IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 December 2022, accepted 3 January 2023, date of publication 9 January 2023, date of current version 13 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3235010

TOPICAL REVIEW

State-of-the-Art Analysis of Deep Learning-Based Monaural Speech Source Separation Techniques

SWATI SONI, RAM NARAYAN YADAV, AND LALITA GUPTA, (Senior Member, IEEE)
Department of Electronics and Communication Engineering, Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh 462003, India
Corresponding author: Swati Soni (swatisoni3131@gmail.com)

ABSTRACT The monaural speech source separation problem is an important application in the signal processing field. But recent interaction of deep learning algorithms with signal processing achieves remarkable performance improvement for speech source separation problems. This paper explores the numerous state-of-the-art deep learning-based monaural speech source separation algorithms in the time-frequency (T-F), time, and hybrid domains. The motivation, algorithm, and framework of different deep learning models for monaural speech source separation are analyzed. The benchmarked algorithms in the T-F domain can be categorized as deep neural networks (DNN), clustering, permutation, multi-task learning, computational auditory sense analysis (CASA), and phase reconstruction-based techniques, whereas the state-of-the-art time-domain approaches can be categorized as CNN, RNN, multi-scale fusion (MSF), and transformer-based techniques. The end-to-end post filter (E2EPF) is a hybrid algorithm combining T-F and time-domain works to achieve enhanced results. Time-domain models have shown improvement in separation performance compared to the T-F and hybrid domain models with small model sizes. Methods in T-F, time, and hybrid domains are compared using *SDR*, $SI-SDR$, $SI-SNR$, PESQ, and *STOI* as quality assessment metrics on some benchmark datasets.

INDEX TERMS Deep-clustering, deep learning, monaural speech source separation, permutation invariant training, time domain speaker separation.

I. INTRODUCTION

The source separation problem occurs due to the undesired signal mixing with the desired signal. The undesired signal includes a speaker signal other than the target speaker, interference, reverberation, and background noises. Automatic voice recognition (to convert speech into text) [1], assisted living (to make appropriate living conditions for older and persons with disabilities) [2], and hearing aids (to improve the hearing capability of the person with hearing loss) [3], and many more are applications of monaural source separation [4], [5], [6], [7], [8], [9]. Hence, many researchers are interested in working on source separation problems due to their widespread applications. Source separation can be categorized as single-channel (monaural) and multichannel categories. In signal and speech processing, monaural speech source separation is challenging because it separates the target speaker from the mixture of speakers and the background noises and interferences in a single microphone recording. Speaker separation [10], [11], [12], [13], speech enhancement [14], [15], and speech de-reverberation and de-noising [16] come under single-channel source separation categories, as in Fig. 1.

Speaker separation allows extracting more than one speaker from the mixture of two and more than two speakers [10], [11], [12], [13]. Speech enhancement improves noisy speech signals' intelligibility and perceptual quality [14], [15] and attempts to separate speech from noisy mixture signals. Speech de-reverberation and de-noising remove reverberation and suppress background noise from the target speaker signal [16], [17]. Speaker separation is the pre-processing stage in many speech-processing applications with multiple speakers, such as multi-speaker automatic speech recognition [18] and multi-speaker emotion recognition [19], [20]. Hence researchers are motivated to work and improve the speaker separation algorithms.

The associate editor coordinating the review of this manuscript and approving it for publication was Mira Naftaly.

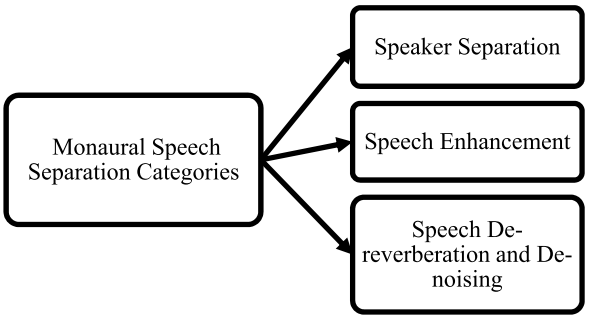

FIGURE 1. Single channel speech separation categories.

Monaural source separation or single-channel source separation works with two learning methods, supervised learning (models can use previous experience to produce outcomes), and unsupervised learning (models do not have previous experience). Existing review articles describe supervised single-channel speaker separation algorithms in either signal processing [21], [22], [23] or in the time-frequency [24], [25] domains. The conventional single channel speaker separation techniques such as computational auditory sense analysis (CASA) [26], non-negative matrix factorization (NMF) [27], [28] in the signal processing domain and deep learning-based deep clustering (DC) [29], deep attractor networks (DANet) [30], permutation invariant training (PIT) [31], in T-F domain have been reviewed in [32]. A comprehensive review with background introduction and formulation of speech separation and components of supervised separation, i.e., learning machines, training targets, and acoustic features, have been introduced with a description of monaural speech enhancement, speaker separation, and speech de-reverberation as well as multi-microphone techniques in [17]. The articles [17], [32], [33], [34] presented interesting reviews of deep learning applied to various problems of speech processing. Nevertheless, these review articles presented speaker separation using deep learning in the T-F domain only in a short portion of the overview. Recently deep learning-based supervised time domain algorithms have achieved significant progress, motivating to review time-frequency, time, and hybrid domain approaches. This paper compares the supervised monaural speaker separation algorithm based on deep learning in T-F, time, and hybrid domains. Available objective performance metrics to evaluate separation models, training objectives, and datasets have been introduced to make the researchers aware of background information for deep-leaning-based speaker separation in T-F, time, and hybrid domains. Before being familiar with deep learning advantages, signal processing-based approaches performed the audio source separation tasks. Signal processing-based speech source separation models can be classified as statistical, clustering, and factorization models, as shown in Fig. 2.

Statistical models include probabilistic models such as Gaussian Mixture Models (GMM) [35], [36], [37], [38], Hidden Markov Models (HMM) [39], [40] and factorial Hidden Markov Models [41], [42], etc. GMM can work well for different gender speakers, and the HMM model separates similar-gender speakers efficiently. GMM and HMM models assume that source energy does not change throughout the change from mixture signal to separated signal. This assumption limits the real-time performance of the models. Factorial HMM models [43], [44], [45], a gain-adapted minimum mean square error estimator [46], and a frame-based gain estimation technique [47] overcome this limitation but compromise increased computational complexity. Clustering methods use computational auditory sense analysis (CASA) [26] and spectral clustering [48], [49], [50] for performing source separation tasks. These methods are based on the principle of auditory sense analysis and attempt to perform separation similar to the human auditory system. The CASA systems aim to separate the mixture of sound sources like human ears do. Hence, the CASA system can be interpreted as a machine listening system [51], [52]. Factorization models make use of the principle of non-negative matrix factorization (NMF) [27], [28]. Considering the source signal non-negative can make its energy unaltered throughout separation with HMM and GMM as in [27], [53], [54], and [55]. However, the energy of real-world sources can be negative or positive.

All these classification-based approaches estimate hard masks to classify each time-frequency (T-F) bin belonging to sources [56]. Due to this hard decision, essential information related to sources can be lost. Signal processing-based approaches fail to work well with real-world scenarios. The success of deep learning applications in various research fields inspires researchers to perform supervised monaural speech source separation in the deep learning domain [57]. Deep learning models with so many hidden layers are suitable for dealing with complex real-world data.

Deep learning-based single-channel speech source separation approaches perform separation in the T-F, time, and hybrid domains. In the T-F domain, DNN, clustering, permutation, multi-task learning, CASA, and phase reconstruction-based approaches are used to separate the speakers from mixture signals. Deep clustering (DC) [29], deep attractor networks (DANet) [30], permutation invariant training (PIT) [31], etc., are benchmarked T-F domain approaches. These methods calculate the spectrums of signals to get into the T-F domain using a Short-Time Fourier Transform (STFT) [58]. The separation can be performed by calculating a mask function and multiplying it with a mixture signal to obtain a clean speech signal. These methods calculate soft mask function instead of hard mask hence getting better separation accuracy than signal processing-based approaches. The STFT is a suboptimal transformation for speech signals because it is not specifically designed for speech signals and can transform any type of signal into the T-F domain. The T-F domain methods only process the magnitude spectrum, leave the phase spectrum unchanged, and can cause phase magnitude decoupling.

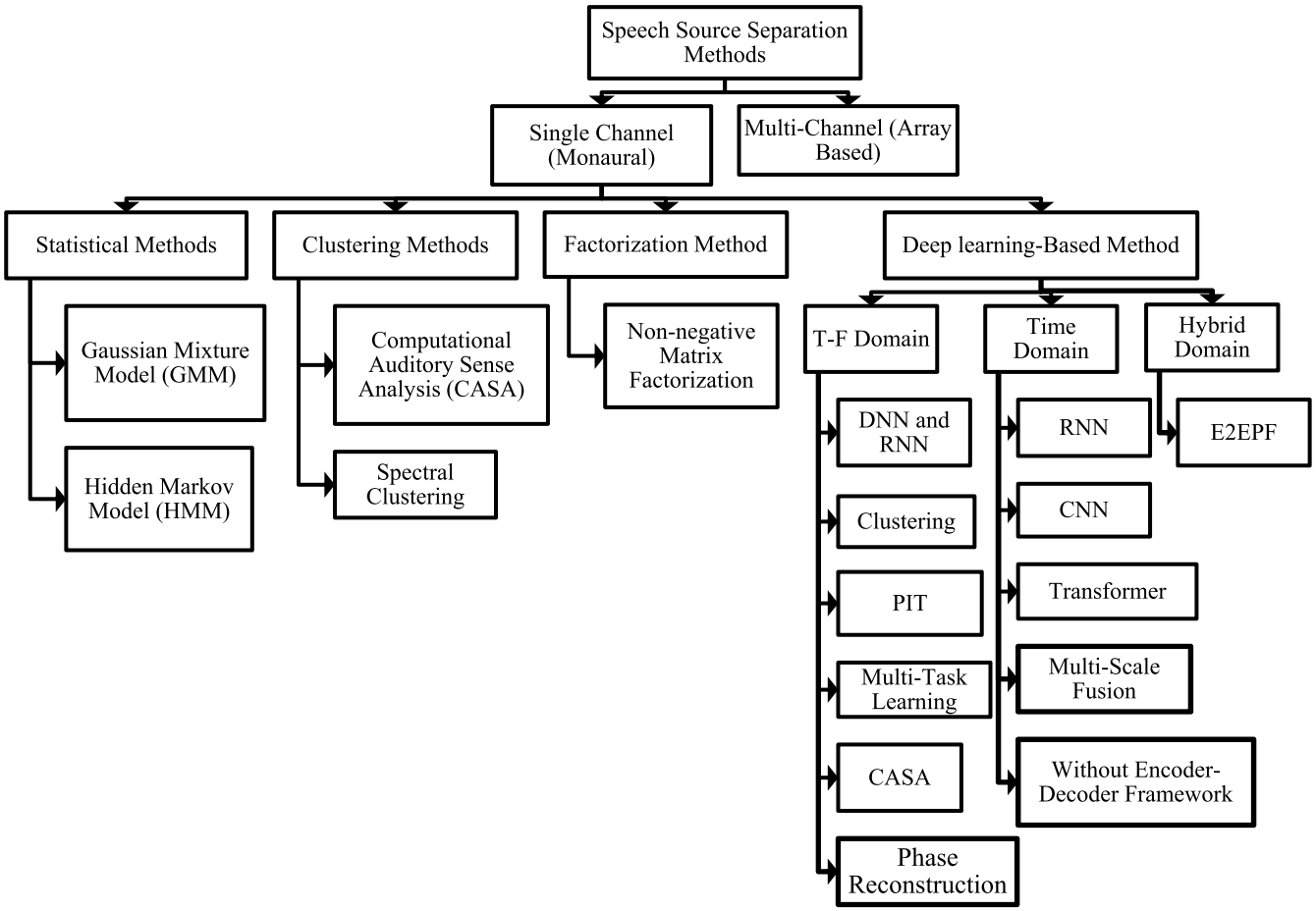

FIGURE 2. Block diagram illustrating classification of single channel speaker separation techniques.

Phase reconstruction-based approaches overcome this limitation with limited performance and increased complexity. The separation accuracy of T-F domain methods increases with increased window size and compromises with the size and complexity of models. STFT calculation, phase magnitude decoupling, and long contextual window are the limitations of T-F domain methods and inspire the researchers to work in the time domain. The time-domain approaches make use of data-driven representation instead of T-F domain spectrograms. In these methods, separate models are designed for data-driven representations and inverse transformation. These methods have an encoder-decoder and separation modules. The encoder module converts the time-domain mixture speech signal into an encoded time-domain mixture signal. The separation module calculates the mask function using the encoder output. The calculated mask functions are multiplied with the mixture signal from the encoder to separate sources. Then decoder transforms the separated sources into an understandable form. Deep learning-based time-domain speech source separation can be categorized as CNN, RNN, and transformer-based approaches and techniques without an encoder-decoder framework. Time-domain audio source separation (TasNet) [59], Convolutional TasNet (ConvTasNet) [60], etc., are examples of time-domain audio source separation work. Time-domain approaches overcome limitations of T-F domain approaches like STFT calculation, magnitude and phase decoupling, and long context window. The end-to-end post filter is the hybrid method performing separation in T-F and time domains. This paper reviews the T-F, time, and hybrid domain deep learning-based monaural audio source separation approaches. Section III explains the performance measures for comparing audio source separation outcomes. Section IV presents existing training objectives to train deep learning models for speech source separation tasks. Section V describes various available datasets for monaural speech source separation frameworks. Section VI reviews the state-of-the-art deep learning-based monaural speech source separation algorithm in the T-F, time, and hybrid domains. Section VII compares the performance of speech source separation approaches using SDR, $SI-SDR$, $SI-SNR$, PESQ, and $STOI$ on different datasets. Section VII concludes T-F, time, and hybrid domain speech source separation algorithms.

II. PERFORMANCE MEASURE

Subjective and objective are two types of performance measures for evaluating speech source separation outcomes.

Subjective measures are scores given by a human personal perspective or viewpoint for the outcomes of the separation tasks. Human perceptual involvement makes subjective measures more reliable standards than objective measures, but they are time-consuming and expensive, which are the reasons for using them rarely. Further, humans can have different perspectives on particular outputs. Objective measures are cheaper and faster and perform a set of calculations for evaluating separation quality by comparing estimated outcomes with the clean separated sources. This paper explains objective measures because of their wide use in research to judge and compare separation accuracy. Commonly used objective metrics for monaural audio source separation algorithms are as follows: Source to distortion ratio (*SDR*) [61], source-to-interference ratio (*SIR*) [61], source-to-artifact ratio (*SAR*) [61], signal-to-noise ratio (*SNR*) [61], scale-invariance *SDR* (*SI* − *SDR*) [62], scale-invariance *SIR* (*SI* − *SIR*) [62], scale-invariance *SAR* (*SI* − *SAR*) [62], scale-invariance *SNR* (*SI* − *SNR*) [60], [61], short time object intelligibility (*STOI*) [63], perceptual evaluation of speech quality (PESQ) [64].

The predicted separated signal $\hat{P}$ from the mixture can be decomposed as target source signal and errors due to interference, noise, and artifact as follows

$$\hat{P} = d_{tar} + e_{int} + e_{noi} + e_{art} \tag{1}$$

In the above equation, d_{tar} is the target source signal with acceptable distortion, e_{int}, e_{noi}, and e_{art} show error due to interference, noise, and artifacts respectively. Numerical performance measures *SDR*, *SIR*, *SAR*, *SNR*, *SI*−*SDR*, *SI*−*SNR*, *SI* − *SIR*, and *SI* − *SAR* are energy ratios expressed in dB, with higher values indicating good results and formulated in TABLE 1. *SDR* can be defined as the ratio of the energy of the target source signal to the energy of the sum of all error signals. Almost all existing speech source separation techniques uses *SDR* to evaluate separation performance. The *SDR* is equivalent to the opposite of the normalized log squared error by the reference signal energy [62], [65]. The *SIR* evaluates the amount of error in predicted signal due to interferences. It computes the correlation between the target and estimated signals by calculating the log of the target signal energy to the interference error signal's energy ratio. The *SAR* represents undesired artifacts in the estimated source signal compared to the actual source signal and can be calculated as a log of the ratio of the energy of the target signal plus the error signals due to interference and noise to the energy of the error signal due to artifacts. To make it independent of noise and interference the formulation of *SAR* in TABLE 1 numerator contains error term due to noise and interference. The *SNR* can be defined as the log of the ratio of target signal plus interference error signal energy to the energy of error due to noise. *SIR*, *SAR*, and *SNR* are introduced as a performance measure for audio signals because *Distortion* = *Interference* + *Artifact* and are separately used for comparing different monaural source separation approaches [61]. Furthermore, the formulation of *SIR*, *SAR*, and *SNR* exhibit a nonlinear relationship which may be irrelevant for proper analysis of machine learning algorithms. These metrics can be made scale invariance to get linear relation between them. In the condition where the estimated signal is a scaled version of the target, scaling the estimate is helpful to get perceptually enhanced output rather than boosting a particular metric. Scale invariance metrics perform scaling to produce outcomes invariance to scale. Suppose the target signal e_{tar} is a scaled version of the predicted target signal $e_{tar} = \alpha d_{tar}$, here α is the scaling factor. In this case predicted signal can be decomposed as $\hat{P} = e_{tar} + e$, where $e = e_{int} + e_{art}$. *SI* − *SDR*, *SI* − *SIR*, and *SI* − *SAR* can be formulated as in TABLE 1. These numerical illustration helps to derive $\|e\|^2 = \|e_{int}\|^2 + \|e_{art}\|^2$. Hence scale invariance metrics produce a direct relationship between signal distortion, interference, and artifact metric.

The metrics can be made scale invariance by normalizing the predicted and clean speech signals to the zero mean before calculation [60]. Scale invariance *SNR* (*SI* − *SNR*) is one of the commonly used performance metrics for source separation approaches [61], [66]. *SI* − *SDR* is equivalent to *SI* − *SNR* when e is only due to the noise and can be illustrated in TABLE 1.

Short-time objective intelligibility (*STOI*) [63] is a performance assessment measure of objective time-domain signal intelligibility for separating monaural audio sources. It evaluates the intelligibility content by calculating the similarity of time-related short-time envelops of the time-domain reference speech signal and predicted speech signal. *STOI* scores can vary from [0, 1] [63]. The higher value of predicted intelligibility represents the better accuracy of separated speech. Nowadays, *STOI* is considered the standard measure for evaluating sound source separation performance [4], [67], [68]. Suppose for one T-F unit, the intermediate intelligibility measure is $v_k(\ell)$ as shown in TABLE 1. Here ℓ is intermediate frequency. The $T_k(m)$ and $P_k(m)$ are T-F units for clean speech signal and processed signal, respectively, for k^{th} DFT bin, and m is the time index belonging to a region of X consecutive T-F units. The $P'_k(m)$ denotes clipped and normalized processed speech signal. Suppose z belongs to the region of all existing frames; the objective intelligibility measure is obtained by taking the average of intermediate intelligibility measure over all bands and frames and represented mathematically by v [63] as formulated in TABLE 1. K is the one-third octave band number, and I is the total number of frames. PESQ is suggested by the International Telecommunication Union (ITU) [69], [64], covers the distortion due to telecommunication networks and measures separated speech signal quality. The PESQ estimates and compares the loudness spectra of desired and separated speech calculated by auditory transformation [69], [70].

PESQ scores range from [−0.5, 4.5], with the higher score representing good quality. PESQ can evaluate only one-way noise distortion or speech perceived by the receiver. It requires complex computations and the whole utterance access, which may be undesired. PESQ is a speech quality measure, while *STOI* is a speech intelligibility

TABLE 1. Performance Metrics for Speech Source Separation Algorithms.

Metrics	Mathematical Formulation	Descriptions
Source to Distortion Ratio (SDR) [61] Source to Interference Ratio (SIR) [61] Source to Artifact Ratio (SAR) [61] Source to Noise Ratio (SNR) [61]	$SDR = 10\ log_{10} \frac{\|d_{tar}\|^2}{\|e_{int} + e_{noi} + e_{art}\|^2}$ $SIR = 10\ log_{10} \frac{\|d_{tar}\|^2}{\|e_{int}\|^2}$ $SAR = 10\ log_{10} \frac{\|d_{tar} + e_{int} + e_{noi}\|^2}{\|e_{art}\|^2}$ $SNR = 10\ log_{10} \frac{\|d_{tar} + e_{int}\|^2}{\|e_{noi}\|^2}$	• These are numerical measures and can be defined with any kind of allowed distortion. • SIR, SAR, and SNR differentiate estimated errors due to interference, artifacts, and noise respectively. • Numerical measures have high precision for low-performance values than for high ones. • SDR involves maximum error term hence widely used metric than SIR, SAR, and SNR
Scale-invariance SDR ($SI-SDR$)[62] Scale-invariance SIR ($SI-SIR$) [62] Scale-invariance SAR ($SI-SAR$) [62] Scale-invariance SNR ($SI-SNR$) [62] [61][60]	$SI-SDR = 10\ log_{10} \frac{\|e_{tar}\|^2}{\|e\|^2}$ $SI-SIR = 10\ log_{10} \frac{\|e_{tar}\|^2}{\|e_{int}\|^2}$ $SI-SAR = 10\ log_{10} \frac{\|e_{tar}\|^2}{\|e_{art}\|^2}$ $SI-SNR = 10\ log_{10} \frac{\|e_{tar}\|^2}{\|e_{noi}\|^2}$	• These are numerical performance measures. • Scale invariance metrics produce solutions with minimum energy and maximize, the correlation between clean speech and predicted signal. • The direct relation between these metrics is relevant for non-stationary deep-learning applications. • Scaling the target than increasing a particular metric perceptually enhances the output.
Short Time Objective Intelligibility ($STOI$) [63]	$v_k(\ell) = \frac{\sum_m \left[(T_k(m) - \frac{1}{X}\sum_z T_k(z)\right] \left[P'_k(m) - \frac{1}{X}\sum_z P'_k(z)\right]}{\sqrt{\sum_m \left[(T_k(m) - \frac{1}{X}\sum_z T_k(z)\right]^2 \sum_m \left[P'_k(m) - \frac{1}{X}\sum_z P'_k(z)\right]^2}}$ $v = \frac{1}{IM}\sum_{k,\ell} m_k(\ell)$	• Measures the intelligibility content of speech signals. • Independent of distortion, interference, and noise effects.

measure [64], [70]. Some metrics can evaluate particular distortion while being meaningless for others. One or more than one numerical metric with intelligibility and quality metrics have been calculated in recent works for a more accurate evaluation of the separation works.

III. TRAINING OBJECTIVES

Training objectives are essential for training neural network models properly. Training targets belong to three categories, i.e., masking, mapping, and signal approximation (SA) based targets [71], [72] as in TABLE 2. Masking-based training targets are ideal time-frequency (T-F) masks that establish the time-frequency relationship of the desired speech signal and mixture signal. Ideal binary mask (IBM), ideal ratio mask (IRM), and complex ideal ratio mask belong to masking-based training targets. The exclusive allocation principle in auditory scene analysis [69] and the auditory masking phenomenon in audition [58] are motivations of the first training target, i.e., an ideal binary mask (IBM) in supervised monaural speech separation [71], [73], [74], [75]. A two-dimensional T-F illustration of the noisy signal is used to represent IBM is given in TABLE 2.

For the signal-to-noise ratio (SNR) greater than the threshold value (th), IBM assigns the value 1 and 0 otherwise. In IBM, the separated speech signal becomes distorted due to hard decisions regarding the masking. Hence IRM or soft mask was introduced to overcome the signal distortion associated with IBM.

In IRM, the time-frequency points of the mixed speech signal represent the ratio of the energy of the target speech signal to the energy of the mixed speech signal [76].

Let $y\,(m)$ is the mixed speech signal, $d\,(m)$ is the desired speech signal, and $\eta\,(m)$ is interference signal.

$$y\,(m) = d\,(m) + \eta\,(m) \tag{2}$$

STFT [58] of the mixture signal $y\,(m)$ can be represented as follows:

$$Y\,(m, n) = D\,(m, n) + N\,(m, n) \tag{3}$$

Here n and m represent the frequency index and time, respectively. $Y\,(m, n)$, $D\,(m, n)$ and $N\,(m, n)$ are the Fourier transform of mixed-signal, desired speech signal, and interference signal respectively. By multiplying the IRM function $M_{IRM}\,(m, n)$ with mixture signal $Y\,(m, n)$ [58] the clean

speech signal can be reconstructed as:

$$D\,(m,n) = Y\,(m,n)\,M_{IRM}\,(m,n) \tag{4}$$

$M_{IRM}(m,n)$ represents the T-F ideal ratio mask function formulated in TABLE 2, and β is a tunable parameter for changing the magnitude value of the mask. $|D\,(m,n)|$ and $|N\,(m,n)|$ are magnitude spectra of clean speech and interference noise, respectively. The ideal ratio mask [77] employs only magnitude information; however, the use of the desired signal spectrum phase information is also essential [60], [67]. Hence cIRM [78] was proposed, which uses both the magnitude and phase information of desired signal spectrogram to recover the target signal. The complex domain mixture and the clean speech signals spectrograms can be written in as:

$$Y\,(m,n) = Y_r\,(m,n) + jY_c\,(m,n) \tag{5}$$

$$\mathrm{D}\,(m,n) = D_r\,(m,n) + \mathrm{j}D_c\,(m,n) \tag{6}$$

$$M_{cIRM}\,(m,n) = M_{cIRM_r}\,(m,n) + jM_{cIRM_c}\,(m,n) \tag{7}$$

where $j \triangleq \sqrt{-1}$ and r denotes real and c denotes imaginary components of STFT. $\hat{M}_{cIRM_r}\,(m,n)$ and $\hat{M}_{cIRM_c}\,(m,n)$ [58] are real and imaginary parts of the estimated cIRM function. M_{cIRM} is the cIRM expressed as follows:

$$\begin{aligned} &M_{cIRM}\,(m,n) \\ &= \frac{Y_r\,(m,n)\,D_c\,(m,n) + Y_c\,(m,n)\,D\,(m,n)}{Y_r^2\,(m,n) + Y_c^2\,(m,n)} \\ &\quad + j\frac{Y_r\,(m,n)\,D_c\,(m,n) - Y_c\,(m,n)\,D_r\,(m,n)}{Y_r^2\,(m,n) + Y_c^2\,(m,n)} \end{aligned} \tag{8}$$

The cost function J_{cIRM} for cIRM is formulated in TABLE 2. Phase-sensitive mask (PSM) [67], [79] is an effectively calculated mask function for speech separation to become aware of the phase of speech signal using the phase information of spectrograms [80]. The T-F domain ideal PSM $M_{Ph}(m,n)$ for speaker separation can be formulated as in TABLE 2. $\theta_y\,(m,n)$ and $\theta_s(m,n)$ represents a phase of mixed-signal and clean speech source for source number s respectively, and $D_s\,(m,n)$ is clean speech signal of s^{th} source.

Mapping-based targets [77] are the spectra of the desired speech signal having the broad value range, i.e., $[0, +\infty)$ for all T-F points. In the mapping-based approaches, the magnitude spectrum of the target speaker trains the deep learning model. The cos function $J_{mapping}$ for mapping-based training targets is formulated in TABLE 2. Here $\hat{D}\,(m,n)$ is a spectrum of the predicted signal of desired speech source.

Hence the value of the cost function should be minimized to reduce the difference between the desired signal and the estimated signal [77]. However, the spectrum of the clean speech signal may take value belongs to the broad range, i.e., $[0, +\infty)$ at every T-F point. Hence mapping-based models are challenging to train [78] and obstacle to produce desired performance. SA based training targets overcome this challenge by estimating the desired speech signal in the range $[0, 1]$ at each T-F point.

Signal approximation-based training targets are signal spectrums calculated by multiplying the estimated mask with the mixture signal in the T-F domain with a range between $[0, 1]$ [58]. In signal approximation (SA) [77], mapping decides the training target, and masking estimates the desired speech. Hence SA is a combination of mapping and masking. Similar to the mapping-based algorithm, the magnitude spectra of the desired signal become the target to train the model. However, the predicted T-F mask and spectrum of mixture signals are multiplied to obtain an estimated speech spectrum as in the masking-based approach. The cost function J_{SA} [58] for the SA-based approach can be formulated as in TABLE 2. Here $\hat{M}_{SA}\,(m,n)$ is the predicted T-F mask used to obtain an estimated spectrum $\hat{D}\,(m,n) = Y\,(m,n)\,\hat{M}_{SA}\,(m,n)$ for the SA-based method.

Hence SA based approach increases accuracy in the source separation problem. SA based training targets considers only real terms while cSA [77] based training target uses both real and imaginary components of the signals to calculate target signals. In the complex domain, the cost functions of the cSA-based method [77] can be calculated as J_1 for the real term and J_2 for the imaginary term as shown in TABLE 2.

TABLE 2. Training Objectives for monaural Speech Source Separation Approaches.

Training Objectives	Mathematical Formulation
IBM [60]	$IBM = \begin{cases} 1, & if\ SNR(m,n) > th \\ 0, & otherwise \end{cases}$
IRM [76]	$M_{IRM}(m,n) = \left(\frac{\lvert D(m,n)\rvert^2}{\lvert D(m,n)\rvert^2 + \lvert N(m,n)\rvert^2}\right)^{\beta}$
cIRM [78]	$J_{cIRM} = \sum_m \sum_n [\left(\widehat{M}_{cIRM_r}(m,n) - M_{cIRM_r}(m,n)\right)^2 + \left(\widehat{M}_{cIRM_c}(m,n) - M_{cIRM_c}(m,n)\right)^2]$
Mapping [77]	$J_{mapping} = \sum_m \sum_n \left(\lvert\widehat{D}(m,n)\rvert - \lvert D(m,n)\rvert\right)^2$
SA [77]	$J_{SA} = \sum_m \sum_n \left[\left(\lvert Y(m,n)\widehat{M}_{SA}(m,n) - D(m,n)\rvert\right)^2\right]$
cSA [77]	$J_1 = \sum_m \sum_n [\widehat{M}_{cSA_r}(m,n)Y_r(m,n) - \widehat{M}_{cSA_c}(m,n)Y_c(m,n) - D_r(m,n)]^2$ $J_2 = \sum_m \sum_n [\widehat{M}_{cSA_r}(m,n)Y_c(m,n) - \widehat{M}_{cSA_c}(m,n)Y_r(m,n) - D_c(m,n)]^2$
PSM [80]	$M_{Ph}(m,n) = \frac{D_s(m,n)\,cos[\theta_y(m,n) - \theta_s(m,n)]}{Y(m,n)}$

$\hat{M}_{cSA_r}(m,n)$ and $\hat{M}_{cSA_c}(m,n)$ are real and imaginary parts of a complex signal approximation mask function.

IV. DATASETS

Monaural speech source separation methods have been worked with various benchmarked datasets. Speech source separation datasets contain mixture, separated, and noise signals to facilitate researchers for separation, enhancement, and de-noising tasks.

WSJ0 [81] corpus is created for automatic speech recognition (ASR) tasks. WSJ0-2mix [81] and WSJ0-3mix [81] are subsets of WSJ0 used to perform two and three-speaker separation tasks in many state-of-the-art techniques. It contains a speech signal of 30 hours spoken by 119 speakers. The WSJ0 hipster ambient mixture (WHAM!) [82] the dataset is a noisy version of the WSJ0-2mix suitable for speech signal de-noising tasks. The WHAM! contains two speaker mixture signals with Noise. Unique noise is added in the background to make it noisy. WSJ0-2mix and WSJ0-3mix are further extended to WSJ0-4mix [83] and WSJ0-5mix [83] by modifying the basic script of the WSJ0 dataset [83]. To create WSJ0-4mix and WSJ0-5mix, four and five speakers, respectively, are randomly selected and mixed at random 0-5 dB SNR values [83]. The WSJ0 hipster ambient mixture reverberant (WHAMR!) [84] is a reverberant and noisy extension of WHAM! [82]. It contains artificial reverberation with noise in the background. Texas Instruments Massachusetts Institute of Technology (TIMIT) [85] is acoustic-phonetics continuous speech corpus [85]. It contains 6300 utterances produced by 630 speakers. Each speaker speaks ten sentences [85]. Telecommunication and signal processing (TSP) [86] consists of 1444 sentences of 2.372 seconds and speaks by 24 speakers. This dataset also included children's speech signals [87]. SSC (Speech Separation challenge) was the standard corpus to evaluate the separation system in ICSLP 2006 [88]. It contains training, testing, and development sets separately. The dataset for training contains 17000 utterances from 34 speakers (18 males, 16 females) [89]. Tasting and development sets consist of separate noise and two talker sentences. Each set of two talker sentences consists of speech at six different SNR values $-9, -6, -3, 0, 3, 6$dB [89]. LibriSpeech [90] from Librivox audiobook [90] is a read corpus for ASR. This dataset has 470 hours of speech signals spoken by 1252 speakers. The LJSpeech [91] dataset consists of 12522 training and 578 testing utterances out of 13100 utterances with 1 to 10-second varying lengths. It is a single-speaker reading passage corpus [92]. The LibriMix [93] derived from LibriSpeech, and WHAM! Noises. It contains two and three-speaker recordings of separated, mixed, and noise signals making it beneficial for deep learning-based source separation tasks. The LibriMix is a freely available dataset. However, WSJ0 is the commercially available dataset. Recent works perform monaural speech source separation on both WSJ0 and LibriMix datasets. The LibriMix dataset can be extended to more than three speakers. Libri5mix, Libri10mix, Libri15mix, and Libri20mix are $5, 10, 15,$ and20 speaker datasets, respectively can be created by using the modified script of the LibriMix dataset [94], [95]. The VCTK dataset contains 109 speakers. Each speaker reads 2-6 seconds long 400 newspaper sentences in native English [96]. VCTK-2mix [96] is an open-source dataset derived from VCTK [96] and WHAM! Noises [96]. It can be used as a test dataset for source separation in a noisy environment and helps to perform cross-dataset experiments [96]. TABLE 3 describes benchmarked datasets for the source separation task

V. DEEP LEARNING-BASED MONAURAL SPEECH SOURCE SEPARATION TECHNIQUES

The impressive performance of deep learning in various research fields motivates the researchers to work on deep learning-based speaker separation problems. recent approaches are available in the T-F, time, and hybrid domains. techniques in the T-F domain transform the speech signal in the T-F domain before processing, while time-domain techniques perform separation in the time domain only. the hybrid method performs the separation in both domains. techniques in T-F, time, and hybrid domains can be explained in the following section.

A. TIME-FREQUENCY DOMAIN SPEECH SOURCE SEPARATION TECHNIQUES

The T-F domain approaches use concepts of clustering, permutation, grouping, and phase reconstruction with deep learning models to perform separation tasks. These approaches can be classified as clustering, permutation, multi-task learning, CASA, and phase reconstruction-based, as presented in the following section.

1) DNN AND RNN-BASED APPROACHES

Deep neural networks (DNN) based approaches are the first to solve speaker separation problems using deep learning [27]. These DNNs are feed-forward networks without recurrent connections. These methods outperform the signal-processing domain speaker separation approaches and motivate researchers to use deep learning for the monaural speech source separation task [27]. NMF uses only positive templates to model the source signals; however, in real-world applications, sources are non-linear and may generate both positive and negative values [97]. Hence non-linear DNN models give more promising results than NMF models.

The DNN models are trained to classify the sources present in the mixture signal. These models capture contextual information by concatenating neighboring features of audio signals, e.g., magnitude spectra, Mel-frequency cepstral coefficients (MFCCs), etc. However, the increase in the number of concatenating neighboring features increases the complexity of the neural network models due to the limitation in incrementing the size of the concatenating window [4]. Hence instead of deep neural networks, recurrent neural networks (RNNs) are used for temporal information of time series audio signals [4]. RNNs employ memory from

TABLE 3. Available Datasets for Monaural Speech Source Separation Techniques.

Datasets	Corpus	Descriptions	Year
WSJ0 (Wall Street Journal) [81]	Commercially available at LDC (Linguistic Data Consortium)	The WSJ0 is created for automatic speech recognition (ASR) tasks and contains speech signals of 90 hours spoken by 119 speakers.	1992
WSJ0-2mix [81]	Commercially available at LDC	Two speaker mixtures are prepared by mixing the utterance of WSJ0 in two subsets and contain training, validation, and test sets.	1992
WSJ0-3mix [81]	Commercially available at LDC	The mixture of three speakers is prepared by mixing the utterance of WSJ0 in two subsets and contains training, validation, and test sets.	1992
WSJ0-4mix and WSJ0-5mix [83]	https://enk100.github.io/speaker_separation/	The WSJ0-4mix and WSJ0-5mix are created by randomly chosen four and five speakers from the WSJ0 corpus and mixed at random SNR values $0-5$ dB.	2020
WHAM! (The WSJ0 Hipster Ambient Mixture) [82]	Available freely at https://wham.whisper.ai/	WHAM! is a noisy version of WSJ0-2mix. To make two-speaker mixtures of WSJ0-2Mix noisy unique noise is added in the background. It is freely available.	2019
WHAM Reverberant (WHAMR!) [84]	Available freely at https://wham.whisper.ai/	WHAMR! is an extension of WHAM! with artificial reverberation and noise in the background.	2019
TIMIT (Texas Instruments Massachusetts Institute of Technology) [84]	Acoustic phonetic continuous speech corpus.	The TIMIT dataset contains 6300 sentences spoken by 630 speakers; each speaker speaks ten sentences.	1993
TSP (Telecommunication and Signal Processing) [86]	Available freely at https://www.mmsp.ece.mcgill.ca/Documents/Data/	The TSP dataset consists of 1444 utterances (half male, half female) spoken by 24 speakers.	2002
SSC (Speech Separation Challenge) [88]	Available freely at http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm	The training set contains 17000 utterances spoken by 34. Tasting and development sets have two talker sentences at six different SNR values $-9, -6, -3, 0, 3, 6$ dB.	2006
LibriSpeech [90]	It is freely available at Open SLR (open speech and language resources)	On the basis of LibriVox audiobooks, a read corpus LibriSpeech is created for ASR. This dataset contains 1000 hours of speech signals	2015
LJspeech [91]	Available freely at https://keithito.com/LJ-Speech-Dataset/	The LJspeech consists of 12522 training and 578 testing utterances out of 13100 utterances with 1 to 10-second varying lengths. It is a single-speaker reading passage corpus.	2017
LibriMix [93]	Available at https://github.com/JorisCos/LibriMix	The LibriMix is derived from the LibriSpeech and WHAM! Noises. It is an open-source alternative to the WSJ0 dataset. This dataset contains 470 hours of speech signals spoken by 1252 speakers.	2020
Libri2mix [93]	Available at https://github.com/JorisCos/LibriMix	Libri2mix is two speaker mixture dataset from the LibriMix corpus. It is a freely available dataset and contains 292 hours of speech signals of the clean and noisy mixture signals.	2020
Libri3mix [93]	Available at https://github.com/JorisCos/LibriMix	Libri3mix is two speaker mixture dataset from the LibriMix corpus. It is a freely available dataset of clean and noisy mixture signals.	2020
Libri5mix, Libri10mix, Libri15mix, Libri20mix [94], [95]	Available at https://github.com/ShakedDovrat/LibriMix	5, 10, 15, and 20 speaker datasets, respectively, can be created using the modified script of the LibriMix dataset.	2021
VCTK [96]	https://datashare.ed.ac.uk/handle/10283/3443	VCTK comprises 109 naïve English speakers with various ancents. Each speaker reads 400 newspaper sentences of 2-6 seconds long duration.	2017
VCTK-2mix [96]	https://github.com/JorisCos/VCTK-2Mix	VCTK-2mix is two speaker mixture dataset created from the VCTK and WHAM! Noises. This dataset contains 9 hours of speech signals spoken by 108 speakers. It is used as a cross-dataset for testing the performance of models.	2020

previous time steps. Hierarchical RNNs, also known as deep recurrent neural networks (DRNNs), can provide information through multiple time scales [78], [98] and outperforms then DNN-based approaches.

The long-short-term memory recurrent neural network (LSTM RNN) method uses cSA as a training target and produces real and imaginary components of the output separately. The detailed working of the LSTM block is given in [98]. Complex domain monaural source separation approaches utilize phase information of the target speech signal to retrieve the target audio signals. The LSTM RNN uses temporal information from time series data. Two parallel LSTM RNNs with similar configurations simultaneously calculate real and imaginary terms in the cSA-based LSTM RNN approach [77]. The combination of features increases network and system efficiencies. The compound features i.e., the amplitude modulation spectrogram (AMS) [99] (calculated using 64-channel gammatone filterbank [100]), relative spectral transformation, and perceptual linear prediction (RASTA-PLP) [101], Mel-frequency cepstral coefficients (MFCC), cochleagram response, and their deltas are extracted using feature extraction unit [69]. Fig. 3 shows the block diagram of the LSTM RNN method based on cSA [77]. During the training stage, LSTM RNN 1 uses real, and LSTM RNN 2 uses imaginary components of the spectrogram of target speech sources. The calculated complex mask and the mixture signal spectrum are multiplied to obtain separated outputs. The predicted complex T-F mask will be updated in each iteration, reducing the variation between desired speech and calculated speech signals. During the test period, the features of the mixture signals are applied as input to the trained LSTM RNNs. Then, the compound module combines the predicted real and imaginary components of the output signal and the reconstruction module reconstructs the estimated output speech. The cSA-based LSTM RNN algorithm has two advantages over the SA-based DNN algorithms; (1) the SA-based DNN approaches utilize only the magnitude spectrum to calculate mask function. Finally, the unprocessed phase spectrum and calculated mask function of the mixture signal are used to reconstruct the separated signal spectrum. However, the cSA-based LSTM RNN method utilizes information regarding both the magnitude and phase of the desired signal to calculate the mask function [77]. (2) The LSTM-RNN efficiently utilizes the temporal information after training LSTM RNN architecture represents good generalization ability [77].

Ensemble learning [76] motivates to train small DNNs and connects them to perform a big task rather than training a big model to perform the big task. Ensemble learning provides very high performance for regression and classification. It tends to combine small models to provide an enhanced range and flexible representation of the generalized problem [76]. Ensembles of DNN are used to form the multi-context network. The training target for each neural network is the ideal ratio mask or signal approximation. Multi-context networks are of two types: multi-context averaging (MCA) and multi-context stacking (MCS) [76]. The MCA network averages all outputs from small ensembles of DNN to obtain final outcome. However, ensembles in MCS at different context lengths are connected serially to produce the final result. The ensemble learning approach is suitable for efficient training but compromises with designing complexity [76].

2) CLUSTERING-BASED APPROACHES

DC [29] is a speaker-independent speech source separation technique that can work with any number of speakers. It transforms T-F bins of spectra of mixture signal into high dimensional embedding space and produces embedding vectors. Then K-means clustering clusters the embedding vectors to separate the sources. It resolves the output dimension mismatch problem of PIT. Objective function measurement between embedding sources instead of ground truth speech signals reduces the efficiency of mapping sources properly. This limitation is overcome by DANet [30]. The DANet also produces high dimensional embedding space, but instead of clustering, it creates attractor points and reduces the distance between T-F bins corresponding to each source. Attractors are the centroid points of sources in embedding space that helps to separate T-F bins belonging to an individual source. Embedding spaces are updated in each iteration to minimize errors in reconstruction. This approach faces a center mismatch problem in which the true attractor points duffers from the estimated attractor point. The center mismatch problem causes the prediction of wrong sources. Anchor DANet (ADANet) [102] approach overcomes the center mismatch problem by considering the anchors instead of attractors in embedding space. Anchors are several reference trainable points used in both the training and test stage to estimate source assignment. ADANet has improved performance as compared to all existing DC-based approaches.

Attention deep clustering network (ADCNet) [103] is a recent T-F domain approach that uses multi-head self-attention and deep clustering to perform speaker separation [103]. Inspiring from human auditory attention, ADCNet optimizes multi-head self-attention and deep clustering simultaneously [103].

This method captures comprehensive information on multiple time scales using multi-head self-attention. Basic deep clustering approach uses k-means clustering which requires number of clusters should be known previously hence not effective for big data [103]. ADCNet uses density-based canopy k-means algorithm to overcome limitation of k-means algorithm. This improved k-means algorithm does not require cluster number previously [103]. Encoder squash-norm deep clustering (ESDC) [12] is a state-of-the-art T-F domain single channel speaker separation method. It enhances discriminative learning ability of high dimensional vectors by performing input feature encoding, embedding vector training, vector normalization, and vector clustering [12]. The node encoder establishes correlation using adjacency-based

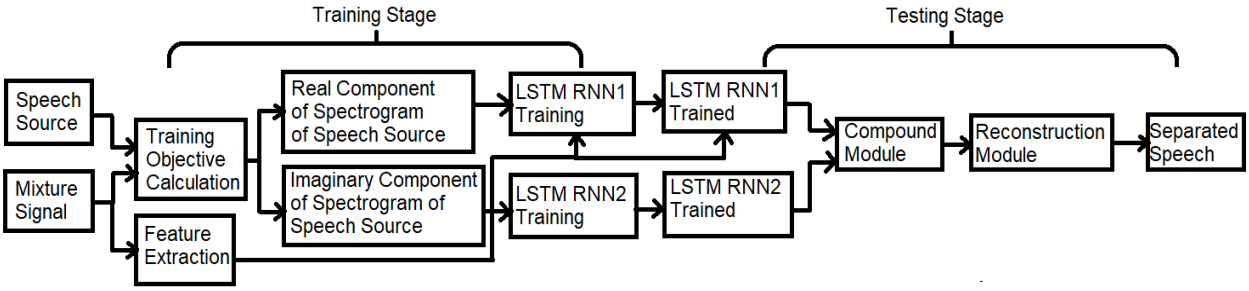

FIGURE 3. The Architecture of cSA-based LSTM RNN [77].

similarity between neighboring information and calculates the scaler product features using input feature vectors. The scaler product features represent the relationship between input vectors. The training stage discriminates these feature vectors to improve the performance of separation approach. Then squash-norm normalization is used in vector normalization stage to increase the discriminative capability of embedding feature vectors. This stage coverts the short vectors to zero vectors and long vectors to unit vectors. Finally, clustering stage clusters the squash-norm embedding vectors using various clustering methods [12].

3) PERMUTATION-BASED APPROACHES

Permutation-based approaches involve permutation invariant training (PIT) [31]. All possible permutations for mixed sources are pooled in PIT, and the lowest error permutation is used to update the network. PIT solves the permutation problem but has an output dimension mismatch problem. Frame level PIT (tPIT) [31], utterance level PIT (uPIT) [77], [80] and constrained uPIT (cuPIT) [104] are the permutation-based approaches. tPIT works at the frame level to perform speech source separation. It needs speaker tracking due to frame level discontinuity. In contrast, real-world problems are at the utterance level. The uPIT overcomes discontinuity of the frame in tPIT by using BLSTM trained at utterance level criterion to align the frames of the same speakers. The cuPIT produces a delta-acceleration coefficient cost function by adding acceleration and weighted delta of output frames.

The CuPIT is the best PIT method, but due to its complexity tPIT and uPIT are frequently used methods. The one and rest permutation invariant training (OR-PIT) [105] is a monaural talker independent multi-speaker speech source separation algorithm and uses tPIT in its architecture. It recursively uses a source separation network to progressively separate sources from the mixture. The source separation network separates one source at a time. The remaining mixture signal is again recursively applied to the separation network to further separate the sources from the mixture signal [105]. In OR-PIT, easy-to-separate speakers are always separated first with high separation quality. However, separation quality degrades with further separation. This approach with iteration termination criterion knows when to stop the iteration [105].

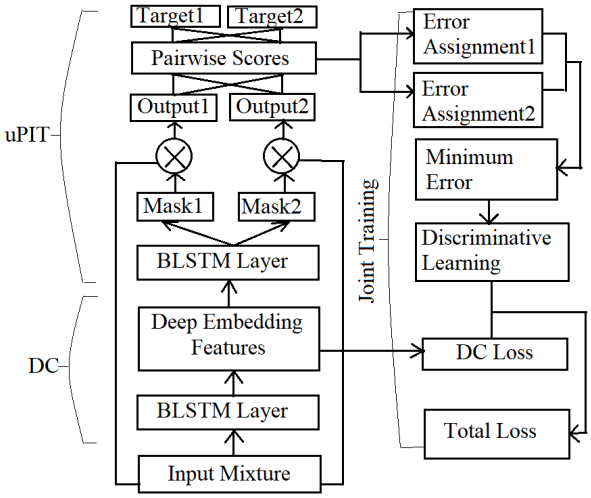

FIGURE 4. The Block Diagram of uPIT-DEF-DL [106].

The uPIT+DEF+DL [104] (uPIT + deep embedding features (DEF) + discriminative learning (DL)) is a T-F domain discriminative learning method with deep embedding features [106]. Single-channel speaker separation can be considered a permutation problem [29], [77].

PIT reduces the distance between the same speech signals but does not increase the distance between different speech signals. tPIT and uPIT have output dimension mismatch problems hence in many approaches PIT is used with DC. The uPIT+DEF+DL is one of the approaches that uses uPIT and DC to perform separation tasks. The Block diagram of uPIT+DEF+DL is shown in Fig. 4 [106]. Deep clustering (DC+) [29] stage extracts deep embedding features (DEF) by producing clusters in embedding space known as embedding vectors. These embedding vectors are used by uPIT stage for separating the sources in the pre-processing stage. Separated signals using DC and uPIT still have possibilities of remixing. Discriminative learning (DL) [107], [108], [109], [110], [111] reduces the chances of remixing of separated signals using discriminative loss function. This algorithm has four stages, DC, uPIT, discriminative learning, and the joint training. In the deep clustering stage, a trained bidirectional long-short term memory (BLSTM) [112] network extracts deep

embedding features (DEF) by projecting each T-F bin of the amplitude spectra of a mixture signal $|Y(m, n)|$ into the D-dimensional embedding vector E_m. The DEF extractor cost function J_{dc} can be formulated using following equation.

$$J_{dc} = \left\| E_m E_m^T - A_s A_s^T \right\|_F^2 = \left\| E_m E_m^T \right\|_F^2 - 2 \left\| E_m^T A_s \right\|_F^2 + \left\| A_s A_s^T \right\|_F^2 \quad (9)$$

A_s is a binary matrix of membership function for source s in each T-F bin. $\|*\|_F^2$ represents the square Frobenius norm. The value of the matrix $A_s = 1$ for the s^{th} source having the maximum energy compared to other sources, and $A_s = 0$ otherwise. The deep embedding vectors from DC are applied as input to uPIT to estimate soft masks for every source. uPIT selects an optimal permutation having a minimum value of mean square error cost function J_{uPIT} at utterance level from all speaker permutations ($\mathcal{P}$).

$$J_{uPIT} = \underset{\theta_s \epsilon N}{argmin} \sum_{s=1}^{S} \left\| |Y(m, n)| \otimes \tilde{M_{Ph}}(m, n) - D_s(m, n) \cos(\theta_y(m, n) - \theta_s(m, n)) \right\|_F^2 \quad (10)$$

uPIT targets to minimize J_{uPIT} to make the output predictions and their corresponding target sources more similar. Discriminative learning (DL) helps to identify the difference between target and interferences by reducing the difference between the predicted and the corresponding target so that possibility of remixing decreases. Suppose the selected permutation is φ^* and has a minimum mean square error value among all permutations. Then DL cost function J_{DL} can be computed as

$$J_{DL} = \varphi^* - \sum_{\varphi \neq \varphi^*, \varphi \epsilon N} \mu\varphi \quad (11)$$

Here φ represents permutation from $\mathcal{P}$ excluding φ^*, $\mu \geq 0$ is the parameter for regularization of φ. When $\mu = 0$, J_{DL}, and J_{uPIT} are the same, this is the condition for no discriminative learning. J_{DL} and J_{uPIT} are jointly calculated in joint training to obtain embedding features effectively [112]. The joint training loss function J_{joint} can be formulated as:

$$J_{joint} = \Upsilon J_{dc} + (1 - \Upsilon) J_{DL} = \Upsilon J_{dc} + (1 - \Upsilon) \left(\varphi^* - \sum_{\varphi \neq \varphi^*, \varphi \epsilon N} \mu\varphi \right) \quad (12)$$

Here $\Upsilon \epsilon [0, 1]$ helps to control J_{dc} and J_{DL} weights. The end-to-end post filter (E2EPF) [112] method with deep attention fusion features reduces residual interferences from pre-separated speech signals [112]. E2EPF uses both magnitude and phase information of pre-separated time-domain signals to maintain correct magnitude and phase values. The E2EPF reduces residual interferences in the output signals of uPIT+DEF+DL.

4) MULTI-TASK LEARNING-BASED APPROACHES

Source separation approaches use multi-task learning (MTL) to perform various tasks simultaneously using similar information from different tasks to improve the model training. Many recent approaches use MTL to obtain models working simultaneously on different tasks of the same information. Convolutional BLSTM DNN (CBLDNN) [113] is a speaker-independent speech separation approach. It uses generative adversarial training (GAT) created by the generative adversarial network (GAN) and MTL. GAN has a generator and discriminator network. The generator of GAN generates speech signals using the observed mapping between the mixture signal feature and mask functions. Then discriminator differentiates generated speech and actual speech features. MTL extracts the fbank-pitch-based features to improve the model's training [113]. This method reduces numerical mean square error and simultaneously increases the perceptual quality of speech. Shifted delta coefficient with multi-task learning using grid LSTM (SDC-MTL-Grid) [114] approach deals with single-channel speaker separation. During end-to-end training shifted delta coefficient (SDC) objective considers the long range of time dynamics to calculate mask functions. These contextual temporal dynamics align the same speaker's frames on the same side [114]. Multi-task learning (MTL) enhances the outcomes of a single task using simultaneous learning of more related tasks. MTL predicts T-F labels like silence labels, single labels, and overlapped labels of mixture signals. SDC and MTL jointly worked with grid LSTM to obtain impressive results and are known as SDC-MTL-Grid. MTL informs SDC about the overlapping regions during the mask estimation because speech separation aims to separate overlapping parts of the mixture signal [114]. Chimera networks incorporate MTL and DC with mask inference [104]. It uses mask inference after the embedding layer. However, in chimera++ [104] network, the mask inference is at the output of the BLSTM hidden layer, which reduces the complexity of the network and increases working speed.

5) COMPUTATIONAL AUDITORY SENSE ANALYSIS (CASA) -BASED APPROACHES

CASA end-to-end (CASA-E2E) [115] is a speaker-independent single-channel speaker separation approach. It uses PIT and DC in two stages of CASA, i.e., in the simultaneous and sequential grouping, respectively. In the simultaneous grouping stage, frame-level PIT trains BLSTM RNN to perform separation at the frame level. In the sequential grouping stage, clustering groups frame-level separated spectra into utterance levels to identify the speakers.

Deep computational auditory sense analysis (deep CASA) employs simultaneous and sequential grouping [116], [117]. The simultaneous grouping works at the frame level to differentiate the desired signal from the mixture. In a situation where more than one speaker is to be separated, then separated frame-level spectrums are applied to the sequential grouping stage to track the desired speaker [116].

Simultaneous grouping in Fig. 5(a) works at the frame level to isolate two speakers. The mixture signal spectrum $Y(m, n)$ is the input to a Dense-UNet [116], [117] to predict two complex ratio masks $cIRM1(m, n)$ and $cIRM2(m, n)$. The masks and the mixture signal $Y(m, n)$ are multiplied to produce two outputs, $\hat{P}_1(m, n)$ and $\hat{P}_2(m, n)$ represents the estimated STFT of the two speakers in a complex domain. Permutation invariant training (PIT) [77], [116] is the popularly used method to train a neural network for more than one target signal

PIT examines permutations for all possible target output signals and permutation having minimum loss optimizes the network during training. Frame level PIT (tPIT) and utterance-level PIT (uPIT) are two types of PIT [116]. In tPIT, there is frame-by-frame variation between permutations of the target output signals. However, in uPIT, each training utterance uses a fixed permutation. In simultaneous grouping, tPIT trains Dense-UNet, and tPIT loss organizes complex outputs $\hat{P}_1(m, n)$ and $\hat{P}_2(m, n)$ into two streams, $\hat{P}_{o1}(m, n)$ and $\hat{P}_{o2}(m, n)$, then inverse STFT of these organized signals produces two time-domain signals $\hat{P}_{o1}(m)$ and $\hat{P}_{o2}(m)$. Signal to noise ratio (SNR) objective helps to properly train the model so that accuracy of separation increases. The sequential grouping stage, as in [116], separates the frame-level predicted spectrum $\hat{P}_1(m, n)$ and $\hat{P}_2(m, n)$ of two speakers. Fig. 5(b) represents a sequential grouping. The input to the sequential grouping stage is a stack of $Y(m, n)$, $\hat{P}_1(m, n)$ and $\hat{P}_2(m, n)$. A temporal convolutional network (TCN) [118] comprises dilated convolutional blocks that propel each frame-level input to a D-dimensional embedding vector $E(w)$. Two-dimensional vector $I\,(w)$ specifies the target labels for TCN training. If output 1 is speaker 1, and output 2 is speaker 2 then $I\,(w) = [0, 1]$ in Dense-UNet, otherwise, $I\,(w) = [1, 0]$. During training, $E(w)$ for the same tPIT pairing arranged closer by a weighted objective function between $E(w)$ and $I\,(w)$, and otherwise to become farther apart.

In simultaneous grouping, the K-means algorithm performs clustering of $E(w)$ and produces a binary value for each frame to arrange the frame-level outputs as the final outputs of deep CASA. The causality of the signal can be considered to make deep CASA causal [119]. But it degrades the separation performance.

Listen and Group [120] approach combines both listening and grouping. It always keeps the order of the output unchanged since it is an autoregressive method. In listening, midlevel representation of magnitude spectrogram of source and mixture signals are simultaneously created. The grouping stage uses these spectra to estimate separated sources.

6) PHASE RECONSTRUCTION-BASED APPROACHES

Sign prediction net [121] is a phase reconstruction-based approach for T-F domain deep learning-based monaural speaker-independent speech source separation approaches. The reconstructed signal's magnitude is not good with phase inconsistency. Hence it predicts the sign and computes the estimated phases [121]. Waveform approximation multiple input spectrogram inverse (WAMISI) [122] uses T-F masking, STFT, and inverse STFT as layers of a deep network to perform multi-speaker monaural speech source separation tasks. It computes loss on reconstructed signal to incorporate error due to phase inconsistency [59]. Sign prediction net and WAMISI use phase spectra of speech signals during separation to overcome phase mismatch problems. These methods are accurate but sacrifice performance. Comparison TABLE 4 shows the advantages and disadvantages of state-of-the-art T-F domain speech source separation approaches.

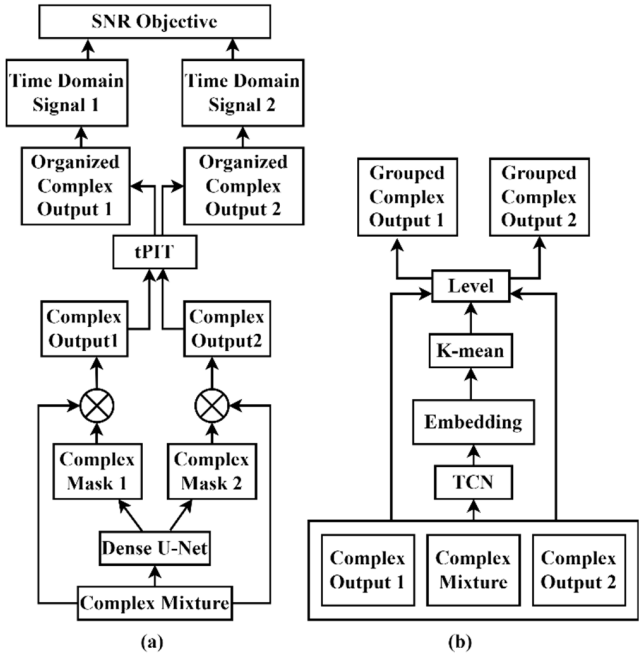

FIGURE 5. Deep CASA Stages. (a) Simultaneous Grouping Stage, (b) Sequential Grouping Stage [116].

B. TIME DOMAIN SPEECH SOURCE SEPARATION TECHNIQUES

The limitations of T-F domain approaches are time-frequency decomposition, long-duration window, and phase magnitude decoupling, which act as obstacles to obtaining the required frequency resolution. Most end-to-end time-domain speech source separation techniques solve these problems using the encoder-decoder framework. The time-domain methods can be efficiently used in real-time applications. These approaches do not use STFT transformation. These techniques work to design encoder separation and decoder modules to perform the separation of speech sources. The encoder transforms the audio signal into a data-driven representation form. The separation module is designed to calculate the mask function using data-driven representation from the encoder. This mask function is multiplied with a mixture signal to separate the speakers. The decoder performs inverse transformation to the encoder and converts the separated

TABLE 4. Advantages and Disadvantages of T-F domain Speech source Separation Techniques.

S. No.	Algorithms	Advantages	Disadvantages
1.	DNN [27]	• First speech source separation work based on deep learning instead of signal processing.	• Permutation and output dimension mismatch problem. • Processes only non-negative input data however real-world signals can be negative or positive. • Does not work with previous time steps.
2.	RNN [4]	• Works at multiple time steps with recurrent connections. • Depending on type of recurrent connection, can processes different temporal context.	• Can separate only fixed number of sources. • More training time due to sequential RNN.
3.	LSTM RNN [77]	• Perform Long context modeling due to LSTM. • Processes signals in complex domain. • Uses processed phase information to separate the sources.	• Phase processing doubles the parameters in model.
4.	Ensemble Learning [76]	• Small DNN can be trained to work at different context. • With ensemble learning DNN are able to work at multiple contexts.	• Undefined relation between the outcomes of ensembles of DNN. • Unable to work at long context information.
5.	Deep Clustering (DC) [29]	• Can be used to separate any number of sources. • Solves both permutation ambiguity and output dimension mismatch problems.	• Objective function defines the relation between embedding vectors not between separated signals. • The value of cluster number should be known previously. • K-means clustering algorithm limits the separation performance.
6.	Attention Deep Clustering (ADCNet) [103]	• This method uses attention mechanism with deep clustering in T-F domain. • Aims to designed the model using attention mechanism similar to human auditory attention. • Overcome the limitation of k-means clustering by using density-based canopy k-means clustering.	• Used multi-head self-attention is not similar to human auditory attention. • Performance is not as improved as time domain methods.
7.	Encoder Squash-Norm Deep Clustering (ESDC) [12]	• Adjacency-based similarity increases correlation information between the embedding vectors. • Embedding vectors normalized using Squash-norm increases discriminative ability of the model.	• State-of-the-art T-F domain method but performance is limited as compared to time domain methods.
8.	Deep Attractor Network (DANet) [30]	• Attractors in high dimension embedding space are optimal instead of direct use of embedding vectors for mask calculation. • Overcome permutation and output dimension mismatch problems.	• Center mismatch problem due to different attractor generation processes in the training and test stage. • Run time attractor calculation increases the complexity of the approach. • K-means clustering limits the separation performance.
9.	Frame Level Permutation Invariant Training (tPIT) [31]	• From all possible permutations lowest error permutation is used to update the network. • Solves permutation problem. • Reduces distance between the same speaker.	• Mismatching of output dimension and can separate only fixed number of sources. • Reduces distance between the same speaker but does not differentiate the different speskers. • Possibility of remixing separated sources. • Frame level separation is undesired for real-world scenario hence need speaker tracking.
10.	Utterance Level Permutation Invariant Training (uPIT) [81]	• Solves level ambiguity problem. • Solves the frame discontinuity problem with PIT. • Work on utterance level instead of frame level. • No need for speaker tracking. • Reduces distance between the same speaker.	• The uPIT and tPIT have similar disadvantages like output dimension ambiguity problem, remixing of separated sources, and processing only a fixed number of sources.
11.	Chimera++ [104]	• Combines both DC and PIT by using multi-task learning.	• Separation model separately outputs the DC and PIT instead of fusing them deeply.

TABLE 4. *(Continued.)* **Advantages and Disadvantages of T-F domain Speech source Separation Techniques.**

		• Takes advantage of both DC and PIT.	• Not overcome the disadvantages of both DC and PIT.
12.	Deep CASA [116]	• Combines both DC and PIT deeply. • Uses frame-level PIT and then DC for speaker tracing.	• K-means clustering limits the algorithm performance. • It works at frame level however real-world speech signals are at utterance level.
13.	uPIT+DEF+DL [106]	• Uses DC, PIT, and discriminative learning. • BLSTM replaces K-means for clustering.	• Separated speech has residual interference.

speaker's signals into an understandable form. Time domain encoder-decoder frameworks for speech source separation can be categorized as recurrent neural networks (RNN), convolutional neural networks (CNN), and transformer-based approaches. However, Wavesplit works in the time domain without an encoder-decoder framework.

1) RNN-BASED APPROACHES

RNN-based techniques use LSTM, BLSTM, and RNN in their separation modules. Time-domain audio separation network (TasNet) [59] is an encoder-decoder framework. The encoder of TasNet estimates weights of mixture signal using one dimensional convolutional (1D Conv) layer followed by ReLU and sigmoid activation functions. The convolutive output of both activation functions is given to the separation module. The separation module consists of deep LSTM layers followed by a fully-connected layer with a soft mask activation function to calculate the mask function. The decoder performs transpose 1D Conv operation on mask and mixture signal multiplication to obtain a time-domain separated signal [59]. The separation module in TasNet LSTM [59] consists of unidirectional LSTM layers to consider causality for real-time systems. TasNet BLSTM [59], [123] uses bidirectional LSTM layers in separation modules for noncausal systems. 1D Conv layer in the encoder of TasNet has a short receptive field less than the length of the input sequence and hence cannot work with utterance level framework [59]. DPRNN [124] replaces the one-dimensional convolutional neural network in TasNet. It is smaller than TasNet and can work with long sequences by constructing a deep network using RNN layers. It consists of a segmentation layer, DPRNN layer, and overlap-add layer. The segmentation layer divides long input sequences into local chunks (intra-chunks) and global chunks (inter-chunks) [124]. In DPRNN stage two RNNs, an intra-chunk and inter-chunk RNN performs iterative and alternative processing of intra-and inter-chunks, respectively. Inter-chunk RNN aggregates the output from intra-chunks to perform utterance-level processing, then the overlap-add stage adds all the segments to obtain a separated source signal. The global processing stage of DPRNN suffers the recurrent connection problem and limits the performance of the approach [124]. It also uses positional encoding to know the sequence order information. Improved transformer [125] integrates RNN instead of positional encoding because positional encoding is not reliable for dual-path networks and creates model divergence during training.

Gated DPRNN [83] separates multiple voices simultaneously using gated neural networks. It mainly focuses on separating an unknown number of multiple speakers. The complexity and performance of two and three-speaker separation approaches decrease quadratically with an increased number of speakers. However, the complexity and performance of Gated DPRNN decrease linearly with an increased number of speakers [83].

2) CNN-BASED APPROACHES

CNN-based approaches use convolutional neural networks in their separation modules. The fully convolutional TasNet (ConvTasNet) [60] consists of only convolutional layers in all processing stages. It consists of an encoder-decoder and separation module similar to TasNet. Instead of a deep LSTM network, the separation module consists of a stacked dilated 1D convolutional block similar to the temporal convolutional network (TCN) [118]. The convolutional operation processes consecutive segments parallelly to increase processing speed and decrease model size. It incorporates global layer normalization (gLN) for causal systems and cumulative layer normalization for noncausal systems [60].

Speaker attractor network (SANet) [126] is an improved version of DANet. It uses TCN, similar to ConvTasNet, to create embedding vectors. Then attractors from these embedding vectors are calculated using mask-weighted average during training and approximated during the test phase using the k-means centroid of the embeddings [126]. In SANet number of speakers during training and testing can be different [126].

Neural architecture search (NAS) [127] is an artificial neural network technique for searching best model structure and minimizing human interaction. NasTasNet provides search space for ConvTasNet using candidate operation. It helps to obtain better design parameters for ConTasNet [127] and reduces GPU utilization with the best architecture. The auxiliary loss method with NAS is better for updating the parameters and achieving a balanced architecture for ConvTasNet [127].

Channel-aware audio separation network (CasNet) [128] is similar to TasNet with a channel encoder and separates the mixture speaker signal with the help of channel embeddings and the FilM technique [128]. It enhances the channel robustness of TasNet models. The channel encoder of CasNet consists of a residual net and a pooling layer. The residual net consists of two sub-blocks, i.e., the convolutional block and the residual block [128]. Convolutional block composed of a 1D convolutional layer followed by ReLU activation and batch normalization operations. The residual block has two convolutional blocks and a squeeze and excited layer [128].

3) TRANSFORMER-BASED APPROACHES

The basic transformer [125] consists of the encoder-decoder with multi-head attention for word-to-speech conversion and uses RNN and convolutional models. Dual-Path Transformer Network (DPTNet) [125] uses an improved transformer to model speech sequences with context-aware modeling of extremely long sequences. It has an encoder, decoder, and separation layer followed by a ReLU encoder activation function. The separation layer is a dual-path network with an improved transformer to calculate the mask function. The encoder output is segmented into overlapped intra- and inters-chunks. Intra-and inter-transformers process segmented chunks at the utterance level [125]. The dual-path transformer stage can be repeated further. A 2D convolutional layer processes the last inter-transformer output to calculate the mask function for each source. Overlap-add transforms the mask function into sequences. Now mask signal is multiplied with the mixture signal to obtain masked encoder features for a particular source. The decoder converts masked encoder features into separated speech signals by performing transposed encoder operations [125].

Globally attentive and locally recurrent (GALR) [129] network takes advantage of both attention and recurrent mechanism alternatively and iteratively. It uses BLSTM for local context modeling and multi-head attention for global context modeling [129]. The GALR is globally attentive and locally recurrent, while DPTNet is locally attentive and globally recurrent [129].

The Sepformer [130] is an RNN-free transformer-based model for speech separation. It consists of multi-head attention and feed-forward layers. It learns both short and long-term dependencies with a dual-path framework similar to DPRNN and uses a multi-scale pipeline, which consists of a transformer [130]. The Sepformer performs permutations between intra-and inter-transformer to model long-term dependencies across chunks [130].

Time-domain adaptive attention network (TAANet) [11] has two attention networks, channel attention and spatial attention for local modeling and the self-attention network for global modeling. For local modeling, it works on frame level with BLSTM, and for global modeling, it works on utterance level. Self-attention can pay more attention to the long-term dependency of the speech sequence by calculating the correlation between all parts at different time scales.

The dual path hybrid attention network (DPHA-Net) [131] is a transformer-based approach and utilizes multistage aggregation training (MAT) strategy [131]. The MAT is multistage training with improved feature selective aggregation ability. Similar to transformer-based approaches, DPHA-Net comprises encoding and chunking, separation, and overlap-add and decoding stages. The encoder consists of 1D- convolutional layer with the ReLU activation function to transfer 1D input sequences to 2D output sequences. The output of the encoder is divided into chunks to produce a 3D processable tensor. DPHA-Net separation module processes this 3D tensor to predict mask function through intra-and inter-chunk processing units [131]. These units have similar architecture and consist of multi-head self-attention (MHSA), element-wise attention (EA), adaptive feature fusion, global layer normalization (gLN), and permutation operation. The separation module of DPHA is repeated in the required number of stages. The outputs of the present and previous stages are aggregated to produce the final outcome of a particular stage and separation module [131]. EA unit consists of two layers of gated recurrent unit (GRU), followed by the sigmoid activation function, then a second GRU to capture the context information at various time steps. The adaptive feature fusion (AFF) unit consists of channel-wise attention and temporal attention operations. AFF enhances the feature extraction capability of the network by suggesting suitable attention and channel characteristics for relevant time steps and channels [131].

4) MULTI-SCALE FUSION-BASED APPROACHES

Real-world speech signals have temporal scale variations due to different word lengths and pronunciations characteristics of people, which motivates the researchers to work with different receptive fields or scales. Multi-scale fusion (MSF) methods in the time domain process and fuse information at various time scales. In these methods, input from the bottom stage is processed with more processing stages in an upward direction before returning to the bottom stage. The successive down sampling and resampling of multi-resolution features (SuDoRM-RF) [132], FurcaNeXt [133], sandglasset [134], and asynchronous fully recurrent convolutional neural network (A-FCRNN) [135] are time domain MSF-based single channel speaker separation methods. SuDoRM-RF has an encoder, decoder, and separation architecture. The encoder and decoder have a 1D Conv layer and a transpose 1D Conv layer, respectively, to work opposite each other. The separation module consists of U Conv blocks to work at multiple scales of the speech signal and to calculate the mask function. U Conv block [132] is similar to U-Net and uses successive down-sampling and up-sampling operations to extract information from multiple resolutions [132].

FurcaNeXt [133] introduces variant of TCN with multiple branches for multiscale feature dynamics. For different temporal receptive field scales, these multiple branches in the network characterize different speech speeds [133].

The Sandglasset [134] has a sandglass-like shape and processes the speech signal at the multi-granularity level. For half-block of the network features, granularity becomes coarser gradually and then becomes finer successively towards the raw signal level. It uses RNN for local modeling and SAN for global modeling.

The A-FRCNN [135] introduces recurrent connections in convolution neural networks and updates the network weights asynchronously. SuDoRM-RF and Sandglasset leave the lateral information between stages unprocessed. The A-FRCNN processes information bottom-up, top-down, and also in lateral directions. It is similar to U-Net with delay [135]. In A-FRCNN, input is first passed in the bottom-up direction through stages, then parallelly fuses between adjacent stages, and finally fuses through bottom stages with skip connections. Information moves upward and becomes coarser in each stage because convolutional layers have different scales [135].

Multi-scale group transfer TasNet (MSGT TasNet) [136] applies self-attention to the small groups of the sequence instead of the whole sequence at a time [136]. This group's self-attention reduces the complexity of the model. In self-attention, any two positions are correlated for a given input sequence. Hence, for longer, input complexity increases quadratically. However, group self-attention correlates with local regions of fixed-length sequences or groups; hence with longer sequences number of groups increases and complexity increases with the increased number of groups [136]. Group self-attention does not perform cross-group correlation and loses global context information. MSGT TasNet uses multi-scale fusion to capture global information. It uses group self-attention on high-resolution scales for local context modeling and low-resolution scales for global context modeling [136].

5) TIME DOMAIN TECHNIQUES WITHOUT ENCODER DECODER FRAMEWORK

Wavesplit [65] is a time-domain speaker separation approach without an encoder-decoder framework. It uses a residual convolutional network consisting of the speaker and separation stack [65]. Fig. 6 represents the block diagram of the Wavesplit approach [65].

The speaker stack is the first stack and uses clustering to create a set of vectors for speaker representation from the mixture signal. These speaker representation vectors are in the time domain and independent of frequency bins. Then K-means clustering on speaker representation vectors results in speaker centroid. The separation stack uses the speaker's centroid and mixture signals as input to separate the speakers. The permutation problem is solved during training with the help of PIT. This way, the speaker and separation stacks are trained simultaneously [65]. At the training stage, the speaker stack creates the vector representation for every speaker and makes similar speaker distance small and different speaker distances large. The separation stack also learns to separate the clean speaker signal using these representations. At the testing stage, the speaker stack identifies the centroid for every speaker representation.

The Wavesplit uses two training objectives, i.e., (1) speaker vector objective and (2) reconstruction objective. The Speaker vector objective learns the vector representation to obtain small intra-speaker and large inter-speaker distances. The Reconstruction objective optimizes the separated speech quality. TABLE 5 illustrates the advantages and disadvantages of state-of-the-art time-domain audio source separation approaches.

C. HYBRID SPEECH SOURCE SEPARATION TECHNIQUE

The hybrid approaches work in both T-F and time domains. GCD-TasNet [137] is a hybrid domain approach. It creates an input feature map using the 1D convolutional layer in the time domain and the STFT spectrogram in the frequency domain. Then concatenated features of both domains are processed by embedding network and clustering approach to calculate the mask function [137]. The embedding network is similar to TCN and enhances the dimension of the input; then, clustering is applied to embedding to calculate the mask function. The decoder consists of transposed 1D convolutional block and ISTFT separating and adding both the results to separate the speech signals [137].

E2EPF with deep attention fusion features [112] in which the speech signal is pre-processed to separate the mixture signal in the T-F domain, then the separated signal is processed in the time domain to improve the separation outcomes. The block diagram of E2EPF is shown in Fig. 7(a) [112]. The E2EPF algorithm consists of an attention mechanism to extract deep attention fusion features [112] of speech signals and post-filter for single-channel speech source separation.

Time domain preprocessed speech signals use input features, both magnitude and phase information, to separate speech sources. The uPIT+DEF+DL [114] is the pre-processing stage and separates the mixture signal primarily in the T-F domain. However, there is residual interference in the separated speech in the pre-processing stage. The E2EPF after the pre-separation stage improves separation performance by reducing residual interferences. An attention module in the fully convolutional E2EPF network uses the feature of a mixed signal and pre-separated signal to calculate the similarity. It reduces the residual interferences from the pre-processed signal. Further, E2EPF solves the magnitude and phase mismatch problem by separating speech signals in the time domain. It has mechanisms for feature extraction [112], attention module [112], and post-filter [112]. In feature extraction, features of the mixed speech signal $Y(m)$ and the pre-processed signals $O_s(m), s = 1, 2, \ldots, S$ where S is the total number of extracted sources.The 1D convolution operation [112] extracts deep features $W_y(m)$ and $W_s(m)$ from the $Y(m)$ and $O_s(m)$ respectively given as:

$$W_y(m) = ReLU(Y(m)U_y(m)) \tag{13}$$

$$W_s(m) = ReLU(O_s(m), U_s(m)) \tag{14}$$

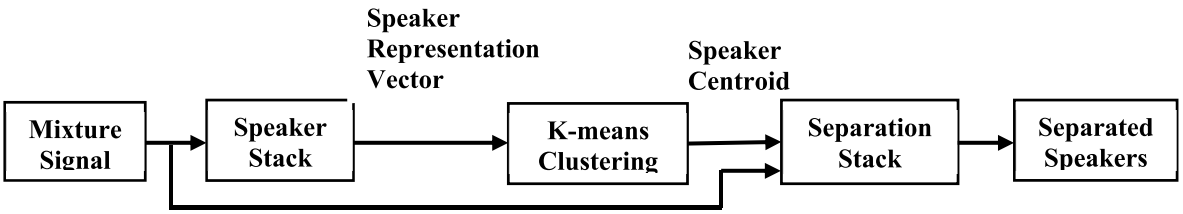

FIGURE 6. Block diagram of Wavesplit [65].

$U_y(m)$ and $U_s(m)$ represent the basis functions of 1D convolution operation [112]. The rectified linear function ReLU(∗) is a nonlinear and optional activation function. Now a day's, attention models can be used successfully to solve sequence-to-sequence learning problems [138], [139], [140], [141], [142]. The attention mechanism works on extracted features and pays more attention to reducing the interferences and improving separation performance. E2EPF applies $W_y(m)$ and $W_s(m)$ to the second 1D convolutional layer and compares the mixed and previously separated speech.

$$W'_y(m) = ReLU(W_y(m)\, U'_y(m)) \tag{15}$$

$$W'_s(m) = ReLU\left(W_s(m)\, U'_s(m)\right) \tag{16}$$

Here $W'_y(m)$ and $W'_s(m)$ are functions representing mixture and separated sources, respectively, and $U'_y(m)$, and $U'_s(m)$ represent the basis functions of second 1D convolutional operation. The correlation $g_{m,m'}(m)$ between $W'_y(m)$ and $W'_s(m)$ can be used to calculate soft mask as attention weight $h_{m,m'}(m)$ by using the global attention mechanism [141] as follows:

$$h_{m,m'} = \frac{exp(g_{m,m'}(m))}{\sum_m{}' exp(g_{m,m'}(m))} \tag{17}$$

$$g_{m,m'}(m) = W_y^{'T}(m)\, W'_o(m) \tag{18}$$

The weighted average of $W'_s(m)$ computes the context function $Co_{m',s}(m)$ as follows:

$$Co_{m',s}(m) = \sum_m h_{m,m'} W'_s(m) \tag{19}$$

The context vectors $Co_{m',s}(m)$ and deep features $W'_y(m)$ of the mixture are applied to the post-filter as attention fusion features. E2EPF in Fig. 7(b) consists of the TCN similar to TasNet [60] and represents a better performance than RNNs in various sequence modeling tasks [60], [115], [143], [144], [145]. The fully convolutional post-filter consists of stacked dilated blocks of 1D convolutional layer (Conv block) with increasing dilation factors $(1, 2, \ldots, 2^{Z-1}$, where Z represents the convolutional block number) for each TCN to capture a large temporal context which can enhance with further repeating the Z (4 times) stacked dilated convolutional blocks [106]. Fig. 7(c) represents the construction of the Conv block [146]. Skip connections maintain input information between the present and successive blocks. The depth-wise separable convolution is generally used for image processing tasks [147] [148], to reduce the number of parameters of models. A nonlinear activation function parametric rectified linear unit (PReLU) [149] improves model fitting with little overfitting risk and almost zero extra computational cost, and the global layer normalization (gLN) [60] is connected after the first 1Dconv and depth-wise 1DConv blocks.

The 1D convolutional layer, followed by ReLU nonlinear function denoted as $\mathcal{F}(*)$ takes the output of the stacked dilated 1D convolutional block. ReLU learns target masks similar to the T-F domain [106]. The predicted mask $Ma_s(m)$ of each source is the output of $F(*)$.

$$Ma_s(m) = F(\left[W_s(m), Co_{m's}(m); W'_y(m)\right]) \tag{20}$$

$$Es_s(m) = W_y(m) \otimes Ma_s(m) \tag{21}$$

$Es_s(m)$ is the estimated separated signal for the target source. The 1D convolutional operator with $U_E(m)$ as basis function reconstructs the predicted signal as follows:

$$\hat{X}_o(m) = Es_o(m) U_E(m) \tag{22}$$

$\hat{X}_o(m)$ is predicted output signal. Hybrid domain method illustrate how to combine T-F and time domain approaches.

VI. PERFORMANCE ANALYSIS OF MONAURAL SPEECH SOURCE SEPARATION TECHNIQUES

Time-domain models present inferior outcomes of all existing deep learning-based monaural speech source separation approaches. Phase magnitude mismatch, time-frequency decomposition, and large window size are disadvantages of T-F domain approaches. Time-domain approaches overcome these difficulties by working in the time domain using the encoder-decoder framework. In this domain, researchers are working to reduce the model's size and increase separation performance with local and global context modeling. Deep learning-based single-channel T-F domain speech source separation approaches use STFT transformation, designed for any type of signal but not specifically for speech signals and may cause suboptimal performance. These approaches process only magnitude spectrograms and have magnitude and phase decoupling problems. Initially, DNN and RNN-based approaches have attempted to design speaker separation models using deep learning. However, these methods are not designed according to speech characteristics

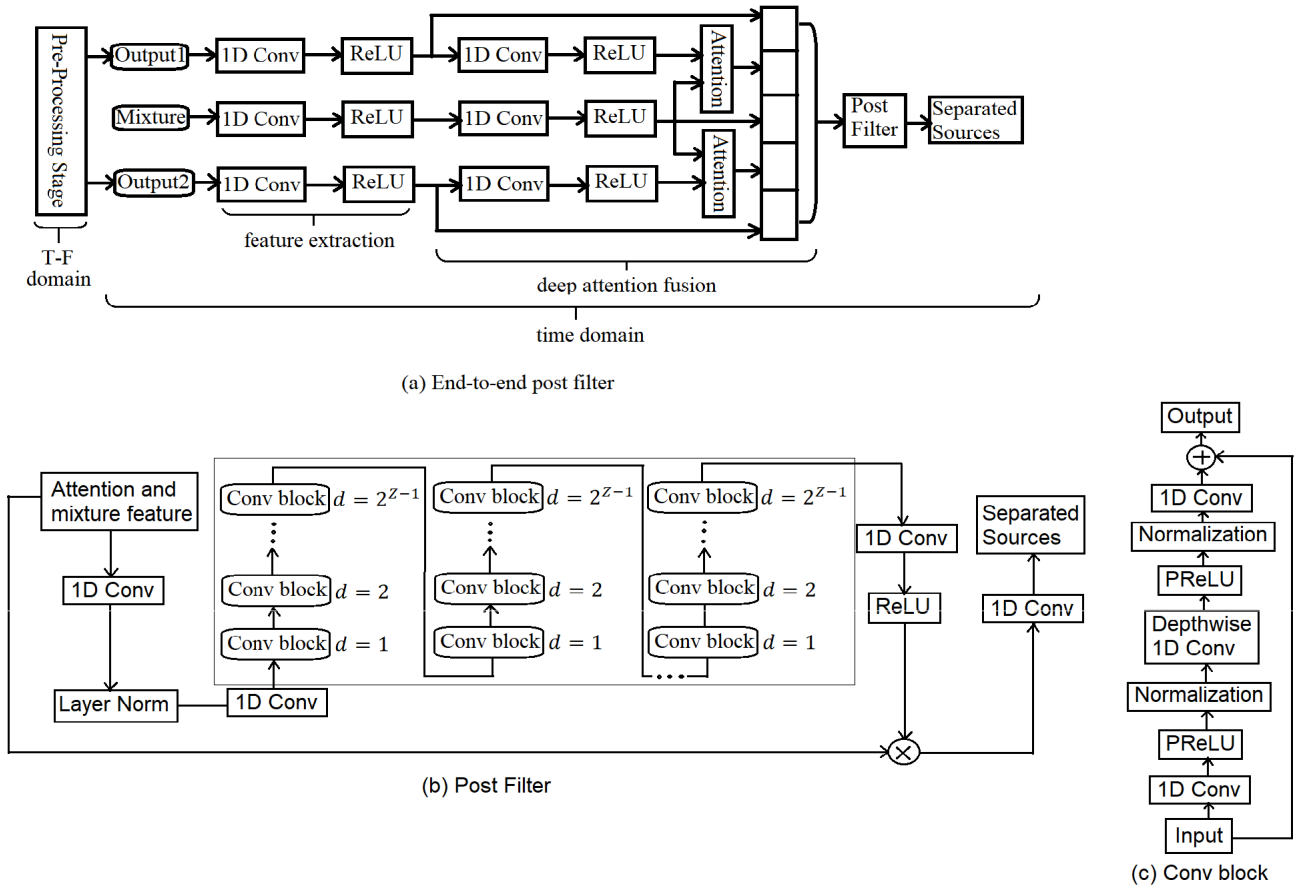

FIGURE 7. The block diagram of E2EPF, (a) E2EPF, (b) Post filter flowchart, (c) Conv block design [112].

and fail to achieve better performance. Clustering-based approaches from DC to ADANet are good initiatives for deep learning-based speech separation and solving permutation and output dimension mismatch problems. But clustering requires the number of speakers should be known previously; hence not useful for real-time applications of speech source separation. Human hearing attention inspires the ADCNet to use multi-head self-attention with density-based canopy k-means clustering. This clustering method self-identifies the number of speakers in the mixture and can work with an unknown number of speakers. ESDC method enhances the discriminative ability of the separation model using adjacency-based similarity and squash-norm normalization of high dimensional embedding vectors. ADCNet and ESDC are state-of-the-art methods in the T-F domain but are not as productive as time domain methods.

PIT-based approaches choose the best permutation to solve the permutation problem and have an output dimension mismatch problem. PIT separates the target speaker from the mixture signal in the T-F domain. But separation with PIT may contain interference because it only reduces the same speaker distance and leaves the distance between different speakers unchanged. Deep clustering is used after PIT to increase the distance between different speakers. Multi-task learning performs multiple tasks simultaneously to improve the training of the model. Monaural speech separation with deep CASA uses tPIT to separate the speakers from the mixture signal in the simultaneous grouping stage and k-means clustering to track the speaker in sequential grouping to provide good separation performance. Here de-noising before the simultaneous grouping reduces further complexity, simultaneously improving performance. Deep CASA is the best T-F domain method but needs speaker tracking due to tPIT. Many recent monaural speaker separation methods use DC and PIT simultaneously to increase separation accuracy. The uPIT+DEF+DL uses uPIT and DC jointly to separate different speech signals, followed by discriminative learning to increase the distance between separated sources and fine-tune the separated speech signal. The T-F domain phase reconstruction approaches try to solve the phase magnitude decoupling problem, but the results are not comparable with time domain approaches. Time-domain approaches use data-driven representation instead of STFT features. These methods have not been analyzed with large data sets because scaling and generalization of large data are impossible with these approaches. Time domain approaches focus on local

TABLE 5. Advantages and Disadvantages of Time Domain Speech Source Separation Approaches.

S. No.	Models	Advantages	Disadvantages
1.	TasNet [59]	• First time-domain audio source separation method. • No upper bound performance due to noisy phase spectrogram. • Outperform over previous T-F domain methods both in causal and non-causal case. • Applicable to real world low latency applications as compared to T-F domain methods.	• Deep LSTM network in separation module limits its applicability. • For smaller length of input segments encoder produces larger output length results unmanageable deep LSTM training. • LSTM network has long temporal dependencies which restricts the separation performance to be consistent. • Sensitive to shift in the input starting point. Shift adversely affect the separation performance.
2.	DPRNN [124]	• Reduced model size by replacing 1D CNN with DPRNN. • Perform separation at sample level. • Perform both local and global modeling with inter and intra chunks respectively. • Performs utterance level processing. • One of the good sequential modeling modules. • System needing long term sequence modeling can use DPRNN.	• Suboptimal performance due to so many intermediate stages to process the information at context level. • RNNs are sequential in nature and can degrades computation parallelization for processing long sequences from large datasets. • Strong global modeling but poor local modeling. • Gradient explosion.
3.	Gated DPRNN [83]	• Uses gated RNN to separate the more than three number of speakers. • Complexity increases linearly rather than quadratically with increased number of speakers as compared to other approaches.	• Good work for unknown number of speakers but still not speaker number independent method.
4.	ConvTasNet [60]	• Consecutive frames can be processed parallelly with convolution to speed up the separation processes. • Depth-wise separation convolution operation reduces number of parameters and computation cost. • Variation in the mixture starting point does not affect consistence performance across entire test set. • Suitable for both real-time and offline, low latency speech processing applications.	• Perform only local modeling. • Due to fixed temporal context length fails to perform long term tracking of individual speakers especially long pause of mixture audio. • Receptive field of 1D CNN is fixed and smaller than the length of the sequence hence sequence level dependency cannot be utilized. • Strong local modeling but poor global modeling.
5.	SANet [126]	• It is an advance version of DANet in time domain. • Number of speakers during training and test phase can be different.	• Can perform only local context modeling.
6.	NasTasNet [127]	• Searches best model architecture using NAS technique. • Minimizes human interaction.	• Increases parameters of the model due to implementation of NAS technique.
7.	CasNet [128]	• Increases channel robustness of the model. • Models become aware about channel information.	• Channel encoder increases size of the model.
8.	DPTNet [125]	• Introduces context aware modeling for direct interaction between speech sequences. • Can model extremally long input sequences. • Uses RNN to learn order information without positional encoding.	• It also includes RNN which degrades parallelization capabilities. • Strong global modeling but poor local modeling. • Worked at fixed context size.
9.	GALR [129]	• DPTNet is globally recurrent and locally attentive and GALR is globally attentive and locally recurrent.	• Changing the attention and recurrent operation order degrades the performance.
10.	DPHA [131]	• It is transformer-based method designed using various attention mechanism.	• Limited performance as compared to other methods.
11.	SuDoRM-RF [132]	• Uses U-Conv block to process successive up and down sampling of information from multiple time steps.	• Correlates only up and down information. • Leave the adjacent information uncorrelated.
12.	FurcaNeXt [133]	• Design variants of TCN to work on multiple time steps.	• Consists of 51.4 Million parameters. • Not reliable for real-world applications due to big model size.
13.	Sandglasset [134]	• First work that can model multi-granularity segments using self-attention network. • Contextual modeling and computational efficiency can be increased with multi-granularity features.	• For same granularity features, residual connections cannot preserve information after bottleneck. • Degraded parallelization capability due to use of RNN.

TABLE 5. *(Continued.)* **Advantages and Disadvantages of Time Domain Speech Source Separation Approaches.**

		• Multi-granularity work with local dependencies like spectral structure, temporal or spectral continuity and timbre of speech sequences.	
14.	A-FCRNN [135]	• Works on up, down, and lateral information using U-Net with delay for improving separation performance.	• Process local information effectively and leaves the global information hence performance is not as expected.
15.	Sepformer [130]	• RNN-free transformer-based architecture. • Consists of multi-head attention and feed forward layers. • Learns both short and long-term dependencies. • Instead of RNN it uses multiscale pipeline composed of transformer.	• Consists of 26 million parameters. • Cannot focus on local dependencies like spectral structure, temporal or spectral continuity and timbre of speech sequences. • Leaves the low performance samples during training. • Worked at fixed context size.
16.	Wavesplit [65]	• First time domain work without an encoder-decoder framework. • Uses clustering and PIT in time domain to perform separation task. • Perform well with long sequences.	• Clustering based method hence not worked well with online applications. • Not perform local modeling. • Output dimension mismatch problem.
17.	TAANet [11]	• Includes local attention network and global attention network. • Channel attention mechanism and special attention mechanism perform local attention and for global attention self-attention mechanism is used. • Both networks are connected serially to focus on local and global information at the frame level.	• Since the self-attention network has a quadratic cost function which can impact long-term processing. • Worked with fixed context size while speech has a different context length.

TABLE 6. **Comparison of *SDR* (in dB) of Monaural Speech Source Separation Techniques on WSJ0-2mix Dataset.**

Processing Domains	Methods	Model Size (in Million)	*SDR* (in dB)
Time-Frequency Domain	DC [66]	13.6	6.7
	uPIT-BLSTM [80]	92.7	10.0
	DANet [30]	9.1	10.8
	ADANet [102]	9.1	10.2
	ADCNet [103]	-	10.9
	Cu-PIT-Grid-RD [104]	47.2	11.0
	Chimera++ [104]	32.9	13.1
	CBLDNN-GAT [113]	39.5	12.0
	WAMISI [122]	39.9	13.6
	SDC-MLT-Grid [114]	-	10.7
	CASA-E2E [115]	-	11.2
	Listen and Group [120]	8.2	11.0
	Sign Prediction Net [121]	-	15.8
	OR-PIT [105]	-	15.0
	Two speaker deep CASA [116]	12.8	15.5
	IRM [60]	-	12.6
	IBM [116]	-	13.5
	IPSM [112]	-	15.3
	WFM [60]	-	13.8
Time-Domain	BLSTM- TasNet [123]	23.6	13.2
	TasNet [59]	8.8	15.2
	GCD-TasNet [137]	10	16.9
	ConvTasNet$_{gLN}$ [60]	5.1	15.8
	DPRNN [124]	2.6	19.0
	DPTNet [125]	2.6	20.2
	Sepformer [130]	26	22.3
	Wavesplit [65]	29	21.2
	TAANet [11]	5.4	20.9
Hybrid Domain	uPIT+DEF+DL+E2EPF [112]	-	17.0

and global context modeling of the speech signal. RNN-based TasNet is the first-time domain approach designed with a deep LSTM network. DPRNN is the time domain speech separation method that performs global modeling effectively

TABLE 7. Comparison of *SDR* (in dB) of Monaural Speech Source Separation Techniques on WSJ0-3mix Dataset.

Processing Domains	Methods	Model Size (in Million)	*SDR* (in dB)
Time-Frequency Domain	uPIT-BLSTM-SI [81]	92.7	7.7
	DANet[30]	9.1	8.9
	ADANet [102]	9.1	9.4
	ADCNet [103]	-	9.56
	uPIT [81]	92.7	7.7
	Listen and Group [120]	56.6	12.5
	Multispeaker Deep CASA [116]	12.8	14.8
	OR-PIT [105]	-	12.9
	uPIT+DEF+DL [112]	-	8.0
Time-Domain	$\text{ConvTasNet}_{\text{gLN}}$ [60]	5.1	13.1
	$\text{ConvTasNet}_{\text{cLN}}$ [60]	5.1	8.2
	Wavesplit [65]	29	17.6
Hybrid Domain	uPIT+DEF+DL+E2EPF+Attension [112]	-	13.0

TABLE 8. Comparison of *SI − SNR* (in dB) of Monaural Speech Source Separation Techniques on WSJ0-2mix Dataset.

Processing Domains	Methods	Model Size (in Million)	$SI-SNR$ (in dB)
Time- Frequency Domain	DC+ [29]	13.6	10.8
	DANet [30]	9.1	10.5
	ADANet [102]	9.1	10.4
	ADCNet [103]	-	10.75
	CBLDNN [113]	39.9	11.0
	Chimera++ [104]	32.9	11.5
	WAMISI-5 [122]	39.9	12.6
	Sign Prediction Net [121]	36.8	15.3
	OR-PIT [105]	-	14..8
	Two speaker deep CASA [116]	12.8	15.2
	IRM [60]	-	12.2
	IBM [116]	-	13.0
	IPSM [112]	-	14.4
Time-Domain	BLSTM- TasNet [123]	23.6	15.3
	TasNet [59]	-	14.6
	$\text{ConvTasNet}_{\text{gLN}}$ [60]	5.1	15.3
	GCD-TasNet [137]	10.0	16.6
	SuDoRM-RF [132]	2.6	18.9
	FurcaNeXt [133]	51.4	18.4
	Gated DPRNN [86]	7.5	20.1
	DPRNN [124]	2.6	18.8
	DPTNet [125]	2.6	20.6
	Sepformer [130]	26	22.6
	TAANet [11]	5.4	20.7
Hybrid Domain	uPIT+DEF+DL+E2EPF+Attension [112]	-	16.6

with dual-path architecture. Existing time domain approaches separate two and three speakers from the mixture. Gated DPRNN can separate more than three speakers but still is not a number of speaker-independent methods.

TABLE 9. Comparison of $SI-SNR$ (in dB) of Monaural Speech Source Separation Techniques on WSJ0-3mix Dataset.

Processing Domains	Methods	Model Size (in Million)	$SI-SNR$ (in dB)
Time- Frequency Domain	DC+ [29]	13.6	7.1
	DANet [30]	9.1	8.6
	ADANet [102]	9.1	9.1
	Listen and Group [120]	56.6	12.1
	Multispeaker Deep CASA [116]	12.8	14.5
	OR-PIT [105]	-	12.6
	uPIT+DEF+DL [112]	-	7.2
	IBM [60]	-	12.5
	IRM [116]	-	12.5
Time-Domain	ConvTasNetgLN [60]	5.1	12.7
	DPRNN [124]	2.6	14.7
	Gated DPRNN [83]	7.5	16.9
Hybrid Domain	uPIT+DEF+DL+E2EPF+Attension [112]	-	12.5

TABLE 10. Comparison of PESQ values of Monaural Speech Source Separation Techniques on WSJ0-2mix Dataset.

Processing Domains	Methods	PESQ
Time-Frequency domain	Sign Prediction Net [121]	3.43
	OR-PIT [105]	3.12
	uPIT-BLSTM [123]	2.84
	ADANet [102]	2.82
	ADCNet [103]	2.89
	Two speaker deep CASA [116]	3.25
	IRM [60]	3.74
	IBM [116]	3.33
	IPSM [112]	3.64
Time-Domain	TasNet [59]	3.25
	ConvTasNet$_{gLN}$ [60]	3.24
	DPRNN [124]	3.49
	DPHA-Net [131]	3.70
Hybrid Domain	uPIT+DEF+DL+E2EPF [112]	3.41

CNN-based approach ConvTasNet is the best model for local context modeling, but due to CNN cannot perform global modeling efficiently. The SANet is implemented in the time domain as an improved T-F domain DANet method. It is categorized as CNN-based because it creates an embedding vector and produces attractors using TCN, similar to ConvTasNet. This method uses the concept of the T-F domain in the time domain. NasTasNet searches for the best model for ConvTasNet using the NAS technique. It can be used with speaker separation models to obtain the best model architecture with minimum human interaction. The Casnet enhances the channel robustness of TasNet models by making them aware of channel information.

Transformer-based approach DPTNet outperforms global modeling of speech source separation with dual-path design. DPTNet is locally attentive and globally recurrent. GALR is globally attentive and locally recurrent. But interchanging the attention and recurrent mechanism in GALR degrades the separation performance compared to DPTNet. The Sepformer is an advancement of the transformer-based approach

TABLE 11. Comparison of *SI − SNR* (in dB) of Monaural Speech Source Separation Techniques on WSJ0-2mix Dataset.

Processing Domains	Methods	PESQ
Time-Frequency Domain	Listen and Group [120]	2.77
	Multispeaker Deep CASA [116]	2.83
	ADCNet [103]	2.25
	OR-PIT [105]	2.60
	uPIT+DEF+DL [112]	2.03
Time-Domain	$\text{ConvTasNet}_{\text{gLN}}$ [60]	2.61
Hybrid Domain	uPIT+DEF+DL+E2EPF+ Attension [112]	2.70

TABLE 12. Comparison of *STOI* (in %) of Monaural Speech Source Separation Techniques on Libri-2mix DATASET.

Processing Domain	Methods	*STOI*
Time-Frequency	Deep CASA [116]	93.20
	IBM [116], [65]	88.44
	IRM [60], [65]	93.33
	IPSM [112]	95.55
Time	DPRNN [124]	92.64
	Gated DPRNN [83]	92.52
	DPHA-Net [131]	98.14
	Gated RNN [83]	92.52

with a built-in attention mechanism. The Sepformer outperforms all the approaches in terms of SDR and SI-SNR with the WSJ0-2mix dataset. The TAANet is also a recent approach and incorporates CNN and attention in its architecture to perform local and global modeling to achieve impressive results.

The DPHA aggregates the output of present and previous stages to calculate present stage output. It extracts the multi-head self-attention features using the EA unit and fuses them using the AFF unit to enhance the feature extraction capability of the model.

Multi-scale fusion-based methods work at different scales and characteristics of speech signals. The SuDoRM-RF uses U Conv block for successive up and down-sampling operations to extract information from multiple time steps. FurcaNeXt proposes a variety of TCNs to work on multiple branches for different speech characteristics. The Sandglasset has a sandglass-like structure and is the only method that works on the multi-granularity level of the speech signal. The A-FCRNN works on lateral information compared to SuDoRm-Rf and Sandglasset using U-Net with delay to improve separation performance. MSGT TasNet creates small groups from input vectors and calculates the correlation between the group elements using multi-head self-attention. This method calculates the cross-group correlation using MSF to reduce the complexity of the model. The information within the group is local context information, and the cross-group is global context information. The Wavesplit is a recent time-domain speech source separation approach that uses concepts of clustering and permutation invariant training in the time domain. It is a multi-speaker separation algorithm with limited performance improvement for more than three speakers and produces separation results on different datasets. It is one of the most efficient speech separation algorithms.

GCD-TasNet is a hybrid domain approach. The GCD-TasNet encodes STFT spectrograms and time domain features from raw input and concatenates the information as the encoder's output separation module consists of an embedding network similar to TCN and clustering operation to separate the speakers from these combined features. E2EPF filter for speaker separation with deep attention features is the state-of-the-art hybrid domain algorithm for deep learning-based speaker separation. The T-F domain uPIT+DEF+DL preliminarily separates the target speech signal from the mixture speech signal, then E2EPF pays more attention using the attention module, and post-filter in the time domain

TABLE 13. Comparison of *SI − SNR* (in dB) of Monaural Speech Source Separation Techniques on Different Dataset.

Model	Datasets					
	WSJ0-2mix	WSJ0-3mix	WHAM!	WHAMR!	Libri2mix	Libri3mix
DC [29]	10.6	7.1	-	-	5.9	-
DANet [30]	10.5	8.6	-	-	8.4	-
uPIT-BLSTM [123]	9.8	-	9.8	-	6.3	-
Chimera++ [104]	11.5	-	10.0	-	12.0	-
DPRNN [124]	18.8	14.7	13.85	10.28	14.1	-
GALR [129]	-	-	-	-	12.2	-
SuDoRM-RF [132]	18.9	-	12.9	-	13.5	-
ConvTasNet [60], [84]	15.3	12.7	12.7	8.3	14.7	12.1
Gated DRNN [65]	20.1	16.9	15.2	12.2	-	-
IRM [60], [65]	-	-	-	-	12.9	13.1
IBM [116], [65]	-	-	-	-	13.7	13.9
Wavesplit [65]	21.0	17.3	15.4	12.0	19.5	15.8
SANet [126]		-	-	-	12.8	-
A-FRCNN [135]	18.3	-	14.5	-	16.7	-
DPTNet [125]	-	-	-	-	14.9	-
DPHA-Net [125]	-	-	14.7	-	16.5	-
Gated DPRNN [83]	20.7	16.9	15.17	12.21	15.2	-

reduces the interference in the pre-separated speech signal. The attention module and post-filter are very good proposals for enhancing the pre-separated speaker signal.

The comparison of results of various single-channel speech source separation deep learning-based approaches in terms of *SDR*, *SI − SNR*, PESQ, and *STOI* on WSJ0-2mix have been illustrated in TABLE 6, 8, 10, and 12, respectively. Similarly, TABLE 7, 9, and 11 compare *SDR*, *SI − SNR*, and PESQ on WSJ0-3mix, respectively. TABLE 13 compares *SI − SNR* values of different deep learning models on different datasets. Comparison tables show that time domain models have a much-reduced size than the T-F domain model with enhanced performance. These methods overcome all the drawbacks of T-F domain approaches like output dimension mismatch, permutation ambiguity, and large context window size with the encoder-decoder framework.

VII. CONCLUSION AND FUTURE SCOPE

This paper comprehensively studies and analyzes deep learning models for monaural speech source separation. Different models are categorized into the T-F, time, and hybrid domains. The methods have been described in brief and some in detail to build the basic concepts of deep learning-based speech source separation work. The comparative analysis of different deep learning-based speech source separation models in terms of SDR, SI-SNR, and PESQ has been provided for the readers to understand the domain better. It is observed that T-F domain methods have several constraints to obtain the required frequency resolution, which time domain methods can overcome. Although numerous approaches are designed for two or three-speaker separation at a particular language dataset, real-world language independent and a number of speaker-independent speaker separations is still a challenging problem. The time domain approaches are still in the primitive stage. Some recent time-domain approaches have been designed to work on more than three speaker separations and some on languages other than English. The future prospects lie in the design of real-world deep learning models for all practical applications. Dataset creation and separation model designing for more than three speakers and multiple languages, improved separation module design for the encoder-decoder framework, improved multi-scale fusion model design that covers all scales of speech signals, designing the attention mechanism analogous to human auditory attention, and implementing the deep learning model using the concepts of T-F, time, and hybrid domain approaches are some areas for future research.

REFERENCES

[1] C. Gao, G. Cheng, T. Li, P. Zhang, and Y. Yan, "Self-supervised pre-training for attention-based encoder–decoder ASR model," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1763–1774, 2022, doi: 10.1109/TASLP.2022.3171967.

[2] S. Chandrakala, S. Malini, and S. V. Veni, "Histogram of states based assistive system for speech impairment due to neurological disorders," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2425–2434, 2021, doi: 10.1109/TNSRE.2021.3125314.

[3] Q. M. Areeb, M. Nadeem, R. Alroobaea, and F. Anwer, "Helping hearing-impaired in emergency situations: A deep learning-based approach," *IEEE Access*, vol. 10, pp. 8502–8517, 2022, doi: 10.1109/ACCESS.2022.3142918.

[4] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio Speech Language Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[5] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, Apr. 2018.

[6] M. Yu, A. Rhuma, S. M. Naqvi, L. Wang, and J. Chambers, "A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 6, pp. 1274–1286, Nov. 2012.

[7] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 125–134, May 2014.

[8] M. S. Khan, S. M. Naqvi, A. Ur-Rehman, W. Wang, and J. Chambers, "Video-aided model-based source separation in real reverberant rooms," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1900–1912, Sep. 2013.

[9] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 5, pp. 895–910, Oct. 2010.

[10] C. Li, Z. Chen, and Y. Qian, "Dual-path modeling with memory embedding model for continuous speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1508–1520, 2022, doi: 10.1109/TASLP.2022.3165712.

[11] J. Cai, K. Wang, J. Yao, and H. Zhou, "Time-domain adaptive attention network for single-channel speech separation," *SSRN*, 2022, doi: 10.2139/SSRN.4084951.

[12] H. M. Tan, K.-W. Liang, Y.-S. Lee, C.-T. Li, Y.-H. Li, and J.-C. Wang, "Speech separation using augmented-discrimination learning on squash-norm embedding vector and node encoder," *IEEE Access*, vol. 10, pp. 102048–102063, 2022, doi: 10.1109/ACCESS.2022.3188712.

[13] H. D. Do, S. T. Tran, and D. T. Chau, "Speech source separation using variational autoencoder and bandpass filter," *IEEE Access*, vol. 8, pp. 156219–156231, 2020, doi: 10.1109/ACCESS.2020.3019495.

[14] P. Hoang, Z.-H. Tan, J. M. De Haan, and J. Jensen, "The minimum overlap-gap algorithm for speech enhancement," *IEEE Access*, vol. 10, pp. 14698–14716, 2022, doi: 10.1109/ACCESS.2022.3147514.

[15] M. Hosseini, L. Celotti, and E. Plourde, "End-to-end brain-driven speech enhancement in multi-talker conditions," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1718–1733, 2022, doi: 10.1109/TASLP.2022.3169629.

[16] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, N. Kamo, and S. Araki, "Switching independent vector analysis and its extension to blind and spatially guided convolutional beamforming algorithms," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1032–1047, 2022, doi: 10.1109/TASLP.2022.3155271.

[17] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Languages Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[18] S. S. Ramya, "DNN based speech quality enhancement and multi-speaker separation for automatic speech recognition system," in *Machine Learning Algorithms for Signal and Image Processing*. Chennai, India: IEEE Press, 2023, pp. 231–246, doi: 10.1002/9781119861850.ch13.

[19] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[20] R. Jin, M. Ablimit, and A. Hamdulla, "Speech separation and emotion recognition for multi-speaker scenarios," in *Proc. 3rd Int. Conf. Pattern Recognit. Mach. Learn. (PRML)*, Jul. 2022, pp. 280–284, doi: 10.1109/PRML56267.2022.9882231.

[21] J. Zeremdini, M. A. Ben Messaoud, and A. Bouzid, "A comparison of several computational auditory scene analysis (CASA) techniques for monaural speech segregation," *Brain Informat.*, vol. 2, no. 3, pp. 155–166, Sep. 2015, doi: 10.1007/s40708-015-0016-0.

[22] T. Pham, Y. S. Lee, Y. A. Chen, and J.-C. Wang, "A review on speech separation using NMF and its extensions," in *Proc. Int. Conf. Orange Technol. (ICOT)*, 2016, pp. 26–29.

[23] K. S. Ananthakrishnan and K. Dogancay, "Recent trends and challenges in speech-separation systems research—A tutorial review," in *Proc. IEEE Region 10 Conf.*, Singapore, Nov. 2009, pp. 1–6.

[24] R. Aihara, G. Wichern, and J. Le Roux, "Deep clustering-based single-channel speech separation and recent advances," *Acoust. Sci. Technol.*, vol. 41, no. 2, pp. 465–471, 2020.

[25] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends Amplification*, vol. 12, no. 4, pp. 332–353, Dec. 2008.

[26] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.

[27] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization with sliding windows and spectral masks," in *Proc. Interspeech*, Aug. 2011, pp. 1773–1776.

[28] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2001, pp. 556–562.

[29] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 31–35, doi: 10.1109/ICASSP.2016.7471631.

[30] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 246–250, doi: 10.1109/ICASSP.2017.7952155.

[31] D. Yu, M. Kolbaek, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 241–245.

[32] Y.-M. Qian, C. Weng, X.-K. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 40–63, Jan. 2018.

[33] K. B. Bhangale and M. Kothandaraman, "Survey of deep learning paradigms for speech processing," *Wireless Pers. Commun.*, vol. 125, no. 2, pp. 1913–1949, Jul. 2022.

[34] A. Singh and T. Ogunfunmi, "An overview of variational autoencoders for source separation, finance, and bio-signal applications," *Entropy*, vol. 24, no. 1, p. 55, Dec. 2021.

[35] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiat, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "Subspace Gaussian mixture models for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4330–4333, doi: 10.1109/ICASSP.2010.5495662.

[36] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2004, pp. 1–4.

[37] A. M. Reddy and B. Raj, "A minimum mean squared error estimator for single channel speaker separation," in *Proc. Interspeech*, Oct. 2004, pp. 2445–2448.

[38] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1766–1776, Aug. 2007.

[39] B. H. Juang and L. R. Rabiner, "Hidden Markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, Aug. 1991.

[40] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Found. Trends Signal Process.*, vol. 1, no. 3, pp. 195–304, 2008.

[41] T. Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space," in *Proc. Interspeech*, vol. 2, Sep. 2006, pp. 89–92.

[42] A. N. Deoras and A. Hasegawa-Johnson, "A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 2004, pp. 1–4.

[43] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 45–66, 2010.

[44] A. Ozerov, C. Fevotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2009, pp. 121–124.

[45] M. H. Radfar, W. Wong, R. M. Dansereau, and W.-Y. Chan, "Scaled factorial hidden Markov models: A new technique for compensating gain differences in model-based single channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 1918–1921.

[46] M. H. Radfar, R. M. Dansereau, and W.-Y. Chan, "Monaural speech separation based on gain adapted minimum mean square error estimation," *J. Signal Process. Syst.*, vol. 61, no. 1, pp. 21–37, Oct. 2010.

[47] M. H. Radfar, W. Wong, W.-Y. Chan, and R. M. Dansereau, "Gain estimation in model-based single channel speech separation," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Sep. 2009, pp. 1–15.

[48] F. Murtagh, "Brief history of cluster analysis," in *Handbook of Cluster Analysis*. Boca Raton, FL, USA: CRC Press, 2015, pp. 21–30.

[49] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," *J. Mach. Learn. Res.*, vol. 7, pp. 1963–2001, Oct. 2006.

[50] H. Jia, S. Ding, X. Xu, and R. Nie, "The latest research progress on spectral clustering," *Neural Comput. Appl.*, vol. 24, nos. 7–8, pp. 1477–1486, 2014.

[51] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.

[52] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 625–638, May 2009.

[53] E. M. Grais and H. Erdogan, "Spectro-temporal post-enhancement using MMSE estimation in NMF based single-channel source separation," in *Proc. Interspeech*, Aug. 2013, pp. 3279–3283.

[54] E. M. Grais and H. Erdogan, "Gaussian mixture gain priors for regularized nonnegative matrix factorization in single-channel source separation," in *Proc. Interspeech*, Sep. 2012, pp. 1518–1521, doi: 10.21437/interspeech.2012-429.

[55] E. M. Grais and H. Erdogan, "Regularized nonnegative matrix factorization using Gaussian mixture priors for supervised single channel source separation," *Comput. Speech Lang.*, vol. 27, no. 3, pp. 746–762, May 2013.

[56] Y. Wang and D. L. Wang, "Cocktail party processing via structured prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1, 2012, pp. 224–232.

[57] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3734–3738.

[58] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. Hoboken, NJ, USA: Wiley, 2018.

[59] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 696–700, doi: 10.1109/ICASSP.2018.8462116.

[60] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[61] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[62] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR—Half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 626–630, doi: 10.1109/ICASSP.2019.8683855.

[63] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4214–4217.

[64] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, document ITU-T 862, ITU-T Recommendation, 2001.

[65] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2840–2849, 2021, doi: 10.1109/TASLP.2021.3099291.

[66] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, Sep. 2016, pp. 545–549.

[67] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

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

[69] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2001, pp. 749–752.

[70] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: MIT Press, 1994.

[71] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[72] B. Moore, *An Introduction to the Psychology of Hearing*. Leiden, The Netherlands: Brill, 2012.

[73] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," in *Proc. Int. Joint Conf. Neural Netw.*, vol. 4, 2001, pp. 2861–2866.

[74] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[75] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. Boston, MA, USA: Springer, 2005, pp. 181–197.

[76] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 5, pp. 967–977, May 2016.

[77] Y. Sun, Y. Xian, W. Wang, and S. M. Naqvi, "Monaural source separation in complex domain with long short-term memory neural network," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 359–369, May 2019, doi: 10.1109/JSTSP.2019.2908760.

[78] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[79] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 708–712.

[80] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[81] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. 2nd Int. Conf. Spoken Lang. Process. (ICSLP)*, New York, NY, USA, Oct. 1992, pp. 1–6.

[82] G. Wichern et al., "WHAM!: Extending speech separation to noisy environments," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019, pp. 1368–1372.

[83] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," 2020, *arXiv:2003.01531*.

[84] M. MacIejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 696–700.

[85] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon, Tech. Rep. N, 1993, vol. 93.

[86] P. Kabal, "TSP speech database," McGill Univ., Montreal, QC, Canada, Database Version 2 (2018-11), 2002.

[87] M. Li, T. Lan, C. Peng, Y. Qian, and Q. Liu, "Multi-layer attention mechanism based speech separation model," in *Proc. IEEE 19th Int. Conf. Commun. Technol. (ICCT)*, Oct. 2019, pp. 506–509.

[88] M. Cooke and T.-W. Lee. (2006). *Speech Separation Challenge*. [Online]. Available: http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm

[89] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 30–44, Jan. 2010.

[90] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[91] K. Ito. (2017). *The LJ Speech Dataset*. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[92] Q. Liu, P. J. B. Jackson, and W. Wang, "A speech synthesis approach for high quality speech separation and generation," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1872–1876, Dec. 2019.
[93] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," 2020, *arXiv:2005.11262*.
[94] S. Lutati, E. Nachmani, and L. Wolf, "SepIt: Approaching a single channel speech separation bound," 2022, *arXiv:2205.11801*.
[95] K. Li, R. Yang, and X. Hu, "An efficient encoder–decoder architecture with top-down attention for speech separation," 2022, *arXiv:2209.15200*.
[96] C. Veaux, J. Yamagishi, and K. Macdonald, "SUPERSEDED—CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," Centre Speech Technol. Res., Univ. Edinburgh, Edinburgh, U.K., 2016. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/2119
[97] M. Hermans and B. Schrauwen, "Training and analyzing deep recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 190–198.
[98] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," in *Proc. 2nd Int. Conf. Learn. Represent.*, 2013, pp. 1–15.
[99] E. Ceolini and S.-C. Liu, "Impact of low-precision deep regression networks on single-channel source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 256–260.
[100] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1486–1494, Sep. 2009.
[101] M. Delfarah and D. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 5, pp. 1085–1094, May 2017.
[102] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 4, pp. 787–796, Apr. 2018, doi: 10.1109/TASLP.2018.2795749.
[103] Y. Jin, C. Tang, Q. Liu, and Y. Wang, "Multi-head self-attention-based deep clustering for single-channel speech separation," *IEEE Access*, vol. 8, pp. 100013–100021, 2020.
[104] Z. Q. Wang, J. le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 686–690, doi: 10.1109/ICASSP.2018.8462507.
[105] N. Takahashi, S. Parthasaarathy, N. Goswami, and Y. Mitsufuji, "Recursive speech separation for unknown number of speakers," in *Proc. Interspeech*, Sep. 2019, pp. 1348–1352, doi: 10.21437/Interspeech.2019-1550.
[106] C. Fan, B. Liu, J. Tao, J. Yi, and Z. Wen, "Discriminative learning for monaural speech separation using deep embedding features," in *Proc. Interspeech*, Sep. 2019, pp. 4599–4603.
[107] C. Fan, B. Liu, J. Tao, J. Yi, and Z. Wen, "Spatial and spectral deep attention fusion for multi-channel speech separation using deep embedding features," 2020, *arXiv:2002.01626*.
[108] C. Fan, B. Liu, J. Tao, Z. Wen, J. Yi, and Y. Bai, "Utterance-level permutation invariant training with discriminative learning for single channel speech separation," in *Proc. 11th Int. Symp. Chin. Spoken Language Process. (ISCSLP)*, Nov. 2018, pp. 26–30.
[109] E. M. Grais, G. Roma, A. J. R. Simpson, and M. D. Plumbley, "Combining mask estimates for single channel audio source separation using deep neural networks," in *Proc. Interspeech*, Sep. 2016, pp. 3339–3343.
[110] P. Sen Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *Proc. 15th Int. Soc. Music Inf. Retr. Conf.*, 2014, pp. 477–482.
[111] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1562–1566.
[112] C. Fan, J. Tao, B. Liu, J. Yi, Z. Wen, and X. Liu, "End-to-end post-filter for speech separation with deep attention fusion features," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1303–1314, 2020.
[113] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, "CBLDNN-based speaker-independent speech separation via generative adversarial training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 711–715.
[114] C. Xu, W. Rao, and E. Chng, "A shifted delta coefficient objective for monaural speech separation using multi-task learning," in *Proc. Interspeech*, Mar. 2018, pp. 3479–3483, doi: 10.21437/Interspeech.2018-1150.
[115] Y. Liu and D. Wang, "A CASA approach to deep learning based speaker-independent co-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5399–5403.
[116] Y. Liu, M. Delfarah, and D. Wang, "Deep CASA for talker-independent monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6354–6358.
[117] Y. Liu and D. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 12, pp. 2092–2102, Dec. 2019.
[118] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 47–54.
[119] Y. Liu and D. Wang, "Causal deep CASA for monaural talker-independent speaker separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2109–2118, 2020.
[120] Z.-X. Li, Y. Song, L.-R. Dai, and I. McLoughlin, "Listening and grouping: An online autoregressive approach for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 4, pp. 692–703, Apr. 2019, doi: 10.1109/TASLP.2019.2892241.
[121] Z.-Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 71–75, doi: 10.1109/ICASSP.2019.8683231.
[122] Z.-Q. Wang, J. Le Roux, D. Wang, and J. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. Interspeech*, Sep. 2018, pp. 2708–2712.
[123] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network," in *Proc. Interspeech*, Sep. 2018, pp. 342–346, doi: 10.21437/Interspeech.2018-2290.
[124] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 46–50.
[125] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Proc. Interspeech*, Oct. 2020, pp. 2642–2646.
[126] F. Jiang and Z. Duan, "Speaker attractor network: Generalizing speech separation to unseen numbers of sources," *IEEE Signal Process. Lett.*, vol. 27, pp. 1859–1863, 2020.
[127] J. H. Lee, J. H. Chang, J. M. Yang, and H. G. Moon, "NAS-TasNet: Neural architecture search for time-domain speech separation," *IEEE Access*, vol. 10, pp. 56031–56043, 2022.
[128] F.-L. Wang, Y.-F. Cheng, H.-S. Lee, Y. Tsao, and H.-M. Wang, "CasNet: Investigating channel robustness for speech separation," 2022, *arXiv:2210.15370*.
[129] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "Effective low-cost time-domain audio separation using globally attentive locally recurrent networks," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 801–808.
[130] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 21–25.
[131] W. Qiu and Y. Hu, "Dual-path hybrid attention network for monaural speech separation," *IEEE Access*, vol. 10, pp. 78754–78763, 2022.
[132] E. Tzinis, Z. Wang, and P. Smaragdis, "SuDoRM-RF: Efficient networks for universal audio source separation," in *Proc. IEEE 30th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2020, pp. 1–6.
[133] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, "FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *Proc. Int. Conf. Multimedia Modeling*. Cham, Switzerland: Springer, 2020, pp. 653–665.
[134] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "Sandglasset: A light multi-granularity self-attentive network for time-domain speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 5759–5763, doi: 10.1109/ICASSP39728.2021.9413837.

[135] X. Hu, K. Li, W. Zhang, Y. Luo, J.-M. Lemercier, and T. Gerkmann, "Speech separation using an asynchronous fully recurrent convolutional neural network," in *Proc. NeurIPS*, vol. 34, 2021, pp. 22509–22522.
[136] Y. Zhao, C. Luo, Z.-J. Zha, and W. Zeng, "Multi-scale group transformer for long sequence modeling in speech separation," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 3251–3257.
[137] G.-P. Yang, C.-I. Tuan, H.-Y. Lee, and L.-S. Lee, "Improved speech separation with time-and-frequency cross-domain joint embedding and clustering," in *Proc. Interspeech*, Sep. 2019, pp. 1363–1367.
[138] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[139] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4945–4949.
[140] X. Hao, C. Shan, Y. Xu, S. Sun, and L. Xie, "An attention-based neural network approach for single channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6895–6899.
[141] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.
[142] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. Dimitriadis, J. Droppo, and Y. Gong, "Single-channel speech extraction using speaker inventory and attention network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 86–90.
[143] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.
[144] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6875–6879.
[145] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1003–1012.
[146] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.
[147] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.
[148] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[149] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

SWATI SONI received the B.Tech. degree in electronics and instrumentation and the M.Tech. degree in digital communication from the Samrat Ashok Technological Institute, Vidisha, Madhya Pradesh, India, in 2010 and 2016, respectively. She is currently pursuing the Ph.D. degree with the Maulana Azad National Institute of Technology, Bhopal. She is also working on performance enhancement of audio source separation using deep learning for her Ph.D. dissertation work. Her current research interests include audio signal processing and deep learning applications.

RAM NARAYAN YADAV received the B.Tech. degree in electronics from the Motilal Nehru National Institute of Technology (MNREC) Allahabad, India, in 1993, the M.Tech. degree in digital communication from the Maulana Azad National Institute of Technology (MACT), Bhopal, India, in 1997, and the Ph.D. degree in electrical engineering from the IIT Kanpur, India, in 2005. In 1997, he joined the Department of Electronics and Communication Engineering, Maulana Azad National Institute of Technology, where he is currently a Professor. He has published more than 70 articles in various international journals and conferences, including journals of IEEE, IET, Elsevier, and Springer. He has supervised 25 master's and ten Ph.D. students. His current research interests include communication systems, image processing, and neural networks and applications. He is also a reviewer of several IEEE and Elsevier journals.

LALITA GUPTA (Senior Member, IEEE) received the B.E. degree in electronics and telecommunication engineering from Pt. Ravishanker Shukla University, Raipur, in 2003, and the M.Tech. degree in digital communication and the Ph.D. degree in electronics and communications engineering from the Maulana Azad National Institute of Technology, Bhopal, India, in 2007 and 2012, respectively. Since July 2004, she has been associated as a Faculty Member with the Maulana Azad National Institute of Technology, where she is currently an Associate Professor with the Department of Electronics and Communication Engineering. She has 30 research publications in various international journals and conferences of repute. Her research interests include signal processing and image processing. She is a member of IETE and ICEIT.