh VA, Braun S (2022) Unsupervised speech enhancement with speech recognition embedding and disentanglement losses. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 391–395. IEEE

Tu Y, Du J, Xu Y, Dai L, Lee CH (2014) Deep neural network based speech separation for robust speech recognition. In: International conference on signal processing proceedings, ICSP, vol 2015-January, no. October, pp 532–536

Tzinis E, Venkataramani S, Wang Z, Subakan C, Smaragdis P (2020a) Two-step sound source separation: training on learned latent targets. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2020-May, pp 31–35

Tzinis E, Wang Z, Smaragdis P (2020b) Sudo RM -RF: efficient networks for universal audio source separation. in: IEEE international workshop on machine learning for signal processing, MLSP, vol 2020-September

Tzinis E, Adi Y, Ithapu VK, Xu B, Smaragdis P, Kumar A (2022) Remixit: continual self-training of speech enhancement models via bootstrapped remixing. IEEE J Sel Topics Signal Process 16(6):1329–1341

Ueda Y, Wang L, Kai A, Xiao X, Chng ES, Li H (2016) Single-channel dereverberation for distant-talking speech recognition by combining denoising autoencoder and temporal structure normalization. J Signal Process Syst 82(2):151–161

Valin J-M, Giri R, Venkataramani S, Isik U, Krishnaswamy A (2022) To dereverb or not to dereverb? Perceptual studies on real-time dereverberation targets, arXiv:2206.07917

van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet: a generative model for raw audio, arXiv preprint arXiv:1609.03499, pp 1–15

Vary P, Eurasip M (1985) Noise suppression by spectral magnitude estimation-mechanism and theoretical limits. Signal Process 8(4):387–400

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 2017:5999–6009

Venkataramani S, Casebeer J, Smaragdis P (2018) End-to-end source separation with adaptive front-ends. In: 2018 52nd asilomar conference on signals, systems, and computers, no. 1, pp 684–688

Veselỳ K, Watanabe S, Žmolíková K, Karafiát M, Burget L, Černockỳ JH (2016) Sequence summarizing neural network for speaker adaptation. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5315–5319. IEEE

Virtanen T (2006) Speech recognition using factorial hidden Markov models for separation in the feature space. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 1, pp 89–92

Virtanen T, Cemgil AT (2009) Mixtures of gamma priors for non-negative matrix factorization based speech separation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 5441, no 3, pp 646–653

von Neumann T, Boeddeker C, Drude L, Kinoshita K, Delcroix M, Nakatani T, Haeb-Umbach R (2020) Multi-talker ASR for an unknown number of sources: joint training of source counting, separation and ASR. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2020-October, pp 3097–3101

Wang D (2008) Time—frequency masking for speech hearing aid design. Trends in amplification, vol 12, pp 332–353. http://www.ncbi.nlm.nih.gov/pubmed/18974204

Wang D, Chen J (2018) Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans Audio Speech Lang Process 26(10):1702–1726

Wang DL, Lim JS (1982) The unimportance of phase in speech enhancement. IEEE Trans Acoust Speech Signal Process 30(4):679–681

Wang Z, Sha F (2014) Discriminative non-negative matrix factorization for single-channel speech separation. In: 2014 IEEE international conference on acoustic, speech and signal processing (ICASSP), pp 3777–3781 https://pdfs.semanticscholar.org/854a/454106bd42a8bca158426d8b12b48ba0cae8.pdf

Wang Y, Wang DL (2013) Towards scaling up classification-based speech separation. IEEE Trans Audio Speech Lang Process 21(7):1381–1390

Wang L, Yoon KJ (2022) Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks. IEEE Trans Pattern Anal Mach Intell 44(6):3048–3068

Wang Y, Han K, Wang D (2013) Exploring monaural features for classification-based speech segregation. IEEE Trans Audio Speech Lang Process 21(2):270–279

Wang Y, Narayanan A, Wang DL (2014) On training targets for supervised speech separation. IEEE/ACM Trans Audio Speech Lang Process 22(12):1849–1858

Wang Y, Du J, Dai L-R, Lee C-H (2016) Unsupervised single-channel speech separation via deep neural network for different gender mixtures. In: 2016 Asia-pacific signal and information processing association annual summit and conference (APSIPA), pp 1–4. IEEE

Wang Y, Du J, Dai LR, Lee CH (2017) A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 25(7):1535–1546

Wang ZQ, Roux JL, Hershey JR (2018a) Alternative objective functions for deep clustering. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2018-April, pp 686–690

Wang J, Chen J, Su D, Chen L, Yu M, Qian Y, Yu D (2018b) Deep extractor network for target speaker recovery from single channel speech mixtures, arXiv preprint arXiv:1807.08974

Wang Q, Muckenhirn H, Wilson K, Sridhar P, Wu Z, Hershey J, Saurous RA, Weiss RJ, Jia Y, Moreno IL (2018c) Voicefilter: targeted voice separation by speaker-conditioned spectrogram masking, arXiv preprint arXiv:1810.04826

Wang Q, Rao W, Sun S, Xie L, Chng ES, Li H (2018d) Unsupervised domain adaptation via domain adversarial training for speaker recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4889–4893. IEEE

Wang ZQ, Tan K, Wang D (2019) Deep learning based phase reconstruction for speaker separation: a trigonometric perspective. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol 2019-May, pp. 71–75

Wang S, Li BZ, Khabsa M, Fang H, Ma H (2020) Linformer: self-attention with linear complexity, vol 2048, no. 2019. arXiv:org/abs/2006.04768

Wang K, He B, Zhu WP (2021) Tstnn: two-stage transformer based neural network for speech enhancement in the time domain. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, vol. 2021-June, pp 7098–7102

Weng C, Yu D, Seltzer ML, Droppo J (2015) Deep neural networks for single-channel multi-talker speech recognition. IEEE/ACM Trans Audio Speech Lang Process 23(10):1670–1679

Weninger F, Hershey JR, Le Roux J, Schuller B (2014) Discriminatively trained recurrent neural networks for single-channel speech separation. In: 2014 IEEE global conference on signal and information processing, GlobalSIP 2014, pp 577–581

Weninger F, Erdogan H, Watanabe S, Vincent E, Le Roux J, Hershey JR, Schuller B (2015) Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR.

Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 9237, pp 91–99

Wichern G, Lukin A (2017) Low-latency approximation of bidirectional recurrent networks for speech denoising. In: IEEE workshop on applications of signal processing to audio and acoustics, vol 2017-October, pp 66–70

Williamson DS, Wang D (2017a) Speech dereverberation and denoising using complex ratio masks. In: IEEE international conference on acoustics, speech, and signal processing (ICASSP) 2017, pp 5590–5594

Williamson DS, Wang D (2017b) Time-frequency masking in the complex domain for speech dereverberation and denoising. IEEE/ACM Trans Audio Speech Lang Process 25(7):1492–1501

Williamson DS, Wang Y, Wang DL (2016) Complex ratio masking for monaural speech separation. IEEE/ACM Trans Audio Speech Lang Process 24(3):483–492

Wisdom S, Tzinis E, Erdogan H, Weiss RJ, Wilson K, Hershey JR (2020) Unsupervised sound separation using mixture invariant training. In: Advances in neural information processing systems, vol 2020-December, june 2020. arXiv:org/abs/2006.12701

Wu JY, Yu C, Fu SW, Liu CT, Chien SY, Tsao Y (2019) Increasing compactness of deep learning based speech enhancement models with parameter pruning and quantization techniques. IEEE Signal Process Lett 26(12):1887–1891

Xia B, Bao C (2014) Wiener filtering based speech enhancement with Weighted Denoising Autoencoder and noise classification, pp 13–29

Xiang Y, Bao C (2020) A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network. IEEE/ACM Trans Audio Speech Lang Process 28:1826–1838

Xiao X, Chen Z, Yoshioka T, Erdogan H, Liu C, Dimitriadis D, Droppo J, Gong Y (2019) Single-channel speech extraction using speaker inventory and attention network. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 86–90

Xiao F, Guan J, Kong Q, Wang W (2021) Time-domain speech enhancement with generative adversarial learning, arXiv preprint arXiv:2103.16149

Xu Y, Du J, Dai LR, Lee CH (2014a) An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process Lett 21(1):65–68

Xu Y, Du J, Dai L-R, Lee C-H (2014b) Cross-language transfer learning for deep neural network based speech enhancement. In: The 9th international symposium on chinese spoken language processing, pp 336–340. IEEE

Xu Y, Du J, Dai L-R, Lee C-H (2014c) Global variance equalization for improving deep neural network based speech enhancement. In: 2014 IEEE China summit & international conference on signal and information processing (ChinaSIP). IEEE, pp 71–75

Xu Y, Du J, Dai LR, Lee CH (2015) A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 23(1):7–19

Yan Z, Buye X, Ritwik G, Tao Z (2018) Perceptually guided speech enhancement using deep neural networks. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, pp 5074–5078

Ye F, Tsao Y, Chen F (2019) Subjective feedback-based neural network pruning for speech enhancement. In: 2019 Asia-pacific signal and information processing association annual summit and conference, APSIPA ASC 2019, pp 673–677

Yu D, Kolbaek M, Tan ZH, Jensen J (2017) Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, pp 241–245

Yul D, Kalbcek M, Tan Z-H, Jensen J (2017) Speaker-independent multi-talker speech separation. In: IEEE international conference on acoustics, speech and signal processing, pp 241–245

Zeghidour N, Grangier D (2021) Wavesplit: end-to-end speech separation by speaker clustering. IEEE/ACM Trans Audio Speech Lang Process 29(4):2840–2849

Zhang XL, Wang D (2016) A deep ensemble learning method for monaural speech separation. IEEE/ACM Trans Audio Speech Lang Process 24(5):967–977

Zhang H, Zhang X, Gao G (2018) Training supervised speech separation system to improve STOI and PESQ directly. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, pp 5374–5378

Zhang Q, Nicolson A, Wang M, Paliwal KK, Wang C (2020a) DeepMMSE: a deep learning approach to mmse-based noise power spectral density estimation. IEEE/ACM Trans Audio Speech Lang Process 28:1404–1415

Zhang L, Shi Z, Han J, Shi A, Ma D (2020b) FurcaNeXt: end-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 11961 LNCS, pp 653–665

Zhang C, Yu M, Weng C, Yu D (2021a) Towards robust speaker verification with target speaker enhancement. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 6693–6697. IEEE

Zhang J, Zorila C, Doddipatla R, Barker J (2021b) Teacher-student mixit for unsupervised and semi-supervised speech separation, arXiv preprint arXiv:2106.07843

Zhao Y, Wang ZQ, Wang D (2019) Two-stage deep learning for noisy-reverberant speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 27(1):53–62

Zhao Y, Wang D, Xu B, Zhang T (2020) Monaural speech dereverberation using temporal convolutional networks with self attention. IEEE/ACM Trans Audio Speech Lang Process 28:1598–1607

Zheng N, Zhang XL (2019) Phase-aware speech enhancement based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 27(1):63–76

Zhou R, Zhu W, Li X (2022) Single-channel speech dereverberation using subband network with a reverberation time shortening target, arXiv preprint arXiv:2210.11089arXiv:2204.08765

Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232

Zolnay A, Kocharov D, Schlüter R, Ney H (2007) Using multiple acoustic feature sets for speech recognition. Speech Commun 49(6):514–525

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.