big data and cognitive computing

MDPI

Review

Speech Separation Using Advanced Deep Neural Network Methods: A Recent Survey

Zeng Wang [1] and Zhongqiang Luo [1,2,*]

[1] School of Automation and Information Engineering, Sichuan University of Science & Engineering, Yibin 644000, China; 324081104122@stu.suse.edu.cn
[2] Intelligent Perception and Control Key Laboratory of Sichuan Province, Sichuan University of Science & Engineering, Yibin 644000, China
* Correspondence: luozhongqiang@suse.edu.cn

Abstract

Speech separation, as an important research direction in audio signal processing, has been widely studied by the academic community since its emergence in the mid-1990s. In recent years, with the rapid development of deep neural network technology, speech processing based on deep neural networks has shown outstanding performance in speech separation. While existing studies have surveyed the application of deep neural networks in speech separation from multiple dimensions including learning paradigms, model architectures, loss functions, and training strategies, current achievements still lack systematic comprehension of the field's developmental trajectory. To address this, this paper focuses on single-channel supervised speech separation tasks, proposing a technological evolution path "U-Net–TasNet–Transformer–Mamba" as the main thread to systematically analyze the impact mechanisms of core architectural designs on separation performance across different stages. By reviewing the transition process from traditional methods to deep learning paradigms and delving into the improvements and integration of deep learning architectures at various stages, this paper summarizes milestone achievements, mainstream evaluation frameworks, and typical datasets in the field, while also providing prospects for future research directions. Through this detailed-focused review perspective, we aim to provide researchers in the speech separation field with a clearly articulated technical evolution map and practical reference.

Keywords: source separation; deep neural network; sequence modeling; U-net; TasNet; Transformer; Mamba

Academic Editor: Giuseppe Ciaburro

Received: 5 October 2025
Revised: 10 November 2025
Accepted: 11 November 2025
Published: 14 November 2025

Citation: Wang, Z.; Luo, Z. Speech Separation Using Advanced Deep Neural Network Methods: A Recent Survey. *Big Data Cogn. Comput.* **2025**, *9*, 289. https://doi.org/10.3390/bdcc9110289

1. Introduction

Language is one of the most important functions that distinguishes humans from other animals. Humans have achieved rapid information exchange through speech: the most convenient and efficient means of communication. With the rapid development of digital signal processing technology, significant attention has been given to speech processing technology. This includes multiple technical fields, such as automatic speech recognition (ASR) [1], speech synthesis [2], which have advanced rapidly and entered the practical application phase.

However, in practical application scenarios, audio captured by microphones typically contains multiple components such as background noise, interference from other speakers, and environmental reverberation. Directly feeding such mixed signals into an automatic

speech recognition (ASR) system would lead to a significant decline in recognition accuracy due to severe acoustic contamination [3]. Therefore, incorporating a speech separation module at the front end of the ASR system to extract the target speaker's clean speech from the mixed signal in advance is of critical importance for enhancing the robustness of the overall recognition system. This module can be regarded as a front-end enhancement component of the ASR system, with its core task being to address the source separation problem of "who is speaking," thereby providing a high-quality speech input foundation for the back-end task of recognizing "what is being said".

Speech separation problems, also known as the "cocktail party problem" [4], separate the speech signals of individual speakers from a mixed signal containing multiple speakers; Cherry first proposed this problem in 1953. Researchers found that humans can easily distinguish and focus on target speech in complex acoustic environments where multiple people speak simultaneously. However, this task is extremely difficult for machines.

Today, speech separation technology is critical for separating and denoising audio signals, ensuring the proper functioning of applications in communication, speech recognition, and speaker identification. With the improvement of computer system computing power and the development of big data technology and machine learning-related theories, deep neural networks have been widely applied in fields such as image recognition, video recommendation, text mining, and audio processing. In the field of speech separation, methods using deep neural networks have achieved better results than traditional separation methods in various competitions. Therefore, the application of deep neural networks to solve speech separation problems has attracted widespread attention from researchers.

While numerous review studies have contributed valuable insights to this field (as summarized in Table 1), the relentless pace of innovation and the extensive scope of speech separation technology present a challenge: even the most contemporary reviews find [5] it difficult to delineate a comprehensive and coherent developmental trajectory. In response to this shortcoming, the present paper endeavors to introduce a novel reviewing viewpoint. It undertakes a systematic retrospective examination of speech separation technologies predicated on deep neural networks, paying particular attention to the single-channel context. By conducting an in-depth technical tracing, this work charts the progression from theoretical underpinnings to cutting-edge models, with a pivotal analysis devoted to the core design philosophies and technical contributions of mainstream architectures, such as U-Net, TasNet, Transformer, and Mamba. It is anticipated that this linear, evolutionary perspective will furnish readers with a lucid and efficient pathway into the subject, thereby aiding in the comprehension of its pivotal advancements and prospective directions.

Table 1. Past and latest reviews.

Refs.	Years	Title
[6]	2018	An overview of lead and accompaniment separation in music
[7]	2018	Supervised speech separation based on deep learning: An overview
[1]	2023	Deep neural network techniques for monaural speech enhancement and separation: state of the art analysis
[8]	2023	A survey of artificial intelligence approaches in blind source separation
[5]	2025	Advances in speech separation: Techniques, challenges, and future trends

Based on a systematic review of the existing literature and methods in the field of speech separation, this paper focuses on deep learning-based single-channel supervised speech separation tasks, providing an in-depth analysis of its developmental trajectory and representative frameworks. The main contributions of this paper can be summarized as follows:

- We conducted a systematic review and evaluation of the developmental trajectory and practical performance of various deep learning techniques in single-channel supervised speech separation.
- We clearly identified existing research gaps and core challenges within the field, while elucidating the critical role of rational deep learning architectures in addressing these issues.
- We prospectively outlined future directions and proposed a series of promising research avenues with the potential to drive sustained advancement in the field.

The remainder of this paper is organized into six sections, as illustrated in Figure 1.

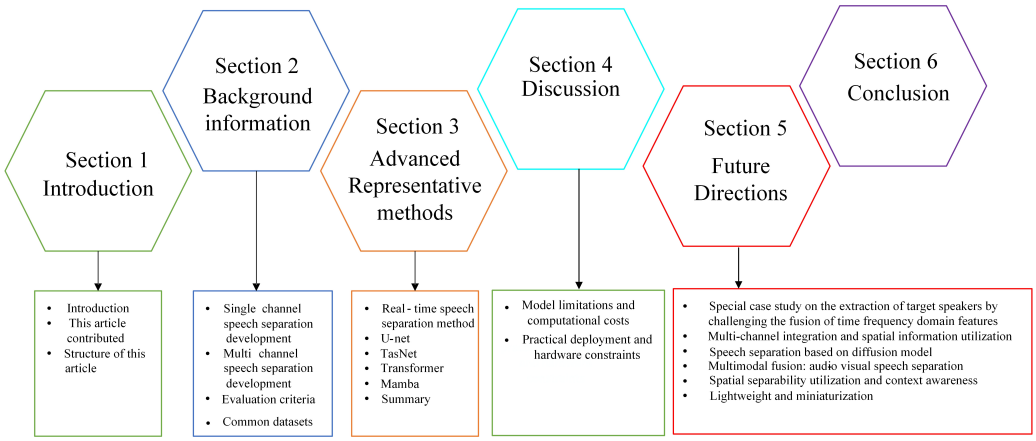


Figure 1. The overall structure of the topics covered in this article.

Section 1 is the introduction. Section 2 discusses the background of speech separation, including evaluation metrics and datasets. Section 3 offers a detailed review of recent advances in speech separation based on advanced models such as U-Net, TasNet, Transformer, and Mamba. Section 4 delves into a discussion on model limitations, computational costs, performance-efficiency trade-offs, and the challenges associated with practical deployment and hardware constraints. Section 5 analyzes the core challenges in the field and outlines promising future research directions. Finally, Section 6 concludes the paper.

2. Background Information

Based on the type of interference source, speech separation tasks are primarily categorized into three types: speech enhancement (where the interference source is noise signals), multi-speaker separation (where the interference source is the speech of other speakers), and reverberation cancellation (where the interference source is the reflected waves of the target speaker's own speech). Additionally, based on the number of microphones used to capture the signal, speech separation methods can be further categorized into single-channel speech separation methods [9] (where the mixed speech signal is captured by a single microphone. This means the system cannot directly obtain spatial directional information of different speakers) and multi-channel speech separation methods [10] (where the mixed speech signal is captured by an array of microphones arranged in a specific geometric configuration, with the microphone array providing critical spatial information).

2.1. Overview of the Development of Single-Channel Speech Separation

The task of single-channel speech separation is to estimate the overlapping speech signals from a monaural mixed signal such as Figure 2.

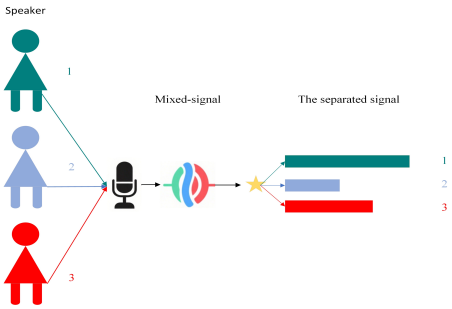


Figure 2. Single-channel speech separation diagram.

Single-channel speech separation usually adopts a linear instantaneous mixing model, and its mathematical expression is shown in Formula (1).

$$\boldsymbol{y} = \sum_{i} \boldsymbol{s}_i + \boldsymbol{n} \quad i \in \{0, ..., C-1\} \tag{1}$$

The mixed signal $\boldsymbol{y} = \{y_0, ..., y_{L-1}\}$ collected by the microphone is a linear mixing in the time domain of the speech signals $\boldsymbol{s}_i = \{s_{i,0}, ..., s_{i,L-1}\}$ ($i \in \{0, ..., C-1\}$) of C speakers and the background noise $\boldsymbol{n} = \{n_0, ..., n_{L-1}\}$. L represents the number of signal sampling points.

Since the 1950s, researchers have been exploring methods for single-channel speech separation. Before the rise of deep learning methods, these methods mainly relied on signal processing and statistical modeling techniques. However, these early methods were often based on ideal assumptions or heavily relied on carefully designed heuristic rules, which limited their separation performance.

Traditional single-channel speech separation methods that focus on signal processing—primarily aimed at speech enhancement in the presence of non-speech noise—have evolved in two main directions, time domain and frequency domain.

Time-domain methods include "parameter and filter methods", which are based on channel parameters and the estimation of excitation parameters, as referenced in citation [11,12]. Another approach is "signal subspace methods," which decompose mixed speech signals into distinct subspaces for target speech and interference noise, as noted in citation [13].

Frequency-domain methods encompass techniques such as spectral subtraction [14], Wiener filtering [15], and algorithms based on minimum statistics [16] or minimum controlled recursive averaging [17]. Generally, these algorithms first utilize non-speech frames to estimate statistical parameters, like prior or posterior signal-to-noise ratios for the interference noise. They then combine these statistics with a predefined noise model to estimate the target speech signal.

However, traditional frequency domain methods encounter significant challenges when addressing non-stationary noise. The statistical characteristics of non-stationary noise, such as its power spectra, make it difficult to accurately estimate the noise using only information from previous non-speech frames. This leads to inaccurate noise updates and two major issues. The first issue is the underestimation of noise energy. Even algorithms based on the minimum mean square error criterion may fail to estimate noise energy accurately in certain scenarios. This can result in substantial residual noise within the separated speech, severely degrading performance metrics. The second issue revolves around the incorrect classification of speech and non-speech frames. These methods typically depend on voice activity detection (VAD) to differentiate between speech frames and non-speech noise frames. However, in conditions of low signal-to-noise ratio and non-stationary noise, the effectiveness of VAD declines sharply.

This deterioration leads to increased classification errors, resulting in speech frames being incorrectly suppressed or lost.

Methods based on the Vector Quantization (VQ) model [18] represent an early data-driven approach, whose core lies in speech separation and modeling the distribution of speech features. Its fundamental assumption is that, in an appropriate feature space, speech features from different speakers occupy distinct regions or cluster centers; mixed speech feature vectors can be regarded as linear combinations or probabilistic mixtures of independent source feature vectors; and the speech feature distribution of each speaker can be approximated by a codebook containing typical spectral feature code vectors. This method systematically introduced the concepts of feature space clustering and sound source modeling into the field of speech separation for the first time, laying the theoretical foundation for subsequent statistical models (GMM, HMM, NMF) and deep learning methods. However, due to its discrete characteristics and linear assumptions, this method has limitations.

To address the discretization issue of the VQ model, the Gaussian mixture model (GMM) [19] extended the discrete codebook to a continuous probability model and achieved successful applications in speech recognition. However, both VQ and GMM assume that each speech frame is statistically independent. This does not fully align with the short-time stationary and frame-dependent characteristics of speech signals.

To overcome the limitations of this assumption, the Hidden Markov Model (HMM) [20] was introduced. By constructing multiple independent HMMs and forming a factor HMM, the relationship between mixed and original speech was described. Nevertheless, this method suffers from exponential growth in computational complexity as the number of speakers increases, relies on phoneme-level annotation data, and performs poorly in scenarios involving rapid speaker switching.

Building upon parameterized temporal models, a method based on non-negative matrix factorization (NMF) [21] shifts toward a more flexible data-driven approach. It assumes that the amplitude spectra of the target speech and interference noise can be represented by a set of basis signals, which are learned through data-driven methods. In the separation stage, the trained basis signals are used to decompose the mixed speech, extract the weight components of the target speech, and reconstruct the signal. However, this method fails to fully utilize the characteristics and regularities of the speech signal itself and has high computational complexity.

Inspired by interdisciplinary research, Computational Auditory Scene Analysis (CASA) [22] (derived from the study of the perception mechanisms of the human auditory system in complex acoustic scenes) points out that the human auditory and visual systems have similarities. CASA performs auditory peripheral analysis on mixed speech to obtain its time–frequency domain representation, and then uses time–frequency segments formed by acoustic features, grouped cues (or trained source background knowledge) for scene combination, and finally reconstructs the waveform signals of the separated speech. However, the algorithmic rules for each step in the CASA system are mostly based on manual design or heuristic summaries from limited data. This results in poor generalization capabilities when faced with various types of noise and complex acoustic environments in practical applications, and it is unable to fully leverage the advantages of big data to learn the complex patterns and dependencies between target speech and interfering noise.

Thanks to the rapid development of computing power, deep learning technology has once again gained attention. Using these methods, the field of speech separation has become increasingly sophisticated. Today's speech separation systems can not only more accurately separate the voice of the target speaker in noisy environments, but also effectively handle overlapping speech, cope with complex acoustic environments, and gradually

achieve real-time or near-real-time processing capabilities. Deliang Wang's team proposed the Ideal Binary Mask (IBM) [23], which is based on the auditory masking effect of CASA and converts speech separation into a binary classification problem of time–frequency units. Lu et al. [24] proposed a mapping-based speech separation method, training multiple autoencoders to learn the mapping relationship between noisy speech and clean speech. Erdogan [25] improved the network structure and mask target, incorporating phase difference information between clean and noisy speech to compensate for model estimation errors. Williamson [26] found that phase is critical for perceived quality, proposing to simultaneously enhance amplitude and phase spectra in the complex domain using deep neural networks. By estimating the real and imaginary parts of the complex ideal ratio mask, they significantly improved multiple metrics.

In speaker-related scenarios, algorithms that directly estimate the time–frequency mask perform well. However, in speaker-target-unrelated or mixed scenarios, the effectiveness of these methods decreases significantly due to the inherent permutation ambiguity problem in speech separation.

When the model needs to separate a mixture of speech from multiple speakers, it outputs an estimated time–frequency mask or speech signal for each speaker. However, during training or inference, the model cannot pre-determine which output channel should correspond to which real target speaker. The model may allocate speaker A's output to channel A in one frame and channel B in the next.

To solve the permutation problem, researchers have proposed various solutions from different angles.

The deep clustering method was proposed by Hershey et al. [27] This method uses deep neural networks to learn a high-dimensional embedding vector for each mixed speech spectrogram's time–frequency (TF) unit. The key idea is to make the embedding vectors of TF units belonging to the same speaker close to each other in the embedding space, while those of different speakers are far apart. Subsequently, clustering algorithms are applied to these embedding vectors to directly estimate the ideal binary mask (IBM) corresponding to different speakers, avoiding the problem of output order permutation.

Permutation Invariant Training was proposed by Yu et al. The PIT method [28] allows neural networks to freely output signals from multiple speaker channels. Its core innovation lies in dynamically enumerating all possible permutations of output channels and target speakers during training, and automatically selecting the permutation that minimizes the loss function (e.g., scale-invariant signal-to-noise ratio loss, SI-SNR loss) as the "correct" label for the current training sample. In this way, the network learns a mapping invariant to the order of output sequences, effectively addressing the issue of uncertainty in training label order.

Furthermore, a fusion method combining deep embedded features and discriminative learning is proposed [29], integrating the advantages of deep clustering (DC) and ranking invariance training (PIT). Combining DCs with the end-to-end optimization framework of PIT through joint training or loss function design achieves better separation performance than single methods.

These methods fundamentally resolve the permutation ambiguity challenge in training multi-speaker separation models. Thanks to continuous innovation and significant breakthroughs in these areas, modern deep learning-based speech separation models have successfully established a solid mathematical framework (such as the loss function mechanism of PIT and the geometric properties of the embedding space in DC) and have been extensively validated through empirical experiments. This has not only greatly advanced the development of speech separation but also laid a reliable theoretical and practical foundation for its practical applications.

To facilitate understanding, the aforementioned content has been presented in a structured format in the Table 2.

Table 2. Comparison of Optimized Speech Separation Methods

Category	Type	Method	Core Concept	Advantages	Limitations
Single-Channel	Traditional	Time-domain [11–13]	Direct parameter estimation or subspace decomposition	Low complexity, minimal hardware	Limited in non-stationary noise
		Frequency-domain [14–17]	Spectral noise statistics estimation	Cost-effective, real-time capable	VAD and noise estimate dependent
		VQ [18]	Discrete codebook feature modeling	Pioneered data-driven approaches	Limited generalization
		GMM [19]	Continuous Gaussian mixture modeling	Superior to VQ for continuous features	Assumes frame independence
		HMM [20]	Temporal dynamic modeling with factorial models	Captures speech temporal structure	Exponential complexity with speakers
		NMF [21]	Spectral basis + weight decomposition	Flexible, assumption-free	Computationally intensive
		CASA [22]	Human auditory perception simulation	Biologically plausible	Rule-limited generalization
	Deep Learning	IBM [23]	Time–frequency binary classification	Intuitive, reliable baseline	Permutation ambiguity
		Autoencoder [24]	End-to-end spectral mapping	Feature learning automation	Permutation issues persist
		PSM/cIRM [25,26]	Complex spectral mask estimation	Enhanced quality	Multi-speaker challenges remain
		Deep Clustering [27]	Embedding space clustering	Solves permutation ambiguity	Computational overhead
		PIT [28]	Minimum loss assignment training	End-to-end efficient	Data hungry
		DC+PIT [29]	Hybrid clustering + PIT framework	Accuracy + efficiency balance	Increased complexity
Multi-Channel	Traditional	ICA/IVA	Statistical independence separation	Blind separation capable	Reverberation sensitive
		Beamforming [30,31]	Spatial filtering	Robust interference rejection	Hardware cost, direction dependency
	Deep Learning	Neural Beamforming [32]	DNN-based weight estimation	Physical + neural integration	Training complexity
		DNN Masks	Spatial + spectral masking	Mature mask approach	Layout sensitivity
		DAE/CDAE	Implicit spatial learning	Delay robust, real-time	Debugging challenges

2.2. Overview of the Development of Multi-Channel Speech Separation

Multi-channel speech separation methods rely on signals collected by microphone arrays with fixed geometric configurations as shown in Figure 3. The spatial distribution of the microphone arrays causes relative time delays and spatial characteristics between the collected signals. Time domain, frequency domain, and spatial information can be comprehensively utilized to separate mixed speech. Multi-channel speech separation methods significantly improve separation performance and robustness in complex acoustic environments. However, this comes at higher costs and algorithm complexity.

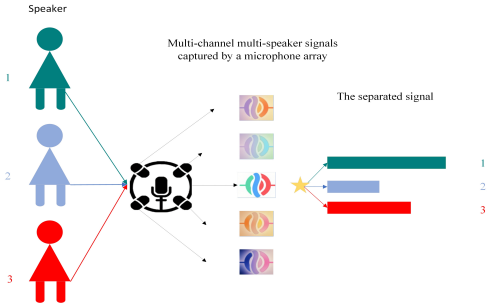


Figure 3. Multi-channel speech separation diagram.

Most of the real scenarios for speech separation applications belong to indoor scenarios. The signals collected by the microphone array include the direct signal from the sound source to the microphone and the reverberant signal after reflection. At this time, a reverberation model needs to be established. If there are C sound sources in the sound field, the signal received by the $m \in \{0, ..., M-1\}$-th microphone in the array can be expressed as follows:

$$y_m(k) = \sum_{i=0}^{C-1} x_{mi}(k) + n_m(k) = \sum_{i=0}^{C-1} g_{mi} \circledast s_i(k) + n_m(k) \quad (2)$$

In formula (2) $x_{mi}(k) = g_{mi} \circledast s_i(k)$, $\circledast$ represents the convolution operation, g_{mi} is the room impulse response (Room Impulse Response, RIR) from sound source i to microphone m, n_m represents the environmental noise at microphone m. It is assumed that the noise signal represented by $n_m(k)$ is uncorrelated with the sound source signal.

Multi-channel speech separation methods can be broadly categorized into two types. The first type does not utilize additional information, such as independent component analysis (ICA) and independent vector analysis (IVA) (often extensions of single-channel methods). The second type utilizes additional information obtained through the geometric structure of microphone arrays, such as the spatial location of speakers and speech time boundaries.

Beamforming technology is a representative algorithm that utilizes additional information. At its core, it functions as a spatial filter, which can be categorized into two types based on its filtering method. The first type is fixed beamforming, where the microphone array is fixed, the direction is fixed, and the filter coefficients remain constant. Examples include filter summation [30] and delay summationv [31]. The second category is adaptive beamforming, where the filter coefficients are adaptively adjusted based on the environmental acoustic characteristics to achieve directional optimization. Examples include linear-constrained minimum variance (LCMV), generalized side lobe cancellation (GSC), and minimum variance without distortion response (MVDR).

Deep learning methods have achieved significant performance improvements in multi-channel speech separation tasks compared to traditional signal processing methods. Based on the model's core role or output type, these methods can be mainly categorized into two types, neural beamformers and DNN time–frequency masking.

Neural Beamformer methods integrate DNN with traditional beamforming principles to optimize the key parameter estimation of beamformers using DNN. Representative work includes the All-Deep-Learning MVDR (ADL-MVDR) beamformer proposed by Zhang [32]. Its core innovation lies in abandoning the traditional MVDR solution, which directly performs matrix inversion and feature decomposition to estimate target direction vectors. Instead, it uses two recurrent neural networks (RNNs) to model these complex processes, directly predicting frame-level beamforming weights. Complex ratio filtering (CRF) technology is also introduced to effectively enhance the joint training stability with the front-end filter estimator.

The DNN-based time–frequency masking estimation method extends the successful paradigm of masking estimation in single-channel speech separation. Deep neural networks are directly employed to estimate masking values for each TF unit of mixed speech. Mask types include phase-sensitive masks (PSM) [25], complex IRM [26], and ideal ratio masks (IRM) [33].

Recently, denoising autoencoders (DAEs) have attracted attention, among which convolutional denoising autoencoders (CDAEs) have shown potential in multi-channel separation, as they can directly handle differences in signal arrival times. Ref. [34] explores the effective-

ness of this method in multi-channel separation tasks. The model maps noisy speech to clean speech in the time domain and learns spatial information in an end-to-end manner.

2.3. Evaluation Criteria

The purpose of speech separation is to improve communication quality. For speech recognition-related applications, speech separation aims to enhance the recognition rate of speech recognition systems. Therefore, the performance evaluation criteria for speech separation are not singular; in practical applications, three evaluation metrics are typically used: signal-level evaluation metrics, perceptual-level evaluation metrics, and application-level evaluation metrics.

Signal-level evaluation metrics include the signal distortion ratio (Source-to-Distortion Ratio, SDR), Signal-to-Noise Ratio (SNR), Signal-to-Interference Ratio (SIR), and System Error Ratio (SAR), which were proposed by Vincent et al. in 2006 as performance evaluation functions for audio blind source separation algorithms [35]. Their definitions are as follows (typically expressed in dB):

$$SDR := 10\log_{10}\frac{\|\boldsymbol{s}_{target}\|^2}{\|\boldsymbol{e}_{interf}+\boldsymbol{e}_{noise}+\boldsymbol{e}_{artif}\|^2} \tag{3}$$

$$SNR := 10\log_{10}\frac{\|\boldsymbol{s}_{target}+\boldsymbol{e}_{interf}\|^2}{\|\boldsymbol{e}_{noise}\|^2} \tag{4}$$

$$SIR := 10\log_{10}\frac{\|\boldsymbol{s}_{target}\|^2}{\|\boldsymbol{e}_{interf}\|^2} \tag{5}$$

$$SAR := 10\log_{10}\frac{\|\boldsymbol{s}_{target}+\boldsymbol{e}_{interf}+\boldsymbol{e}_{noise}\|^2}{\|\boldsymbol{e}_{artif}\|^2} \tag{6}$$

The processed source estimation signal $\hat{S}$ is decomposed into four parts, including the true source signal $\boldsymbol{s}_{target}$, interference from other source signals $\boldsymbol{e}_{interf}$, interference from additive noise $\boldsymbol{e}_{noise}$, and interference caused by human factors in the separation algorithm $\boldsymbol{e}_{artif}$.

In practical applications, the energy in the separated source estimation signal that does not belong to the true source signal is regarded as distortion. SDR then becomes the following formula:

$$SDR := \frac{1}{C}\sum_{i=0}^{C-1} 10\log_{10}\frac{E\left(|s_i(k)|^2\right)}{E\left(|\hat{s}_i(k)-s_i(k)|^2\right)} \tag{7}$$

Here, $E(\cdot)$ denotes the mean value operation, C denotes the number of sound source signals, $\hat{s}_i(k)$ and $s_i(k)$ denote the first estimated sound source signal i and the second true sound source signal i obtained at time k, respectively.

Maximizing the scale-invariant source-to-noise ratio (SI-SNR) is commonly used as an evaluation metric for source separation.

$$\begin{cases} \boldsymbol{s}_{target} := \dfrac{\langle \hat{\boldsymbol{s}}, \boldsymbol{s}\rangle \boldsymbol{s}}{\|\boldsymbol{s}\|^2} \\ \boldsymbol{e}_{noise} := \hat{\boldsymbol{s}} - \boldsymbol{s}_{target} \\ \text{SI-SNR} := 10\log_{10}\dfrac{\|\boldsymbol{s}_{target}\|^2}{\|\boldsymbol{e}_{noise}\|^2} \end{cases} \tag{8}$$

Among these, $\hat{\boldsymbol{s}} \in \mathbb{R}^{1\times T}$ and $\boldsymbol{s} \in \mathbb{R}^{1\times T}$ represent the estimated and original clean sources, respectively, while $\|\boldsymbol{s}\|^2 = \langle \boldsymbol{s}, \boldsymbol{s}\rangle$ denotes the signal power, to ensure scale invariance, $\hat{\boldsymbol{s}}$ and $\boldsymbol{s}$ are normalized to a zero mean before computation.

Commonly used perceptual-level evaluation metrics include the Perceptual Evaluation of Speech Quality (PESQ) [36] and Short-Term Objective Intelligibility (STOI) [37], which measure intelligibility scores and human speech quality scores, respectively, both of which are closely related to human auditory perception.STOI calculates the correlation coefficient between the short-term envelopes of the target speech and the processed speech, with values ranging from [0, 1], where higher values indicate better intelligibility of the processed speech; PESQ uses a cognitive model to calculate the disturbance between the target speech and the processed speech, with values ranging from -0.5 to 4.5, where higher values indicate higher quality of the processed speech.

Application-level evaluation involves selecting relevant application systems based on specific scenarios to test the separated speech.

2.4. Commonly Used Data Sets

In related studies, scholars have adopted diverse datasets to support different types of speech processing tasks. Table 3 shows the commonly used datasets for speech separation tasks.

The TIMIT dataset [38] contains high-quality recordings from 630 speakers of 8 American English dialects, each reading up to 10 phonetically rich sentences (6300 segments). It is annotated with phoneme and word boundaries and covers all English phonemes.

The WSJ0-2mix dataset [39] is constructed from the WSJ0 corpus by mixing speech segments from multiple speakers in the time and frequency domain to simulate multi-person dialogues. It uses varying signal-to-noise ratios and time offsets for realism and complexity.

The MIR-1K dataset [40] is the first public dataset for vocal separation, consisting of 1000 Chinese pop song clips recorded at 16 kHz.

The iKala dataset [41] addresses MIR-1K's limitations (short clips, lack of non-vocal regions) by providing 352 30-s clips with longer instrumental solos.

The MUSDB18 dataset [42] contains 150 full-length music tracks (10 h) across various genres, provided with isolated stems for drums, bass, vocals, and other instruments. It includes separate "train" (100 songs) and "test" (50 songs) sets.

The LibriSpeech dataset [43] contains approximately 1000 h of 16kHz read English speech.

The VCTK dataset [44] includes speech from 50 speakers with different accents, each recording a large number of sentences in various contexts.

The LibriTTS dataset [45] is derived from LibriSpeech, containing 585 h of 24 kHz speech data from 2456 speakers and corresponding text.

The WHAM! dataset [46] enhances the WSJ0-2mix dataset by matching two-speaker utterances with unique, realistic noise backgrounds recorded in non-stationary environments like cafes and bars.

The SMS-WSJ dataset [47] is a multi-channel dataset generated from WSJ0 and WSJ1, featuring two overlapping speakers with Gaussian noise and supporting 2 to 6 channels.

The LibriMix dataset [48] is an open-source alternative to WSJ0-2mix, based on LibriSpeech and WHAM! noise, containing mixtures of two or three speakers with environmental noise.

Mixed datasets and Synthetic datasets. Mixed datasets are constructed by integrating subsets from multiple existing datasets; synthetic datasets are algorithmically generated or created using software tools by the authors for specific research purposes. For detailed usage of these datasets in model training, please refer to the relevant descriptions in the respective original literature.

Table 3. Commonly used datasets for speech separation tasks.

Name	Refs.	Year	Content
TIMIT	[38]	1993	8 US English dialects from 630 speakers; 6300 audio clips.
WSJ0	[39]	2007	Mixed-speech dataset generated from WSJ0 corpus.
MIR-1K	[40]	2009	1000 Chinese pop song clips (16 kHz); 352 CD-quality 30-s excerpts.
iKala	[41]	2015	352 thirty-second music clips with instrumental solos.
LibriSpeech	[43]	2015	1000 h of 16 kHz English speech.
MUSDB18	[42]	2017	150 full music tracks (10 h total duration).
VCTK	[44]	2019	Speech from 50 speakers across different regions.
LibriTTS	[45]	2019	585 h of 24 kHz speech from 2456 speakers with text.
WHAM!	[46]	2019	WSJ0-2mix speech combined with unique noise backgrounds.
SMS_WSJ	[47]	2019	Multi-channel dataset (2–6 channels) based on WSJ0 and WSJ1.
LibriMix	[48]	2020	LibriSpeech speakers mixed with WHAM! noise samples.

3. Advanced Representative Methods

End-to-end separation models based on deep neural networks, such as Wave-U-Net, Conv-TasNet, Transformer, and Mamba, have achieved remarkable success. In the following sections, this paper will introduce the development of speech separation based on deep learning architecture methods, and further discuss and analyze the latest progress of several representative technologies.

3.1. Progress in Real-Time Speech Separation Methods Based on Traditional Methods

In recent years, with the widespread adoption of remote work, online meetings, and smart assistants, there has been a surge in demand for low-latency, high-fidelity voice interaction. Addressing the critical technical challenge of real-time speech separation in complex acoustic environments, researchers have actively explored and proposed numerous efficient online algorithms.

Traditional speech separation methods remain widely studied and applied. Although deep learning (DL) methods have brought revolutionary advancements to the field, their inherent "black-box" nature (i.e., the difficulty in explaining the model's decision-making process) remains a significant challenge. As discussed in Section 2 traditional methods are typically built on rigorous mathematical derivations and a solid theoretical foundation.

Auxiliary function-based independent vector analysis (AuxIVA) converges quickly and has low computational costs, and its online algorithms have been widely applied. However, low latency requirements typically necessitate the use of shorter short-time Fourier transform (STFT) frames, which conflicts with separation performance (requiring frame lengths longer than the reverberation time).

To address this contradiction, researchers introduced the Weighted Prediction Error (WPE) algorithm [49] as an effective online microphone signal de-reverberation method, which achieves this through recursive updates of multi-channel linear prediction filters. A common strategy is to cascade AuxIVA after WPE (WPE + IVA) [50]. The advantage of this combination lies in its low computational complexity and high modularity, facilitating joint algorithm optimization.

Algorithmic improvements to enhance efficiency have also been explored. Some AuxIVA extensions, such as [51], employ Iterative Projection (IP) [52] to update the de-reverberation matrix. However, IP involves matrix inversion, which is computationally intensive. To improve efficiency, the Iterative Source Steering (ISS) algorithm [53] was proposed for batch AuxIVA.ISS is a matrix inversion-free update rule that can be generalized to other blind source separation (BSS) methods. Ref. [54] combined autoregressive estimation to propose a novel online AuxIVA algorithm without matrix inversion operations, significantly improving the speed of online BSS.

Ueda et al. explored the deep integration of WPE and independent vector extraction (IVE). They introduced a log-likelihood function with a forgetting factor for the online joint

optimization of the two methods and derived an efficient algorithm (online WPE × IVE) [55]. Compared to the separately optimized online WPE + IVE, this joint method achieves higher separation accuracy with shorter STFT frames. Additionally, to address scale ambiguity and transfer function error issues, the study proposed a spatially regularized online joint optimization algorithm.

Mo et al. [56] proposed a method combining online constrained beamforming (CBF) with non-causal sample truncation-based independent vector analysis (NST-IVA). This method utilizes long STFT analysis windows to ensure performance while reducing algorithm latency by truncating non-causal samples in the filter and shortening the convolution operation length used to update the de-reverberation filter.

He et al. noted that the sparsity of speech signals in the time–frequency (TF) domain can aid in constructing effective weights. Based on this, they proposed a weighted sparse component analysis (SCA) method [57]. This method employs a weighting scheme that only requires projecting the microphone array received signals onto microphone gains related to the sound source direction. The researchers designed two weighting minimization algorithms for sound source estimation in instantaneous mixing and convolution mixing scenarios, respectively.

3.2. Progress in Speech Separation Methods Based on the U-Net Architecture

The U-Net architecture initially achieved success in medical image segmentation and was subsequently introduced into the field of speech processing. Its core structure features a symmetric encoder-decoder architecture and is named for its distinctive U-shaped topology.

The U-Net architecture is illustrated in Figure 4 below.

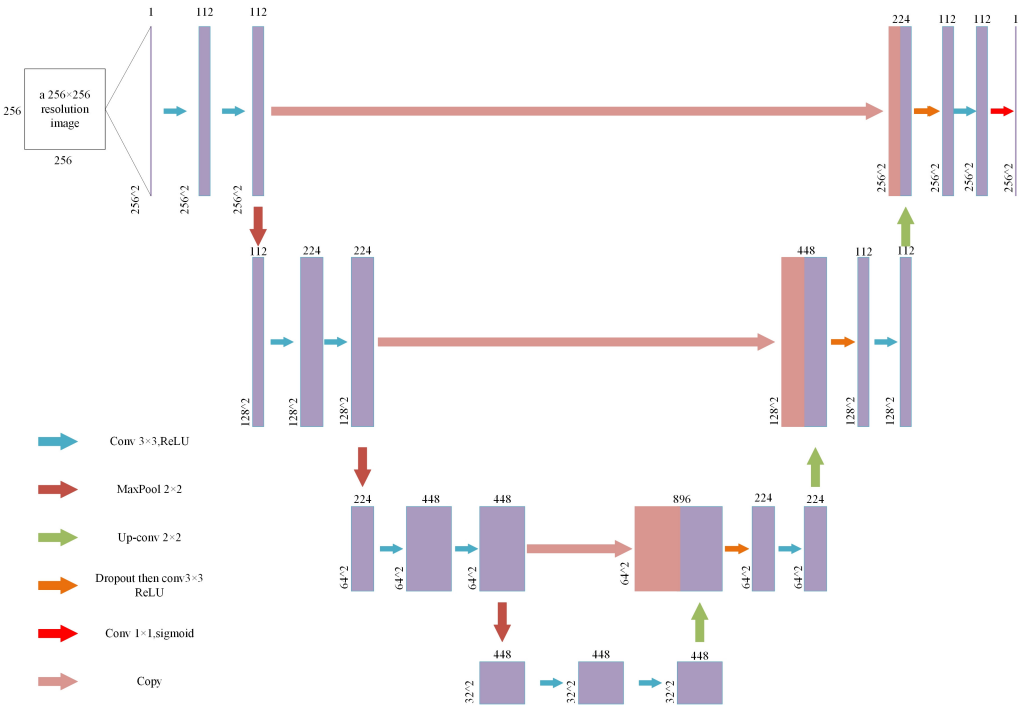


Figure 4. U-Net architecture.

Although U-net may not always be explicitly mentioned in current speech separation tasks, it is essential to emphasize that the most critical feature of the U-Net architecture lies in its skip connections between corresponding encoder and decoder layers. These connections enable the direct propagation of shallow, high-resolution features to deeper network levels.

Initially, the U-Net architecture was applied to segment the spectrogram of audio signals [58]. However, the inherent information loss and artifacts in time–frequency trans-

formation processes (such as short-time Fourier transform and its inverse transform) significantly limit the performance of spectrum-based speech separation methods.

In order to avoid the limitations of time–frequency transformation, researchers explored the path of end-to-end source separation directly in the time domain. Ref. [59] studied end-to-end source separation in the time domain and proposed Wave-U-Net, which is a one-dimensional time-domain adaptation of the U-Net architecture. Wave-U-Net is a deep learning model specifically designed for audio source separation tasks, operating on one-dimensional time-series data such as extracting vocals or accompaniment from mixed music signals. The model processes raw waveform inputs directly, aiming to capture temporal dependencies across audio samples.

The overall architecture is illustrated in the accompanying Figure 5. The input mixed audio waveform undergoes successive downsampling along the encoder path, progressively yielding multi-scale feature representations. The deeper layers capture global semantic information of the audio content, while the shallow features preserved via skip connections retain fine-grained structural details of the waveform. In the decoder path, the model learns to leverage these cross-rscale features to effectively distinguish characteristic patterns of different sound sources.

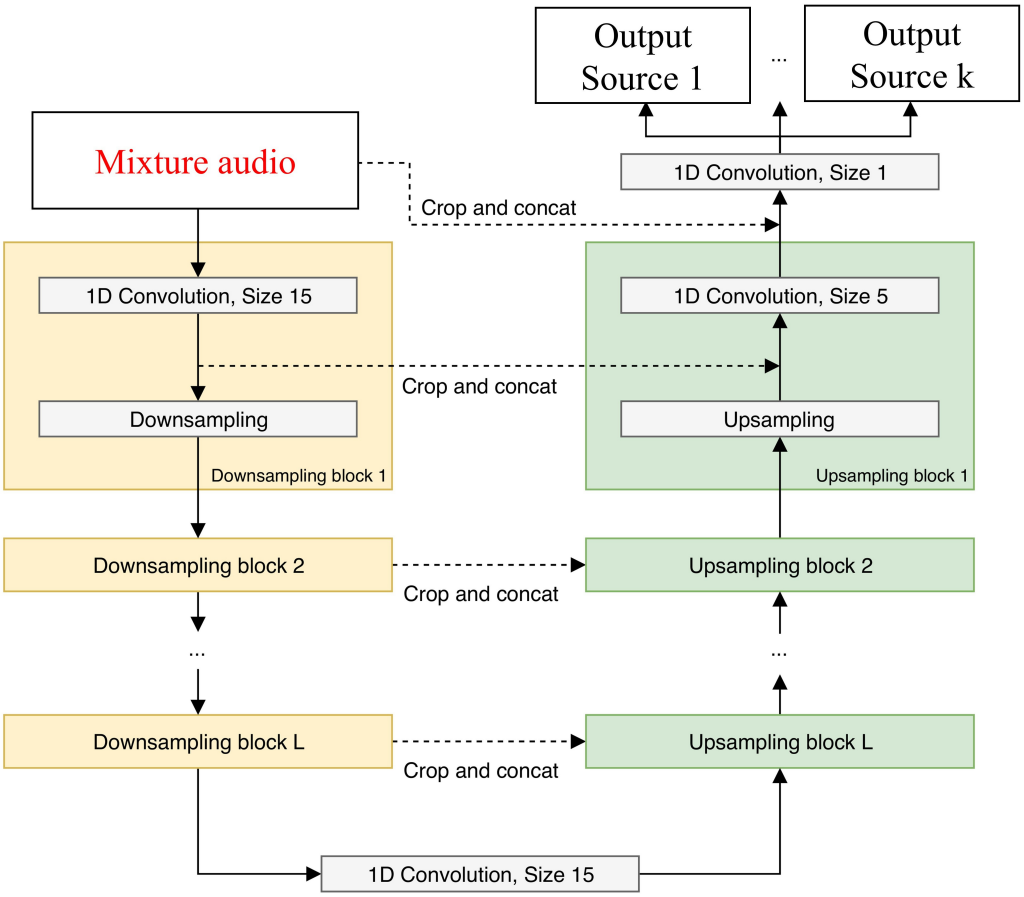


Figure 5. Wave-U-Net architecture.

At the output stage, instead of directly predicting target waveforms, Wave-U-Net estimates a time-varying soft mask for each target source. The final separated source signals are obtained by performing element-wise multiplication between each estimated mask and the original mixed waveform.

The wave-U-Net architecture retains the essential framework of the U-Net encoder-decoder structure along with skip connections, while introducing key modifications to better

accommodate the characteristics of audio waveform data—such as its one-dimensional nature and temporal sensitivity. A detailed comparison of their differences is provided in the Table 4.

Macartney et al. [60] further investigated the performance differences of Wave-U-Net across different layer numbers and variants, finding that model performance peaks around 9 to 10 layers. They speculate this may be related to the model's receptive field size and suggest that the optimal receptive field for speech signals may be smaller than that for music signals.

Table 4. Contrasting U-Net and Wave-U-Net Designs.

Characteristic	U-Net (for Images)	Wave-U-Net (for Audio)
Data Domain	2D Spatia	1D Temporal
Basic Operations	2D Convolution/Transposed Convolution	1D Convolution/Transposed Convolution
Downsampling	Reduces spatial size (H, W)	Reduces temporal length (L)
Upsampling	Restores spatial size (H, W)	Restores temporal length (L)
Skip Connection	Feature map concatenation	Feature map summation
Output	2D map (pixel labels)	1D waveform (samples)
Primary Concern	Spatial detail vs. context trade-off	Temporal/phase precision and artifact avoidance
Primary Objectives and Loss Functions	Predicted Spectrum-True Spectrum, L1 Norm	Reconstructed Target Waveform, MSE (Mean Square Error)

Since its initial proposal, the U-Net architecture has undergone years of development, achieving diverse improvements and widespread application in the fields of speech enhancement and separation. Researchers have conducted systematic exploration into key issues such as its long-range dependency modeling capability, feature fusion mechanisms in skip connections, as well as computational efficiency and real-time performance.

In terms of long-range dependency modeling, Défossez et al. from Facebook AI Research [61] proposed an end-to-end model named Demucs based on a masking approach. This model employs a convolutional U-Net as its backbone and incorporates a bidirectional long short-term memory (BiLSTM) network between the encoder and decoder to effectively capture long-term contextual dependencies in speech signals. Fu et al. [62] further introduced Uformer—an expanded complex-real dual-path network based on U-Net—capable of performing simultaneous speech enhancement and dereverberation in both the complex and magnitude domains. This model comprehensively utilizes temporal attention (TA) and dilated convolution (DC) to achieve cooperative modeling of temporal information at both global contextual and local detailed levels.

Regarding the optimization of skip connections, the simple feature concatenation in the original U-Net can easily lead to the propagation of noise or redundant information from the encoder to the decoder. To address this, researchers have proposed a series of intelligent feature selection mechanisms. The study cited as [63] introduced a learnable local internal attention mask into the skip connections, allowing the model to dynamically focus on speech-active regions while suppressing background noise. Visualization results indicate that this structure can autonomously learn functions similar to voice activity detection (VAD). He et al. [64] designed a dual-branch attention module (DBAUNet) that extracts spatial and channel features in parallel, thereby providing richer contextual representations for the skip connections. To further improve the precision of feature fusion, Zhang et al. [65] proposed a channel-spectral attention mechanism, which coordinately regulates information flow from both channel and spectral dimensions, achieving adaptive feature enhancement across the full frequency band.

Aiming at deployment on edge devices and meeting real-time interaction requirements, the lightweight design and low-latency optimization of U-Net have become research hotspots. Ref. [66] proposed a Hybrid Lightweight Time–Frequency Analysis network (HLTFA), which embeds a Lightweight Attention Fourier Module (LAFM) within the U-

Net framework to achieve efficient processing of global information in time–frequency representations. Bulut et al. [67] designed a streamlined U-Net architecture for real-time applications, demonstrating that optimized time-domain methods can also meet stringent low-latency requirements, thereby overcoming the conventional perception that time-domain processing is less efficient. The UL-UNAS model presented in [68] systematically explores efficient convolutional structures and innovatively introduces Adaptive Parametric ReLU (APReLU) and Channel Transform Attention (cTFA) modules, significantly reducing model complexity while maintaining performance.

In the domain of multi-channel speech processing, U-Net exhibits excellent scalability. Ref. [69] employs a Wave-U-Net structure with a cross-channel attention mechanism. By independently encoding information from each channel and leveraging differences in time delay and power between target speech and interference signals to generate masks, it maintains high performance even in environments with strong reverberation and noise. TRUNet [70] further directly estimates filters from multi-channel input spectra, achieving end-to-end separation. This model integrates a spatial processing network with cross-microphone channel attention and a spectral-temporal processing network, capturing spatial diversity and spectral-temporal diversity, respectively. The causal MIMOU-net neural beamformer proposed in [71] deeply integrates a MISO U-net with a beamforming structure. By explicitly incorporating beamforming operations at the network output, it achieves efficient multi-channel speech enhancement without requiring explicit spatial features such as IPD.

A summary of the aforementioned content is provided in Table 5.

Table 5. Currently, advanced speech separation methods based on the U-Net architecture, only list the optimal results when evaluating multiple application scenarios. Among them, "-" indicates that the corresponding data cannot be found or the data is ambiguous.

Refs	Year	Description	Target	Dataset	Metrics	Model, Param. (M)	MACS (G/s)
[58]	2017	The U-Net architecture is proposed to address source separation tasks.	Transfer and Adaptive Refinement of the U-Net Architecture	iKala	NSDR Vocal: 11.094, NSDR Instrumental: 14.435, SIR Vocal: 23.960, SIR Instrumental: 21.832, SAR Vocal: 17.715, SAR Instrumental: 14.120	U-Net , -	-
[59]	2018	Wave U-Net, which is a one-dimensional temporal-domain improvement of the U-Net architecture.		Mixed dataset	Med.SDR 4.46	Wave-U-Net-M4, -	-
[60]	2018	Discusses performance differences of Wave-U-Net variants with different layers		VCTK	PESQ: 2.41, CSIG: 3.54, CBAK: 3.23, COVL: 2.97, SSNR: 9.87	Wave-U-Net (9-layer), -	-
[61]	2021	Incorporates bidirectional LSTM	Long-range dependency modeling	MusDB	-	Demucs, Baseline model size > 1014 MB	-
[62]	2022	Achieves collaborative modeling of global context and local detail-level temporal information		Mixed dataset	PESQ: 2.4501 (SNR [−5, 0]), 2.7472 (SNR [0, 5]), 2.9511 (SNR [5, 10])	UFormer, 9.46	-
[63]	2019	Introduces learnable local internal attention mask in skip connections	Optimization of skip connections	VCTK	PESQ: 2.53, CSIG: 3.77, CBAK: 3.12, COVL: 3.14, SSNR: 7.42	Wave-U-Net WITH ATTENTION (16 kHz, no aug), -	-
[64]	2022	Dual-branch attention module, extracts spatial and channel features in parallel		VCTK	PESQ: 2.84, CSIG: 4.14, CBAK: 3.47, COVL: 3.50, SSNR: –	DBAUNet, 0.66	-
[65]	2023	Proposes channel-spectrum attention mechanism		ICASSP 2022 DNS Challenge	SIG: 2.988, BAK: 3.974, OVRL: 2.713	U2Net, 8.947	-
[68]	2025	Explores efficient convolutional structures	Lightweight design and low-latency optimization	Mixed dataset	PESQ: 3.09	UL-UNAS, 0.17	0.03
[69]	2020	Adopts Wave-U-Net structure with cross-channel attention mechanism	Multi-channel	Synthetic datasets	SDRI: 18.032, STOI: 0.961, PER: 39.323	-	-
[70]	2021	Estimates filter from multi-channel input spectrum		Mixed dataset	SDRI: 9.38, SIRI: 12.87, PESQi: 0.22	TRUNER-MagPhase, 29	-
[71]	2021	Deeply integrates MISO U-net with beamforming structure		Mixed dataset	PESQ: 1.919, STOI: 0.857, E-TOI: 0.759, SI-SNR: 7.935	MIMO-U-net + BF + PF, 197	-

3.3. Progress in Speech Separation Methods Based on the TasNet Architecture

The Time-Domain Audio Separation Network (TasNet) [72] is one of the most successful deep learning methods in the field of speech separation in recent years. Its core advantage lies in performing end-to-end separation directly in the time domain without the need for short-time Fourier transform (STFT), thereby enabling the simultaneous utilization of both amplitude and phase information of sounds.

TasNet adopts an autoencoder architecture (As shown in Figure 6) and consists of three main components. Encoder learns to map input waveforms to high-dimensional feature representations. Separator generates masks corresponding to the number of target sound sources based on the encoded features. Decoder reconstructs the separated time-domain speech signals by performing a dot product operation between the masks and the features output by the encoder.

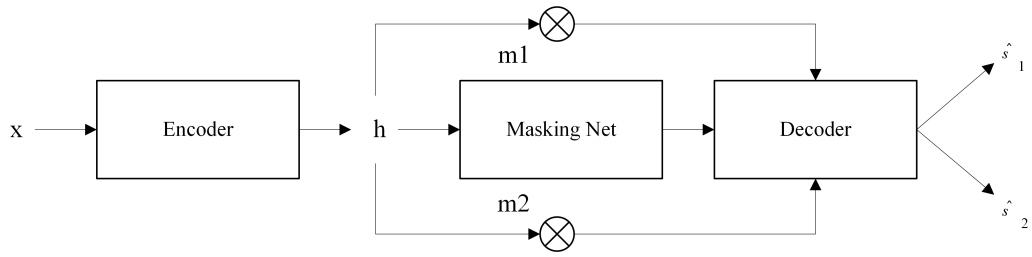


Figure 6. Autoencoder Architecture.

Heitkaemper et al. [73] conducted an in-depth analysis of TasNet. They found that in a noise-free single-channel scenario, its performance gains mainly stem from high temporal resolution and the design of the temporal loss function. However, in reverberant environments, using only the temporal loss function can also improve separation. Although TasNet outperforms traditional time–frequency domain methods in both causal and non-causal implementations, the long short-term memory (LSTM) network used in its separation module has obvious problems. The training complexity is high, and to achieve high temporal resolution, the encoder needs to process short input waveform segments, resulting in a long encoder output sequence, which increases the complexity of training based on LSTM networks.

To overcome these limitations, Luo et al. [74] proposed the fully convolutional time-domain audio separation network Conv-TasNet in 2019. Its core improvement is to replace the LSTM in the separation module with a time-convolutional network (TCN) and introduce depthwise separable convolution. TCN can effectively model long-term sequence dependencies, while depthwise separable convolution significantly reduces the model size.Conv-TasNet outperforms state-of-the-art time–frequency masking methods (e.g., IBM, IRM, WFM) in both objective distortion metrics and subjective perception evaluations on the two-speaker separation task. Additionally, its compact model size and low minimum latency make it an ideal solution for offline and real-time speech separation.

The success of Conv-TasNet has inspired extensive research. Kavalerov et al. [75] evaluated Conv-TasNet on speech/non-speech separation and general sound separation tasks, and systematically compared different masking network architectures and combined transformations. Deng et al. [76] integrated Conv-TasNet into a generative adversarial network (GAN) framework, proposing Conv-TasSAN. This model consists of a Conv-TasNet separator and an encoder-TCN-CNN discriminator, employing an adversarial architecture based on speech-level permutation invariance training (uPIT), with speech objective metrics as the discriminator target, enabling the separator to better learn the source signal distribution. Lee et al. [77] attempted to apply neural architecture search to Conv-TasNet (NAS-TasNet) to automatically search for optimal network structures.

The success of Conv-TasNet highlights the importance of the encoder-decoder structure. Researchers have begun exploring the use of easily understandable deterministic encoders (such as filter banks based on signal processing principles or auditory perception features) to replace fully network-learned encoders. For example, Ditter et al. [78] proposed replacing the learned encoder with a Gammatone filter bank, sparking discussions about learning vs. deterministic encoding.

Conv-TasNet itself still has a key limitation, it can only focus on information within fixed-length speech segments and struggles to integrate sentence-level context, especially when mixed audio contains long silences. To address this issue, Luo et al. [79] proposed a dual-path recurrent neural network (DPRNN-TasNet). This method divides the input sequence into overlapping chunks and applies RNNs on two paths (As shown in Figure 7), intra-chunk and inter-chunk. This design enables the network to effectively model long-range dependencies (integrating sentence-level information) while reducing the sequence length processed by each RNN to the square root of the original length, achieving sublinear complexity. DPRNN-TasNet achieves excellent separation performance but has high computational resource requirements and weak real-time performance. Studies also indicate that reducing the window size can yield more precise waveform encoding but further increases computational costs.

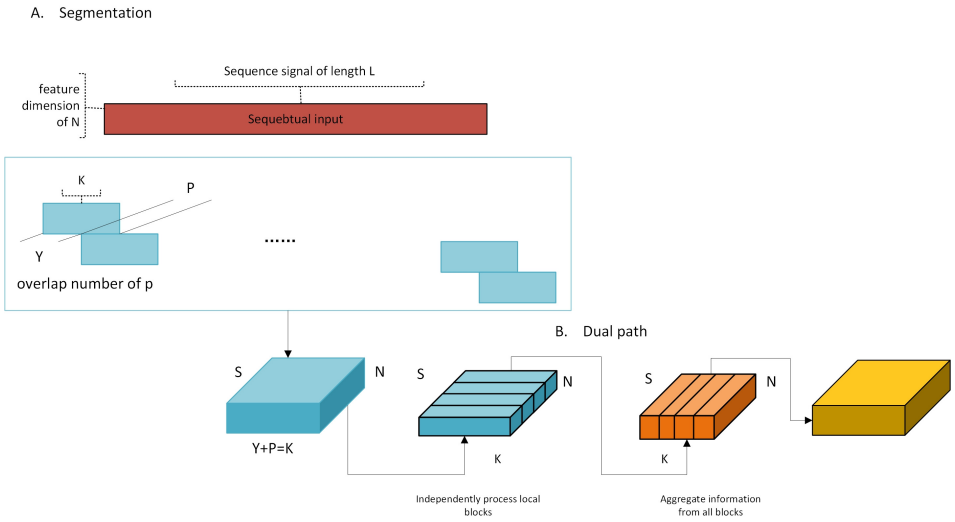


Figure 7. As shown in the figure, it demonstrates a dual-path architecture.

It is worth noting that subsequent time-domain separation models based on RNNs or Transformers have adopted similar dual- or multi-path architectures for reasons of performance and efficiency.

In order to address the high computational complexity of DPRNN and the challenge of modelling long sequences in Conv-TasNet, researchers have proposed new solutions. For example, Wang et al. [80] proposed the use of graph convolutional networks (GCNs) to capture the high-resolution details of waveforms, thereby alleviating the computational burden caused by the size of the window. Sato et al. [81] introduced state space modelling (SSM) into the Conv-TasNet architecture. SSM has been proven to effectively model long-term dependencies, thereby reducing the number of expansion convolution layers required and lowering model complexity. Additionally, by increasing the encoder window length and stride, computational costs are further reduced, and the performance loss can be compensated by overparameterizing the encoder. The paper [82] proposed an improved DConv-TasNet model to enhance generalisation and feature extraction capabilities. This model utilises a deep-expanded encoder/decoder and employs dilated convolutions to expand the receptive field. It also applies improved convolution blocks in the separation module to enhance channel-dimensional feature extraction. Work in [83] Wazir proposed Deep Speak Net, which

is based on an improved Conv-TasNet architecture. This model enhances gradient flow by improving the autoencoder structure (by introducing residual and skip connections), thereby stabilising training and improving separation quality and efficiency.

From the original TasNet to Conv-TasNet and its numerous variants (such as DPRNN-TasNet, Conv-TasSAN, Beam-TasNet [84]), time-domain end-to-end speech separation methods have continuously evolved in terms of performance, efficiency, and applicability. Current research focuses on optimizing long sequence modeling, reducing computational complexity, integrating prior knowledge (such as auditory features), and exploring new architectures (such as SSM and GCN) to drive the broader practical application of this technology.

A summary of the aforementioned content is provided in Table 6.

Table 6. Currently, advanced speech separation methods based on the TasNet architecture, only list the optimal results when evaluating multiple application scenarios. Among them, "-" indicates that the corresponding data cannot be found or the data is ambiguous; when "*" appears alone, it means there are multiple cases of the data, and the original literature should be consulted; when "*" follows a specific data point, it indicates that the data is sourced from other literatures.

Refs	Year	Innovation or Name of the Method	Experimental Dataset	SI-SNRi	SDRi	Model, Param. (M)	MACS (G/s)
[72]	2018	The information from the raw waveform is expressed in a learnable convolutional latent space.	WSJ0-2mix	10.8	11.1	TasNet-BLSTM, 23.6	-
[74]	2019	Fully Convolutional Time-Domain Audio Separation Network	WSJ0-2mix	15.3	15.6	Conv-TasNet-gLN, 5.1	* 10.5
[75]	2019	Different combinations of network architecture concealment	WSJ0-2mix	*	*	*	-
[76]	2020	Integrating Conv-TasNet into Generative Adversarial Networks	WSJ0-2mix	15.1	15.4	Conv-TasSAN, *	-
[77]	2022	Neural Architecture Search (NAS) Strategy Reinforcement	WSJ0-2mix	17.50	17.74	NAS-Tasnet-GD, 6.3	-
[79]	2020	Dual-path recurrent neural network	WSJ0-2mix	18.8	19.0	DPRNN-TasNet, 2.6	* 88.5
[78]	2022	Fully Convolutional Time-Domain Audio Separation Network	WSJ0-2mix	15.9	-	MP-GTF-128, 16.1	-
[82]	2024	Deeply expanded encoder/decoder	WSJ0-2mix	19.2	19.1	DConv-TasNet, -	-
[83]	2025	Autocoders were used to simultaneously employ residual and skip connections.	TIMIT	12.4	12.1	EncoderLINEAR, 1.8	-

3.4. Progress in Speech Separation Methods Based on the Transformer Architecture

The Transformer is a deep learning model architecture proposed by Vaswani et al. [85] in 2017, which fundamentally transformed the task mechanism of natural language processing (NLP). Its core innovation lies in the self-attention mechanism, which effectively models long-range dependencies and complex modalities in sequence data. This feature is also advantageous for tasks such as speaker recognition and speech processing, as it can capture subtle contextual differences and complex interrelationships in acoustic signals. The general architecture of the Transformer is depicted in Figure 8.

The Transformer was quickly introduced into the field of speech processing by researchers. Gulati et al. [86] proposed the Conformer architecture based on Transformer, which combines a self-attention layer for global context modelling with a convolutional module for capturing local correlations. This hybrid approach has achieved significant success in various speech processing tasks.

Focusing on the field of speech separation, the use of the Transformer architecture for speech separation has become a trend. Chen et al. [87] proposed a dual-path Transformer network (DPTNet) for end-to-end single-channel speech separation.DPTNet integrates recurrent neural networks (RNNs) into the Transformer framework, enabling direct context-aware modelling of speech sequences without relying on position encoding to learn sequence information. SepFormer, proposed by Subakan et al. [88], is a model specifically designed for speech separation. Inspired by the DPRNN dual-scale framework, it replaces RNNs with Transformers.SepFormer employs a multi-scale learning strategy to effectively capture short-term and long-term dependencies in speech signals, achieving high efficiency and excellent performance [89]. Subsequently, Subakan et al. [90] further explored this

framework on more realistic and challenging datasets. In this context, they investigated alternative attention mechanisms such as Longformer, Linformer, and Reformer.

A major challenge in speech separation is handling an unknown number of speakers. Inspired by SepFormer, Chetupalli et al. [91] addressed the problem of speaker separation in single-channel mixed speech with an unknown number of speakers. They integrated an Encoder-Decoder Attractor (EDA) module into the separation network to tackle this issue. Lee et al. [92] further extended this method and proposed SepTDA. It combines a three-path approach with LSTM-enhanced self-attention blocks, claiming to more effectively capture local and global contextual information. Its core is the introduction of a TDA (Transformer-based Decoder Attractor) computation module, which extends EDA based on the Transformer architecture.

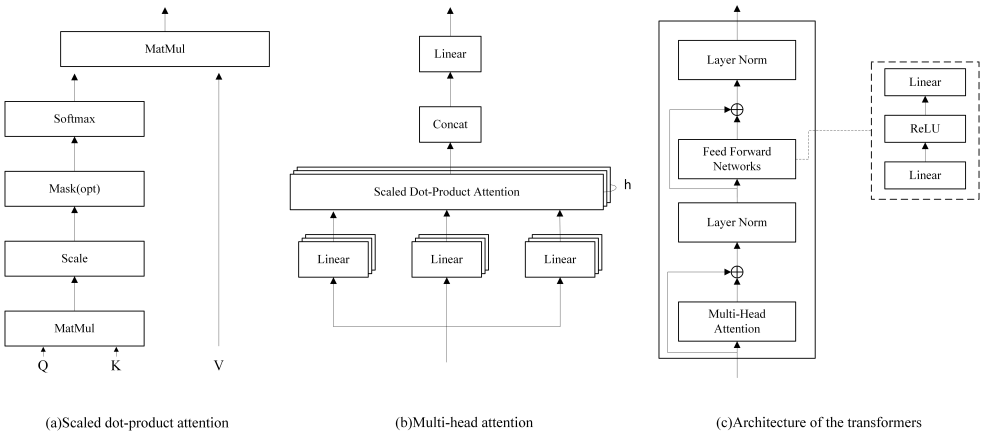


Figure 8. Transformer main framework.

Researchers continue to explore hybrid models that combine the advantages of Transformers and other architectures to improve performance or efficiency. Zhao and Ma [93] proposed MossFormer, which incorporates gated single-head attention and convolution-enhanced joint self-attention mechanisms to more effectively model long speech sequences. It aims to push the performance limits of single-channel speech separation. Building upon MossFormer, Zhao et al. [94] proposed MossFormer2. It abandons traditional recurrent connections and adopts a recurrent module based on feed-forward sequential memory networks (FSMN) to capture recurrent patterns without using recurrent connections. Shin et al. [95] proposed the Separation Reconstruction Transformer (SepReformer) to achieve more efficient time-domain separation. Using an asymmetric encoder-decoder structure, it employs efficient global and local Transformer unit modules based on the ESSD framework. The encoder processes features at different temporal resolutions, with intermediate features used for skip connections; the decoder progressively reconstructs fine-grained information from time-bottleneck features, focusing on consonant discrimination features prone to loss in separated speech. The authors demonstrate that the SepRe method can be integrated into Conv-TasNet and the original SepFormer separator with good results. Wang et al. [96] further proposed a lightweight dual-path Conformer network. The encoder part uses adaptive multi-scale depth-separable convolutions, enabling the model to dynamically adjust the convolution kernel size according to the input features to obtain multi-scale features; the separator network part uses depth-separable kernel convolution modules. Wang et al. [97] proposed a hybrid dual-path network combining Conformer and Transformer architectures for separating vocals in the waveform domain. The network aims to effectively capture the complex temporal dependencies of signals, overcoming the limitations of Sepformer in capturing local features of signals.

To reduce computational complexity and training costs, Liu et al. [98] proposed a short sequence encoder-decoder network (ESEDNet). Its encoder consists of multiple convolution and downsampling layers, significantly reducing the length of high-resolution sequences; the decoder utilises encoded features to reconstruct the target speaker's fine-grained speech sequence. By combining with MTRFormer, ESEDNet can efficiently obtain separation masks on short encoded feature sequences.

A summary of the aforementioned content is provided in Table 7.

Table 7. Currently, advanced speech separation methods based on the Transformer, only list the optimal results when evaluating multiple application scenarios. Among them, "-" indicates that the corresponding data cannot be found or the data is ambiguous; when "*" follows a specific data point, it indicates that the data is sourced from other literatures.

Refs.	Year	Innovation or Name of the Method	Experimental Dataset	SI-SNRi	SDRi	Model, Param. (M)	MACS (G/s)
[87]	2020	Dual-path Transformer network	WSJ0-2mix Libri-2mix	20.2 18.2	20.6 18.4	DPTNet, 2.69	* 102.5
[88]	2021	Proposed the SepFormer model	WSJ0-2mix	20.4	20.5	SepFormer, 26	59.5
[89]	2022	Incorporating super-resolution techniques after the decoder to address information loss caused by downsampling.	WSJ0-2mix	22.0	22.1	SFSRNet, 59	-
[91]	2023	Integrating encoder-decoder attractors to handle an unknown number of speakers.	WSJ0-2mix	21.2	21.4	SepEDA2 *, 12.5	81.0
[92]	2024	The TDA computing module is designed to effectively handle an unspecified number of speakers.	WSJ0-2mix	-	-	SepTDA2, 12.5	-
[93]	2023	The self-attention-based module tends to emphasize longer-range, coarser dependencies.	WSJ0-2mix	22.8	-	Mossformer(S), 10.8	-
[94]	2024	Introducing the innovative RNN-free recurrent module, FSMN.	WSJ0-2mix	24.1	-	Mossformer2, -	* 55.7
[95]	2024	Adopts an asymmetric encoder-decoder architecture.	WSJ0-2mix	24.2 22.4	24.4 22.6	SepReformer-L, 59.4 SepReformer-T, 3.5	155.5 10.4
[98]	2025	Combining a short-sequence encoder-decoder framework with a multi-temporal resolution Transformer-based separation network.	Libri2Mix	13.24	14.08	ESEDNet, 2.31	7.13

3.5. Progress in Speech Separation Methods Based on the Mamba Architecture

State Space Models (SSMs) have emerged as a unique sequence modeling framework and have achieved significant progress in the field of speech separation in recent years.

Gu and Dao [99] proposed the Mamba model based on a novel selective state space mechanism.

Mamba is a novel deep learning architecture specifically designed for sequence modelling. Unlike earlier SSMs, the core innovation of Mamba lies in its input-dependent selection mechanism. This mechanism significantly improves sequence modelling performance while maintaining linear computational complexity across sequence lengths. The emergence of Mamba as a general-purpose sequence modelling backbone offers an opportunity to develop more computationally efficient and resource-efficient solutions for speech separation, potentially achieving performance on par with current state-of-the-art Transformer-based models. Figure 9 illustrates the key architectural components of Mamba.

Mamba's key advantages lie in its outstanding computational and memory efficiency, especially when processing long sequences, where it has a clear advantage over Transformer-based attention mechanisms. Additionally, in cyclic or real-time processing scenarios, Mamba requires significantly less internal state to maintain. Based on the above advantages of Mamba, researchers have quickly applied it to speech separation tasks and proposed several innovative architectures.

Building upon the successful TF-GridNet [100,101] architecture as Figure 10.

SP-Mamba [102] replaces the bidirectional long short-term memory (BiLSTM) component with bidirectional Mamba blocks. However, SP-Mamba retains the self-attention mechanism of Transformer-based models, resulting in high overall computational overhead. Li et al. [103] proposed a U-Net-based model called SepMamba, whose core consists of bidirectional Mamba layers. SepMamba outperforms mainstream models with similar parameter counts (including

Transformer-based models) in terms of performance, while demonstrating significant computational advantages in terms of multiplication-addition operations (MACs), peak memory usage, and inference latency. DPMamba [104] adopts a dual-path architecture to model the local (short-term) and global (long-term) features of speech sequences separately. Its core lies in combining bidirectional Mamba modules to process sequences in both forward and backwards directions, thereby fully leveraging contextual information. Speech Conv-Mamba [105] embeds the Mamba module into a U-net, combining temporal dilated convolution (Temporal Dilated Convolution) and Mamba to construct a separation network. Although its separation performance is slightly inferior to SepFormer, it has obvious advantages in model efficiency (computational complexity and parameter count).

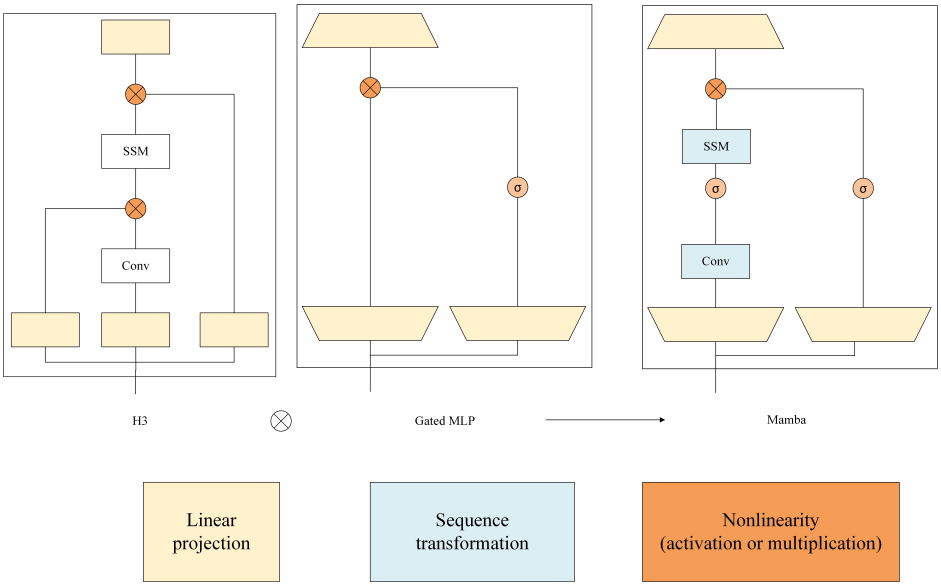


Figure 9. Mamba module

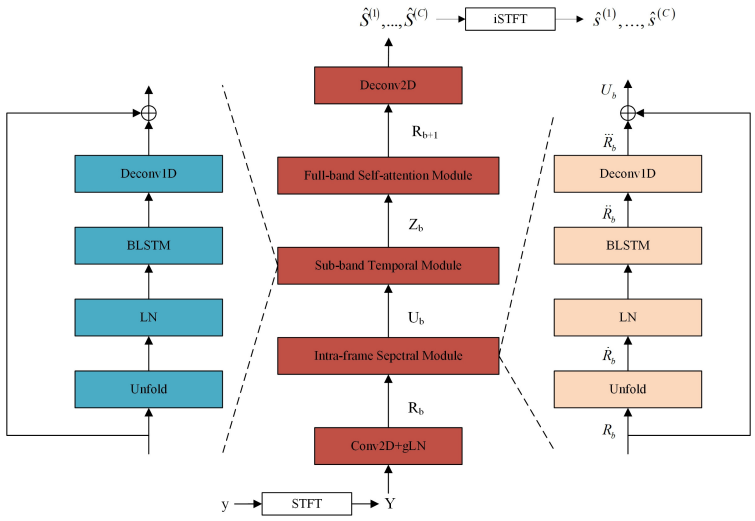
Jiang et al. [106] systematically investigated the performance and efficiency of Mamba across multiple speech-related tasks. They specifically noted that when the speech duration exceeds a certain threshold, Mamba demonstrates significant improvements in both memory usage and processing speed. This threshold is determined by the resolution of speech tokens. Their findings indicate that Mamba shows substantial advantages in high-resolution tasks such as speech separation, while offering limited or negligible benefits in low-resolution tasks like automatic speech recognition (ASR). In the field of speech separation, the researchers proposed Mamba-TasNet, a architecture inspired by Conv-TasNet. This model processes encoded features of mixed signals and ultimately generates masks corresponding to S sound sources. Notably, experimental results demonstrate that within the Mamba framework, a dual-path architecture is unnecessary—excellent performance can be achieved using a single-path design alone.

DeFT-Mamba [107] focuses on general multi-channel sound separation and polyphonic audio classification. It combines a dense frequency–time (DeFT) attention network with Mamba. Specifically, it uses gated convolutional blocks to capture local spatiotemporal relationships, a position-mixed Mamba to capture global spatiotemporal relationships, and introduces a classification-based source counting method to identify the number of sound sources present in the mixed signal.

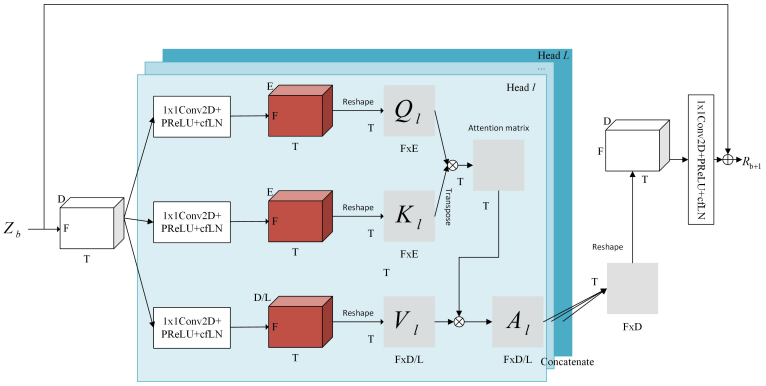
Mamba and its derivative models demonstrate strong potential in speech separation tasks. These models not only challenge or even surpass Transformer-based state-of-the-art (SOTA) models (such as SepFormer) in terms of performance but also typically exhibit significant

advantages in computational efficiency (MACs, memory, latency) and model size, providing a new technical approach for efficient speech separation in resource-constrained scenarios.

A summary of the aforementioned content is provided in Table 8.



(**a**) main architecture



(**b**) Full-band Self-attention Module

Figure 10. TF-GridNet architecture.

Table 8. Currently, advanced speech separation methods based on the Mamba, only list the optimal results when evaluating multiple application scenarios. Among them, "-" indicates that the corresponding data cannot be found or the data is ambiguous; when "*" follows a specific data point, it indicates that the data is sourced from other literatures.

Refs.	Year	Innovation or Name of the Method	Experimental Dataset	SI-SNRi	SDRi	Model, Param. (M)	MACS (G/s)
[100]	2023	A Novel Multipath Deep Neural Network Operating in the Time–Frequency (T-F) Domain	WSJ0-2mix WHAM! Libri-2mix	22.8 16.9 19.8	23.1 17.2 20.1	TF-GridNet, 14.43	* 445.56
[102]	2024	Based on the TF-GridNet architecture, the BLSTM module is replaced with a bidirectional Mamba module.	WSJ0-2mix WHAM! Libri-2mix	22.5 17.4 19.9	22.7 17.6 20.4	Spmamba, 6.14	238.69
[103]	2025	Integrate Mamba layers into the U-Net architecture to learn the multi-scale structure in audio.	WSJ0-2mix	22.7	22.9	SepMamba(S), 7.2	12.46
[104]	2025	Replace Transformer with Mamba	WSJ0-2mix	23.4	23.6	DPMamba(L), 59.8	-
[105]	2025	Embed Mamba into the U-Net architecture.	Mixed dataset	15.01	15.84	Speech Conv-Mamba, 2.4	8.87
[106]	2025	Adhering to the Conv-TasNet architecture Mamba	WSJ0-2mix	22.4	22.6	Mamba-TasNet(M), 15.6	-

3.6. Summary

We have created a visual summary of the aforementioned deep learning approaches based on U-Net, TasNet, Transformers, and Mamba to highlight the relationships among different methods, as illustrated in the Figure 11.

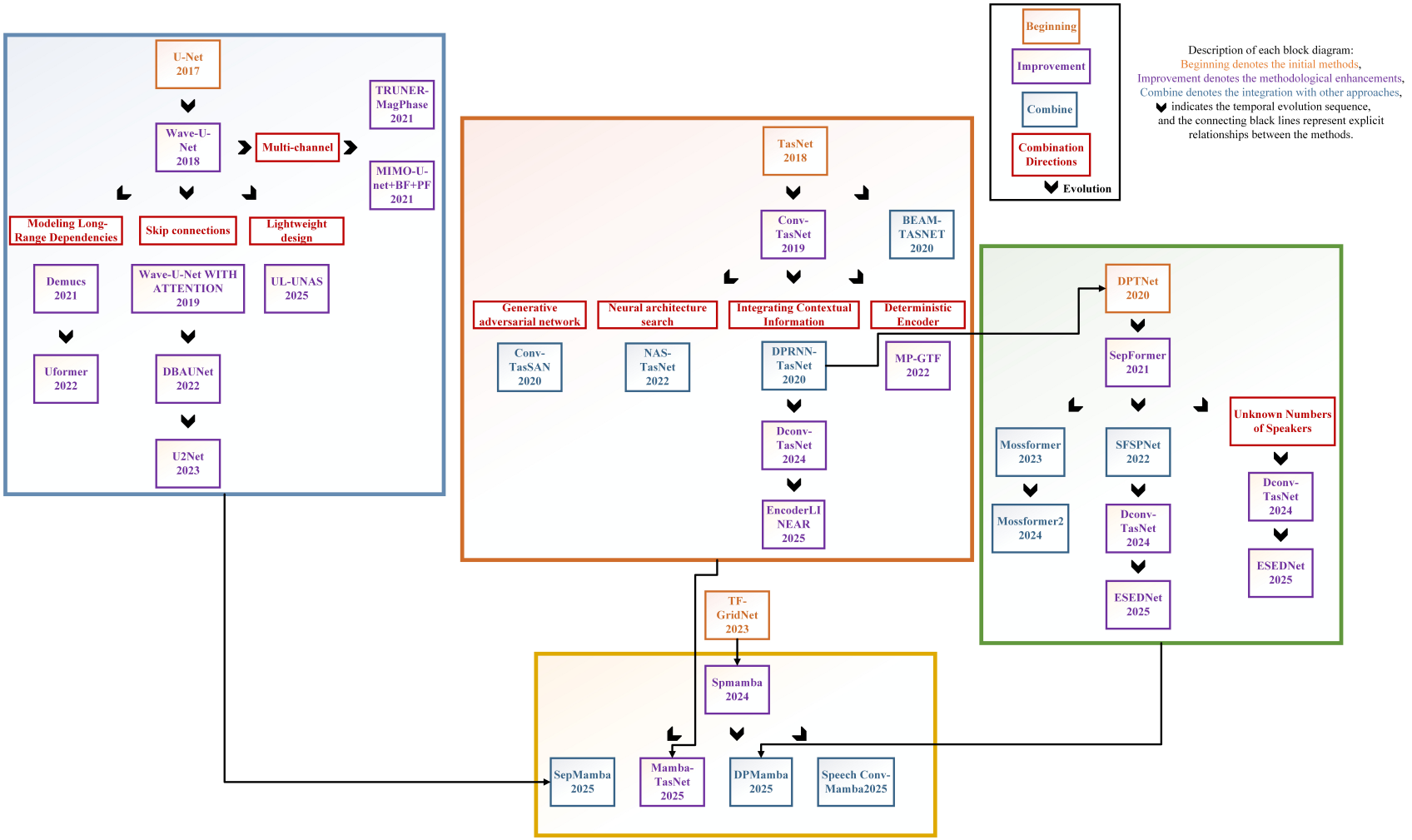


Figure 11. A schematic diagram illustrating the developmental trajectory and interconnections among different methods.

We posit that this series of models represents the current evolutionary trajectory of speech separation research. Understanding these models can assist readers in establishing a foundational comprehension of the field.

4. Discussion

U-net, TasNet, Transformer, and Mamba have achieved remarkable success in speech separation tasks.

We believe that this series of models represents the main advancements in current single-channel supervised speech separation research. Understanding these models can help readers establish a foundational understanding of this field. This section will systematically elucidate their strengths and limitations, followed by corresponding critical discussions.

4.1. Foundation

Speech signals are typical time-series data. In traditional speech separation methods, manual feature extraction often relies on signal processing techniques such as the Fourier transform. In contrast, deep learning-based models can learn the complex nonlinear mapping from mixed speech to target speech in an end-to-end manner, thereby significantly outperforming traditional methods in terms of performance. The core architectures of these models are predominantly built upon recurrent neural networks (RNNs), convolutional neural networks (CNNs), or their improved variants. As described in Section 3, current mainstream approaches generally construct a universal "encoder-separator-decoder" pipeline by integrating and optimizing various network components: the encoder is responsible for feature extraction and downsampling, the separator performs the discrimination and separation of speech features based on this, and the decoder handles signal reconstruction or upsampling. Through this structure, the model can automatically learn highly discriminative intermediate representations from the mixed speech and use them to reconstruct the target speech with high quality.

U-Net and TasNet, as representative models in the early stages of speech separation, hold pioneering and pivotal roles, and their contributions deserve renewed emphasis. The skip-connection structure introduced by U-Net is still widely utilized today to effectively integrate multi-scale contextual information. Meanwhile, TasNet's approach of directly mapping raw waveforms into a feature space has profoundly influenced the architecture of many contemporary speech separation models. The subsequent introduction of the Transformer mechanism can largely be seen as enhancing and optimizing key components built upon these two foundations. Thanks to its powerful global modeling capability and self-attention mechanism, the Transformer has demonstrated outstanding performance in sequence tasks such as speech separation—playing a critical role in strengthening contextual relationships across time and improving discriminative ability at the feature level.

Furthermore, Mamba was proposed based on a selective state space model (SSM). This model, along with other supplementary methods, primarily aims to address the quadratic complexity ($O(N^2)$) of the self-attention mechanism in Transformers or to improve the original context-aware modules. Mamba is capable of efficiently modeling extremely long-range dependencies, with computational complexity and memory usage growing only linearly with sequence length ($O(N)$). This characteristic is particularly advantageous for processing long-duration speech signals.

4.2. Critical Analytical Perspective

4.2.1. The Inherent Limitations of Model

We posit that the fundamental limitations of current speech separation models stem from the inherent constraints of their basic architectural components—namely, the intrinsic limitations of RNNs, CNNs, and Transformers themselves (Mamba, as an emerging model, has limitations that still require systematic evaluation [106]). RNN-based models require the sequential transmission of information through multiple intermediate states, leading to a time complexity that increases linearly with sequence length and difficulty in parallel computation due to state dependencies. Although various improved RNN variants have been proposed, their overall performance still lags significantly behind that of the Transformer.

CNN-based models, on the other hand, are constrained by their limited receptive fields. Even with enhanced structures such as dilated convolutions, they struggle to adequately capture long-range dependencies in audio signals while maintaining high sensitivity to local details. Although the Transformer exhibits powerful modeling capabilities, its core self-attention mechanism suffers from quadratic complexity ($O(N^2)$) with respect to sequence length, and is prone to insufficient video memory, limiting its scalability in long-sequence tasks. In recent years, numerous methods have been proposed to reduce the computational complexity of the Transformer, but such approaches often introduce information loss during simplification. Achieving a better balance between efficiency and performance remains a critical direction for further exploration in the field of speech separation.

Furthermore, while hybrid architectures that integrate multiple basic models can combine their respective advantages, they also introduce issues such as excessive parameter scale and high structural sensitivity. Inappropriate design of a single layer in such models may lead to overfitting or gradient anomalies, further complicating model optimization and stable training.

4.2.2. Computational Costs, and Performance Efficiency of the Model

As evidenced by the survey in Section 3, unless a task explicitly requires lightweight design, most models tend to employ deeper network architectures and larger parameter counts in their design and performance comparisons to achieve superior separation performance—this aligns with the fundamental characteristic of deep learning models where performance generally improves with increasing capacity. However, it is important to note that some models with relatively low parameter counts can require extremely high MACs (Multiply-Accumulate Operations) in practical computation.

This phenomenon is particularly prevalent in models based on the Transformer's self-attention mechanism (or similar structures). Due to the quadratic complexity ($O(N^2)$) of the core attention mechanism with respect to sequence length, even models with modest parameter counts can suffer from significantly impacted training and inference speeds as input sequence length increases, forming a critical bottleneck in practical deployment.

4.2.3. Deployment Challenges and Hardware Limitations

When analyzing the hardware compatibility and system integration constraints of speech separation models, "causal speech separation" constitutes a core concept that must be clearly distinguished. Most current research prioritizes performance-oriented models belonging to the non-causal category, which can utilize complete contextual information including future frames when processing the current speech segment. It is noteworthy that many studies no longer systematically evaluate the performance of their causal variants, resulting in a relative scarcity of comparative data in this regard.

Causal speech separation imposes strict temporal constraints during modeling: when processing the signal at any given moment, the model can only utilize information from

the current and past instances, completely excluding any influence from future frames. The fundamental objective of this approach is to achieve extremely low processing latency while maintaining reasonable separation performance, thereby meeting the practical requirements of scenarios with stringent real-time demands, such as real-time communication and online processing.

To realize causal separation, specially designed causal architectures—such as causal recurrent neural networks (RNNs) or causal convolutional networks—are typically employed. These ensure the correct propagation of temporal dependencies at the computational architecture level. For further discussion on the practical challenges related to hardware adaptation, inference efficiency, and system integration faced by causal models during deployment, it is recommended to consult relevant literature such as [108,109].

5. Future Directions

Deep neural network-based speech separation technology has achieved excellent performance [1], but the following issues remain.

Future Directions Robustness challenges in complex acoustic environments and limitations in data-dependent generalisation capabilities. Currently, these models have reached a certain bottleneck in their performance on specific datasets (WSJ0-2mix), but their performance on other datasets is not yet at an excellent level. This reveals that the current datasets do not fully reflect real-world speech interaction problems.

Difficulties in handling multi-speaker scenarios, when multiple people (three or more) speak simultaneously, the system struggles to accurately segment speech boundaries and assign speaker identities, even with the most advanced deep learning models, resulting in insufficient separation and matching rates.

Computational efficiency and real-time performance bottlenecks. Real-time interaction requires end-to-end latency below 500 ms, which current models struggle to achieve.

However, with further research, this paper believes that the following speech separation methods hold significant potential for addressing these challenges in the future.

5.1. Challenges in Time–Frequency Domain Feature Fusion

Compared to time-domain methods, T-F domain separation methods typically exhibit stronger robustness and are highly correlated with the acoustic structure of speech. The introduction of TF-GridNet [100] has reignited researchers' interest in T-F domain separation models. As a complex spectral mapping network for monaural and multispeaker speech separation, TF-GridNet employs longer STFT window lengths and offsets than time-domain models, demonstrating outstanding speech separation performance. However, the inherent characteristics of the traditional short-time Fourier transform (STFT) as a non-learnable general-purpose signal transformation tool limit the optimisation potential for speech separation. To overcome this limitation, it is typically necessary to introduce a learnable auxiliary encoder after the STFT and design specialised T-F domain separation methods.

Yang et al. [110] proposed the T-F Domain Path Scanning Network (TFPSNet), which follows this approach. The encoder converts the mixed waveform into T-F features, the separation layer predicts the mask vectors for each source signal, and the decoder reconstructs the source waveforms via inverse STFT (iSTFT). For the dual-path structure tailored to TF-domain modelling, Saijo [111] proposed TF-Locoformer, which replaces RNN with Transformer for global modelling and uses convolution for local modelling, while introducing a novel normalisation layer. The core idea of the dual-path network is to divide long sequences into segments and iteratively learn local features within segments and global features across segments. However, the fixed segment length results in a fixed receptive field, limiting the model's flexibility. To address this issue, Qian [112] proposed

the Multi-Scale Time Delay Sampling (MTDS) method, which gradually expands the receptive field through time delay sampling, effectively integrating multi-scale information from fine to coarse in sequence feature learning. Zhai [113] proposed a triple-path RNN model by integrating temporal and frequency domain features to improve separation performance. Wang [114] designed a dual-domain joint encoding module that enhances the feature encoding capability of the temporal domain network by combining temporal and frequency information.

With the recent development of time–frequency domain research, frequency domain features directly reflect speech energy, while time domain features are learned from the latent embedding space. Research combining the advantages of both has garnered widespread attention.

5.2. A Study of Exceptions in Target Speaker Extraction

Target Speaker Extraction (TSE) is a special case of speech separation, aiming to separate the voice of a specific speaker from mixed audio using reference speech, particularly suitable for scenarios where the number of speakers is unknown. How to extend high-performance TSE systems into general-purpose speech separation (SS) systems has attracted much attention [115].

Liu et al. [116] proposed the X-SepFormer model to address the speaker confusion problem. This model optimizes the reconstruction capability of extracted speech and effectively reduces confusion. To address the issue of mismatched speaker embeddings, Zeng et al. [117] proposed the SEF-Net model for target speaker extraction. This model uses time-domain, speaker-independent embeddings and offers a new solution that does not rely on pre-trained embeddings. SEF-Net uses a dual-weight shared convolutional encoder to process mixed speech and reference speech separately, and employs a cross-multihair attention (CMHA) mechanism in the Transformer decoder to implicitly utilise speaker information from the reference speech. However, its limitation lies in the requirement that the reference speech be of the same length as the mixed speech. Based on SEF-Net, Zeng [118] further proposed a general speaker-embedding-free target speaker extraction framework, USEF-TSE.USEF-TSE employs a unified encoder to process both types of speech, using the mixed speech encoding to query the reference speech encoding, and enhances the target speech features through the CMHA module.

5.3. Multi-Channel Fusion and Spatial Information Utilization

Single-channel speech separation has always faced a fundamental challenge: severe insufficiency of information dimensions. When all sound source signals are mixed into a single observed signal, a significant amount of information about the original sound sources and their spatial locations is irreversibly lost during the mixing process. Estimating multiple unknown source signals from a single observed signal is a highly underdetermined and ill-posed inverse problem. The core of multi-channel speech separation lies in effectively utilizing the spatial information captured by microphone arrays, which is the key to overcoming the limitations of single-channel separation and significantly improving speech separation performance.

Quan and Li [119] designed a neural network called SpatialNet for multi-channel joint speech separation, denoising, and reverberation removal. SpatialNet fully utilizes narrowband and cross-band spatial information, and is claimed to have minimal spectral generalization issues, performing exceptionally well in cross-language tasks. Kalkhorani and Wang [120] conducted an in-depth study on the fundamental reasons behind the performance differences between TF-GridNet and SpatialNet (especially in single-channel separation and long-speech scenarios). They proposed a new DNN architecture,

TF-CrossNet, based on their findings. They pointed out that the differences primarily stem from two aspects: the self-attention module in TF-GridNet implements global modeling, while SpatialNet processes each frequency point independently, and the RNNs (e.g., LSTM) used in TF-GridNet can implicitly capture spatial information. To address this, TF-CrossNet employs global multi-head self-attention to capture cross-frequency and cross-embedding correlations, and introduces Random Block Positional Encoding (RBPE) for long sequences to tackle the out-of-distribution generalization issues associated with conventional position encoding.

Traditional multi-channel methods typically extract microphone inter-phase differences (IPD) and combine them with amplitude spectra to provide spatial cues. Shin et al. [121] proposed TF-CorrNet for multi-channel continuous speech separation (including de-reverberation and denoising). TF-CorrNet directly utilizes microphone inter-correlations to capture spatial and spectral context, simplifying the learning process and eliminating the need for explicit connections between different features or reliance on raw signal components.

5.4. Speech Separation Based on Diffusion Models

Score-based generative modelling (SGM), also known as diffusion-based modelling, involves sampling by defining a forward process that progressively distorts target samples into noise and an inverse generative process based on a score function (the gradient of the log probability density). Crucially, the function is typically unknown but can be approximated using a DNN with simple training strategies.

Lu et al. [122] proposed a conditional diffusion probabilistic model for speech enhancement. Scheibler et al. [123] introduced DiffSep, a novel single-channel source separation method based on a denoising diffusion model and fractional matching of stochastic differential equations (SDEs).DiffSep designs a continuous-time diffusion mixing process and trains a neural network to approximate its marginal probability's fractional function, employing an improved training strategy to address model mismatch and source permutation ambiguity. Dong et al. [124] proposed EDSep, an improved SDE-based method whose training and sampling processes are customised for different signal features, aiming to enhance separation efficiency and quality.

5.5. Multimodal Fusion: Audio-Visual Speech Separation

Although existing methods primarily rely on audio signals, humans rely on both auditory and visual cues to understand speech in complex acoustic environments. Inspired by this, audiovisual (AV) speech separation aims to integrate visual information such as facial attributes and actions to improve separation performance [125]. Audio and visual modalities are complementary, and their fusion can provide stronger robustness than pure audio methods, especially in noisy environments, and can assist in resolving alignment ambiguities.

Early AV separation studies primarily relied on T-F masking [126,127], followed by a shift toward time-domain models [128,129]. Time-domain models employ a learnable encoder-decoder architecture, achieving success in speaker separation and target speaker extraction. These models typically have more parameters and operate on shorter signal windows, resulting in low latency but high computational costs.

Kalkhorani et al. [130] introduced the AVTFGridNet model, an audiovisual version of the TF-GridNet framework. This model uses cross-attention for audiovisual fusion to capture complementary cues and employs a signal-to-noise ratio (SNR) scheduling strategy to gradually increase the contribution of the visual stream during training. Considering the differences between speech and noise interference, Pan et al. [131] further proposed

the scene-aware model SAV-GridNet.SAV-GridNet is a cascaded model. First, a classifier identifies the interference type (speech or noise), then a dedicated AV-GridNet expert model trained for that scenario is applied.

Work in Lee [132] proposed an audiovisual speech separation framework, AVDiffuSS, based on a diffusion model, which fuses audiovisual modal information during speech generation. To address the challenge of selecting fusion layers due to the difference in encoding speeds between audio and video modalities (audio typically converges faster), the paper [133] proposed the Encoding Speed Synchronisation Network (EPS-Net). EPS-Net allows for independent encoding of modalities and establishes communication between corresponding encoding layers to gradually synchronise encoding speeds, achieving progressive information fusion while preserving modal uniqueness.

5.6. Utilization of Spatial Separability and Context Awareness

Utilizing spatial characteristics and contextual information to address speech separation is no longer novel, but this paper reiterates its importance.

Traditional signal processing often employs a two-stage strategy of "spatial alignment followed by filtering" to address acoustic problems. Inspired by this classic design approach, Lee et al. [134] proposed an alignment and filtering network (AFnet) specifically designed for deep learning-based speech processing tasks. The core innovation of AFnet lies in decoupling the complex noise reduction problem into two more manageable subtasks. First, the relative transfer function (RTF) is used to encode key spatial information and perform spatial alignment on the signal. This step aims to more accurately locate and focus on the target sound source. Subsequently, filtering is performed based on the spatial alignment. The goal of this design is to more effectively utilize the separability of sound sources in the spatial domain.

On the other hand, the human auditory system demonstrates remarkable capabilities in noisy "cocktail party" scenarios, actively utilizing contextual information (such as semantics and prosody) of target speech for selective listening. Inspired by this biological mechanism, Qian et al. [135] explored the potential application of contextual information in computational models. They proposed a novel neural network architecture whose key lies in automatically learning and extracting effective contextual embeddings from mixed speech signals themselves. These embeddings encode potentially useful information in mixed speech (which may include speaker identity, speech content fragments, and other contextual cues).

5.7. Lightweight Miniaturization While Maintaining Performance

Developing lightweight models with small dimensions that can run in real time is an important research direction, involving innovations in model structure and compression techniques.

Subband/multiband networks process different frequency bands using independent or shared subnetworks to reduce parameters [136]. Luo et al. [137] proposed the GroupComm method to design ultra-lightweight models, where large feature vectors are split into groups, and inter-group modules are applied to capture cross-group dependencies. When applied to DPRNN-TasNet, the performance was comparable, but the model size was reduced by 23.5 times. Tan [138] utilised selective mutual learning (SML), where two networks are trained simultaneously and learn through supervision and mutual knowledge selection, achieving good performance with only 0.9 million parameters. Elminshawi et al. [139] proposed Slim-TasNet, whose computational graph supports dynamic inference via different subsets, enabling adaptive trade-offs between performance and efficiency during inference.

The paper [140] proposed Grouped Time Convolutional Recurrent Networks (GTCRN), which simplifies DPCRN using grouping strategies and enhances performance through subband feature extraction and temporal recurrent attention modules. Yang et al. [141] introduced the ultra-lightweight real-time speech enhancement model FSPEN, which extracts global and local features using full-band and subband structures and enhances modelling capabilities through frame-to-frame path expansion.

In terms of model compression, Li et al. [142] proposed SepPrune, a structured pruning framework specifically designed for compressed deep speech separation models. This framework first analyses the model's computational structure to identify bottleneck layers, then introduces a differentiable masking strategy to achieve gradient-driven channel selection, and finally prunes redundant channels and fine-tunes the remaining parameters to restore performance.

6. Conclusions

This paper provides a systematic review of deep neural network-based speech separation methods, with a focused examination of the developmental trajectory in single-channel speech separation tasks, aiming to offer valuable reference for researchers to comprehend the current research landscape and future directions.

This study first categorizes mainstream model architectures, then delves into the strengths and limitations of various approaches, and finally analyzes critical issues including causal modeling and computational efficiency. Through a comprehensive analysis of core advancements and trends in single-channel speech separation, this survey aims to present researchers with a clear technological panorama while providing constructive insights for the further development of deep neural networks in speech separation and related tasks.

Author Contributions: Conceptualization, Z.W.; Formal analysis, Z.W.; Data curation, Z.W.; Writing—original draft preparation, Z.W.; Writing—review and editing, Z.W. and Z.L.; Supervision, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Exploration and Practice of the Path to Improve the Quality of Master's Degree Cultivation of Electronic Information Students Empowered by Numerical Intelligence JG202405, in part by the Sichuan Science and Technology Program under Grant 2025YFHZ0006, in part by the Scientific Research and Innovation Team Program of Sichuan University of Science and Engineering under Grant SUSE652A011.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: Not applicable

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ochieng, P. Deep neural network techniques for monaural speech enhancement and separation: State of the art analysis. *Artif. Intell. Rev.* **2023**, *56*, 3651–3703. [CrossRef]
2. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards end-to-end speech synthesis. *arXiv* **2017**, arXiv:1703.10135.
3. Nayeem, M.; Tabrej, M.S.; Deb, K.J.; Goswami, S.; Hakim, M.A. Automatic Speech Recognition in the Modern Era: Architectures, Training, and Evaluation. *arXiv* **2025**, arXiv:2510.12827. [CrossRef]
4. Cherry, E.C.; Taylor, W. Some further experiments upon the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* **1954**, *26*, 554–559. [CrossRef]
5. Li, K.; Chen, G.; Sang, W.; Luo, Y.; Chen, Z.; Wang, S.; He, S.; Wang, Z.Q.; Li, A.; Wu, Z.; et al. Advances in speech separation: Techniques, challenges, and future trends. *arXiv* **2025**, arXiv:2508.10830. [CrossRef]

6. Rafii, Z.; Liutkus, A.; Stöter, F.R.; Mimilakis, S.I.; FitzGerald, D.; Pardo, B. An overview of lead and accompaniment separation in music. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1307–1335. [CrossRef]
7. Wang, D.; Chen, J. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1702–1726. [CrossRef] [PubMed]
8. Ansari, S.; Alatrany, A.S.; Alnajjar, K.A.; Khater, T.; Mahmoud, S.; Al-Jumeily, D.; Hussain, A.J. A survey of artificial intelligence approaches in blind source separation. *Neurocomputing* **2023**, *561*, 126895. [CrossRef]
9. He, P.; She, T.; Li, W.; Yuan, W. Single channel blind source separation on the instantaneous mixed signal of multiple dynamic sources. *Mech. Syst. Signal Process.* **2018**, *113*, 22–35. [CrossRef]
10. Drude, L.; Hasenklever, D.; Haeb-Umbach, R. Unsupervised training of a deep clustering model for multichannel blind source separation. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: New York, NY, USA, 2019; pp. 695–699.
11. Gannot, S.; Burshtein, D.; Weinstein, E. Iterative and sequential Kalman filter-based speech enhancement algorithms. *IEEE Trans. Speech Audio Process.* **1998**, *6*, 373–385. [CrossRef]
12. Kim, J.B.; Lee, K.; Lee, C. On the applications of the interacting multiple model algorithm for enhancing noisy speech. *IEEE Trans. Speech Audio Process.* **2000**, *8*, 349–352. [CrossRef]
13. Ephraim, Y.; Van Trees, H.L. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 251–266. [CrossRef]
14. Boll, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, *27*, 113–120. [CrossRef]
15. Lim, J.S.; Oppenheim, A.V. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE* **2005**, *67*, 1586–1604. [CrossRef]
16. Martin, R. Spectral subtraction based on minimum statistics. In Proceedings of the EUSIPCO-94, Edinburgh, UK, 13–16 September 1994.
17. Cohen, I.; Berdugo, B. Speech enhancement for non-stationary noise environments. *Signal Process.* **2001**, *81*, 2403–2418. [CrossRef]
18. Gersho, A.; Cuperman, V. Vector quantization: A pattern-matching technique for speech coding. *IEEE Commun. Mag.* **1983**, *21*, 15–21. [CrossRef]
19. Roweis, S. One microphone source separation. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: South Lake Tahoe, NV, USA, 2000; Volume 13.
20. Virtanen, T. Speech recognition using factorial hidden Markov models for separation in the feature space. In Proceedings of the Interspeech, Pittsburgh, PA, USA, 17–21 September 2006.
21. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [CrossRef]
22. Wang, D.; Brown, G.J. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*; Wiley-IEEE Press: Hoboken, NJ, USA, 2006.
23. Wang, Y.; Wang, D. Towards scaling up classification-based speech separation. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 1381–1390. [CrossRef]
24. Lu, X.; Tsao, Y.; Matsuda, S.; Hori, C. Speech enhancement based on deep denoising autoencoder. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013; Volume 2013, pp. 436–440.
25. Erdogan, H.; Hershey, J.R.; Watanabe, S.; Le Roux, J. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; IEEE: New York, NY, USA, 2015; pp. 708–712.
26. Williamson, D.S.; Wang, Y.; Wang, D. Complex ratio masking for monaural speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *24*, 483–492. [CrossRef]
27. Hershey, J.R.; Chen, Z.; Le Roux, J.; Watanabe, S. Deep clustering: Discriminative embeddings for segmentation and separation. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; IEEE: New York, NY, USA, 2016; pp. 31–35.
28. Yu, D.; Kolbæk, M.; Tan, Z.H.; Jensen, J. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; IEEE: New York, NY, USA, 2017; pp. 241–245.
29. Fan, C.; Liu, B.; Tao, J.; Yi, J.; Wen, Z. Discriminative learning for monaural speech separation using deep embedding features. *arXiv* **2019**, arXiv:1907.09884. [CrossRef]
30. Kajala, M.; Hamalainen, M. Filter-and-sum beamformer with adjustable filter characteristics. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; IEEE: New York, NY, USA, 2001; Volume 5, pp. 2917–2920.
31. Klemm, M.; Craddock, I.; Leendertz, J.; Preece, A.; Benjamin, R. Improved delay-and-sum beamforming algorithm for breast cancer detection. *Int. J. Antennas Propag.* **2008**, *2008*, 761402.

32. Zhang, Z.; Xu, Y.; Yu, M.; Zhang, S.X.; Chen, L.; Yu, D. ADL-MVDR: All deep learning MVDR beamformer for target speech separation. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: New York, NY, USA, 2021; pp. 6089–6093.
33. Wang, Y.; Narayanan, A.; Wang, D. On training targets for supervised speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1849–1858. [CrossRef]
34. Tawara, N.; Kobayashi, T.; Ogawa, T. Multi-Channel Speech Enhancement Using Time-Domain Convolutional Denoising Autoencoder. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 86–90.
35. Vincent, E.; Gribonval, R.; Févotte, C. Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1462–1469. [CrossRef]
36. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; IEEE: New York, NY, USA, 2001; Volume 2, pp. 749–752.
37. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. A short-time objective intelligibility measure for time–frequency weighted noisy speech. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; IEEE: New York, NY, USA, 2010; pp. 4214–4217.
38. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Tech. Rep. N* **1993**, *93*, 27403.
39. Garofolo, J.S.; Graff, D.; Paul, D.; Pallett, D. *CSR-I (Wsj0) Complete*; Linguistic Data Consortium: Philadelphia, PA, USA, 2007.
40. Hsu, C.L.; Jang, J.S.R. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *18*, 310–319.
41. Chan, T.S.; Yeh, T.C.; Fan, Z.C.; Chen, H.W.; Su, L.; Yang, Y.H.; Jang, R. Vocal activity informed singing voice separation with the iKala dataset. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; IEEE: New York, NY, USA, 2015; pp. 718–722.
42. Rafii, Z.; Liutkus, A.; Stöter, F.R.; Mimilakis, S.I.; Bittner, R. *The MUSDB18 Corpus for Music Separation*; Zenodo: Geneva, Switzerland, 2017.
43. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; IEEE: New York, NY, USA, 2015; pp. 5206–5210.
44. Yamagishi, J.; Veaux, C.; MacDonald, K. *CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (Version 0.92)*; University of Edinburgh, The Centre for Speech Technology Research (CSTR): Edinburgh, UK, 2019; pp. 271–350.
45. Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R.J.; Jia, Y.; Chen, Z.; Wu, Y. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv* **2019**, arXiv:1904.02882.
46. Wichern, G.; Antognini, J.; Flynn, M.; Zhu, L.R.; McQuinn, E.; Crow, D.; Manilow, E.; Roux, J.L. Wham!: Extending speech separation to noisy environments. *arXiv* **2019**, arXiv:1907.01160. [CrossRef]
47. Drude, L.; Heitkaemper, J.; Boeddeker, C.; Haeb-Umbach, R. SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition. *arXiv* **2019**, arXiv:1910.13934.
48. Cosentino, J.; Pariente, M.; Cornell, S.; Deleforge, A.; Vincent, E. Librimix: An open-source dataset for generalizable speech separation. *arXiv* **2020**, arXiv:2005.11262.
49. Nakatani, T.; Yoshioka, T.; Kinoshita, K.; Miyoshi, M.; Juang, B.H. Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; IEEE: New York, NY, USA, 2008; pp. 85–88.
50. Ueda, T.; Nakatani, T.; Ikeshita, R.; Kinoshita, K.; Araki, S.; Makino, S. Low latency online blind source separation based on joint optimization with blind dereverberation. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: New York, NY, USA, 2021; pp. 506–510.
51. Sunohara, M.; Haruta, C.; Ono, N. Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; IEEE: New York, NY, USA, 2017; pp. 216–220.
52. Ono, N. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In Proceedings of the 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 16–19 October 2011; IEEE: New York, NY, USA, 2011; pp. 189–192.
53. Scheibler, R.; Ono, N. Fast and stable blind source separation with rank-1 updates. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: New York, NY, USA, 2020; pp. 236–240.

54. Nakashima, T.; Ono, N. Inverse-free online independent vector analysis with flexible iterative source steering. In Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 7–10 November 2022; IEEE: New York, NY, USA, 2022; pp. 749–753.
55. Ueda, T.; Nakatani, T.; Ikeshita, R.; Kinoshita, K.; Araki, S.; Makino, S. Blind and spatially-regularized online joint optimization of source separation, dereverberation, and noise reduction. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 1157–1172. [CrossRef]
56. Mo, K.; Wang, X.; Yang, Y.; Makino, S.; Chen, J. Low algorithmic delay implementation of convolutional beamformer for online joint source separation and dereverberation. In Proceedings of the 2024 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 26–30 August 2024; IEEE: New York, NY, USA, 2024; pp. 912–916.
57. He, Y.; Woo, B.H.; So, R.H. A Novel Weighted Sparse Component Analysis for Underdetermined Blind Speech Separation. In Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; IEEE: New York, NY, USA, 2025; pp. 1–5.
58. Jansson, A.; Humphrey, E.; Montecchio, N.; Bittner, R.; Kumar, A.; Weyde, T. Singing voice separation with deep u-net convolutional networks. In Proceedings of the 18th ISMIR Conference, Suzhou, China, 23–27 October 2017.
59. Stoller, D.; Ewert, S.; Dixon, S. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv* **2018**, arXiv:1806.03185.
60. Macartney, C.; Weyde, T. Improved speech enhancement with the wave-u-net. *arXiv* **2018**, arXiv:1811.11307. [CrossRef]
61. Défossez, A.; Usunier, N.; Bottou, L.; Bach, F. Music source separation in the waveform domain. *arXiv* **2019**, arXiv:1911.13254.
62. Fu, Y.; Liu, Y.; Li, J.; Luo, D.; Lv, S.; Jv, Y.; Xie, L. Uformer: A unet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: New York, NY, USA, 2022; pp. 7417–7421.
63. Giri, R.; Isik, U.; Krishnaswamy, A. Attention wave-u-net for speech enhancement. In Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019; IEEE: New York, NY, USA, 2019; pp. 249–253.
64. He, B.; Wang, K.; Zhu, W.P. DBAUNet: Dual-branch attention U-Net for time-domain speech enhancement. In Proceedings of the TENCON 2022-2022 IEEE Region 10 Conference (TENCON), Hong Kong, China, 1–4 November 2022; IEEE: New York, NY, USA, 2022; pp. 1–6.
65. Zhang, Z.; Xu, S.; Zhuang, X.; Qian, Y.; Zhou, L.; Wang, M. Half-Temporal and Half-Frequency Attention U 2 Net for Speech Signal Improvement. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: New York, NY, USA, 2023; pp. 1–2.
66. Cao, Y.; Xu, S.; Zhang, W.; Wang, M.; Lu, Y. Hybrid lightweight temporal-frequency analysis network for multi-channel speech enhancement. *EURASIP J. Audio Speech Music Process.* **2025**, *2025*, 21. [CrossRef]
67. Bulut, A.E.; Koishida, K. Low-latency single channel speech enhancement using u-net convolutional neural networks. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: New York, NY, USA, 2020; pp. 6214–6218.
68. Rong, X.; Wang, D.; Hu, Y.; Zhu, C.; Chen, K.; Lu, J. UL-UNAS: Ultra-Lightweight U-Nets for Real-Time Speech Enhancement via Network Architecture Search. *arXiv* **2025**, arXiv:2503.00340.
69. Ho, M.T.; Lee, J.; Lee, B.K.; Yi, D.H.; Kang, H.G. A Cross-Channel Attention-Based Wave-U-Net for Multi-Channel Speech Enhancement. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; Volume 2020.
70. Aroudi, A.; Uhlich, S.; Font, M.F. TRUNet: Transformer-recurrent-U network for multi-channel reverberant sound source separation. *arXiv* **2021**, arXiv:2110.04047.
71. Ren, X.; Zhang, X.; Chen, L.; Zheng, X.; Zhang, C.; Guo, L.; Yu, B. A Causal U-Net Based Neural Beamforming Network for Real-Time Multi-Channel Speech Enhancement. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 1832–1836.
72. Luo, Y.; Mesgarani, N. Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: New York, NY, USA, 2018; pp. 696–700.
73. Heitkaemper, J.; Jakobeit, D.; Boeddeker, C.; Drude, L.; Haeb-Umbach, R. Demystifying TasNet: A dissecting approach. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: New York, NY, USA, 2020; pp. 6359–6363.
74. Luo, Y.; Mesgarani, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1256–1266. [CrossRef] [PubMed]
75. Kavalerov, I.; Wisdom, S.; Erdogan, H.; Patton, B.; Wilson, K.; Le Roux, J.; Hershey, J.R. Universal sound separation. In Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019; IEEE: New York, NY, USA, 2019; pp. 175–179.

76. Deng, C.; Zhang, Y.; Ma, S.; Sha, Y.; Song, H.; Li, X. Conv-TasSAN: Separative Adversarial Network Based on Conv-TasNet. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 2647–2651.
77. Lee, J.H.; Chang, J.H.; Yang, J.M.; Moon, H.G. NAS-TasNet: Neural architecture search for time-domain speech separation. *IEEE Access* **2022**, *10*, 56031–56043. [CrossRef]
78. Ditter, D.; Gerkmann, T. A multi-phase gammatone filterbank for speech separation via tasnet. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: New York, NY, USA, 2020; pp. 36–40.
79. Luo, Y.; Chen, Z.; Yoshioka, T. Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: New York, NY, USA, 2020; pp. 46–50.
80. Wang, T.; Pan, Z.; Ge, M.; Yang, Z.; Li, H. Time-domain speech separation networks with graph encoding auxiliary. *IEEE Signal Process. Lett.* **2023**, *30*, 110–114. [CrossRef]
81. Sato, H.; Moriya, T.; Mimura, M.; Horiguchi, S.; Ochiai, T.; Ashihara, T.; Ando, A.; Shinayama, K.; Delcroix, M. Speakerbeam-ss: Real-time target speaker extraction with lightweight Conv-TasNet and state space modeling. *arXiv* **2024**, arXiv:2407.01857.
82. Shi, H.; Wu, S.; Ye, M.; Ma, C. A speech separation model improved based on Conv-TasNet network. *Proc. J. Phys. Conf. Ser.* **2024**, *2858*, 012033. [CrossRef]
83. Wazir, J.K.; Sheikh, J.A. Deep Speak Net: Advancing Speech Separation. *J. Commun.* **2025**, *20*.
84. Ochiai, T.; Delcroix, M.; Ikeshita, R.; Kinoshita, K.; Nakatani, T.; Araki, S. Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: New York, NY, USA, 2020; pp. 6384–6388.
85. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: South Lake Tahoe, NV, USA, 2017; Volume 30.
86. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100. [CrossRef]
87. Chen, J.; Mao, Q.; Liu, D. Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation. *arXiv* **2020**, arXiv:2007.13975.
88. Subakan, C.; Ravanelli, M.; Cornell, S.; Bronzi, M.; Zhong, J. Attention is all you need in speech separation. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: New York, NY, USA, 2021; pp. 21–25.
89. Rixen, J.; Renz, M. Sfsrnet: Super-resolution for single-channel audio source separation. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 11220–11228.
90. Subakan, C.; Ravanelli, M.; Cornell, S.; Grondin, F.; Bronzi, M. Exploring self-attention mechanisms for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 2169–2180. [CrossRef]
91. Chetupalli, S.R.; Habets, E.A. Speech Separation for an Unknown Number of Speakers Using Transformers with Encoder-Decoder Attractors. In Proceedings of the Interspeech, Incheon, Republic of Korea, 18–22 September 2022; pp. 5393–5397.
92. Lee, Y.; Choi, S.; Kim, B.Y.; Wang, Z.Q.; Watanabe, S. Boosting unknown-number speaker separation with transformer decoder-based attractor. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: New York, NY, USA, 2024; pp. 446–450.
93. Zhao, S.; Ma, B. Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: New York, NY, USA, 2023; pp. 1–5.
94. Zhao, S.; Ma, Y.; Ni, C.; Zhang, C.; Wang, H.; Nguyen, T.H.; Zhou, K.; Yip, J.Q.; Ng, D.; Ma, B. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: New York, NY, USA, 2024; pp. 10356–10360.
95. Shin, U.H.; Lee, S.; Kim, T.; Park, H.M. Separate and reconstruct: Asymmetric encoder-decoder for speech separation. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 52215–52240.
96. Wang, C.; Liu, S.; Chen, S. A Lightweight Dual-Path Conformer Network for Speech Separation. In Proceedings of the CCF National Conference of Computer Applications, Harbin, China, 15–18 July 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 51–64.
97. Wang, C.; Jia, M.; Li, M.; Ma, Y.; Yao, D. Hybrid dual-path network: Singing voice separation in the waveform domain by combining Conformer and Transformer architectures. *Speech Commun.* **2025**, *168*, 103171. [CrossRef]

98. Liu, D.; Zhang, T.; Christensen, M.G.; Ma, B.; Deng, P. Efficient time-domain speech separation using short encoded sequence network. *Speech Commun.* **2025**, *166*, 103150. [CrossRef]
99. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752. [CrossRef]
100. Wang, Z.Q.; Cornell, S.; Choi, S.; Lee, Y.; Kim, B.Y.; Watanabe, S. TF-GridNet: Making time–frequency domain models great again for monaural speaker separation. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: New York, NY, USA, 2023; pp. 1–5.
101. Wang, Z.Q.; Cornell, S.; Choi, S.; Lee, Y.; Kim, B.Y.; Watanabe, S. TF-GridNet: Integrating full-and sub-band modeling for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 3221–3236. [CrossRef]
102. Li, K.; Chen, G.; Yang, R.; Hu, X. Spmamba: State-space model is all you need in speech separation. *arXiv* **2024**, arXiv:2404.02063. [CrossRef]
103. Avenstrup, T.H.; Elek, B.; Mádi, I.L.; Schin, A.B.; Mørup, M.; Jensen, B.S.; Olsen, K. SepMamba: State-space models for speaker separation using Mamba. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; IEEE: New York, NY, USA, 2025; pp. 1–5.
104. Jiang, X.; Han, C.; Mesgarani, N. Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; IEEE: New York, NY, USA, 2025; pp. 1–5.
105. Liu, D.; Zhang, T.; Wei, Y.; Yi, C.; Christensen, M.G. Speech Conv-Mamba: Selective Structured State Space Model With Temporal Dilated Convolution for Efficient Speech Separation. *IEEE Signal Process. Lett.* **2025**, *32*, 2015–2019. [CrossRef]
106. Jiang, X.; Li, Y.A.; Florea, A.N.; Han, C.; Mesgarani, N. Speech slytherin: Examining the performance and efficiency of mamba for speech separation, recognition, and synthesis. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; IEEE: New York, NY, USA, 2025; pp. 1–5.
107. Lee, D.; Choi, J.W. DeFT-Mamba: Universal multichannel sound separation and polyphonic audio classification. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; IEEE: New York, NY, USA, 2025; pp. 1–5.
108. Venkatesh, S.; Benilov, A.; Coleman, P.; Roskam, F. Real-time low-latency music source separation using hybrid spectrogram-tasnet. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: New York, NY, USA, 2024; pp. 611–615.
109. Mishra, H.; Shukla, M.K.; Priyanshu; Dengre, S.; Singh, Y.; Pandey, O.J. A Lightweight Causal Sound Separation Model for Real-Time Hearing Aid Applications. *IEEE Sens. Lett.* **2025**, *9*, 6003504. [CrossRef]
110. Yang, L.; Liu, W.; Wang, W. TFPSNet: Time–frequency domain path scanning network for speech separation. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: New York, NY, USA, 2022; pp. 6842–6846.
111. Saijo, K.; Wichern, G.; Germain, F.G.; Pan, Z.; Le Roux, J. TF-Locoformer: Transformer with local modeling by convolution for speech separation and enhancement. In Proceedings of the 2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC), Aalborg, Denmark, 9–12 September 2024; IEEE: New York, NY, USA, 2024; pp. 205–209.
112. Qian, S.; Gao, L.; Jia, H.; Mao, Q. Efficient monaural speech separation with multiscale time-delay sampling. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: New York, NY, USA, 2022; pp. 6847–6851.
113. Zhai, Y.H.; Hua, Q.; Wang, X.W.; Dong, C.R.; Zhang, F.; Xu, D.C. Triple-Path RNN Network: A Time-and-Frequency Joint Domain Speech Separation Model. In Proceedings of the International Conference on Parallel and Distributed Computing: Applications and Technologies, Jeju, Republic of Korea, 16–18 August 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 239–248.
114. Wang, L.; Zhang, H.; Qiu, Y.; Jiang, Y.; Dong, H.; Guo, P. Improved Speech Separation via Dual-Domain Joint Encoder in Time-Domain Networks. In Proceedings of the 2024 International Conference on Electronic Engineering and Information Systems (EEISS), Changsha, China, 13–15 January 2024; IEEE: New York, NY, USA, 2024; pp. 233–239.
115. Hao, F.; Li, X.; Zheng, C. X-TF-GridNet: A time–frequency domain target speaker extraction network with adaptive speaker embedding fusion. *Inf. Fusion* **2024**, *112*, 102550. [CrossRef]
116. Liu, K.; Du, Z.; Wan, X.; Zhou, H. X-sepformer: End-to-end speaker extraction network with explicit optimization on speaker confusion. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: New York, NY, USA, 2023; pp. 1–5.
117. Zeng, B.; Suo, H.; Wan, Y.; Li, M. Sef-net: Speaker embedding free target speaker extraction network. In Proceedings of the Interspeech, Dublin, Ireland, 20–24 August 2023; pp. 3452–3456.
118. Zeng, B.; Li, M. Usef-tse: Universal speaker embedding free target speaker extraction. *IEEE Trans. Audio Speech Lang. Process.* **2025**, *33*, 2110–2124. [CrossRef]
119. Quan, C.; Li, X. SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 1310–1323. [CrossRef]

120. Kalkhorani, V.A.; Wang, D. TF-CrossNet: Leveraging global, cross-band, narrow-band, and positional encoding for single-and multi-channel speaker separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 4999–5009. [CrossRef]
121. Shin, U.H.; Ku, B.H.; Park, H.M. TF-CorrNet: Leveraging Spatial Correlation for Continuous Speech Separation. *IEEE Signal Process. Lett.* **2025**, *32*, 1875–1879. [CrossRef]
122. Lu, Y.J.; Wang, Z.Q.; Watanabe, S.; Richard, A.; Yu, C.; Tsao, Y. Conditional diffusion probabilistic model for speech enhancement. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: New York, NY, USA, 2022; pp. 7402–7406.
123. Scheibler, R.; Ji, Y.; Chung, S.W.; Byun, J.; Choe, S.; Choi, M.S. Diffusion-based generative speech source separation. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: New York, NY, USA, 2023; pp. 1–5.
124. Dong, J.; Wang, X.; Mao, Q. EDSep: An Effective Diffusion-Based Method for Speech Source Separation. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; IEEE: New York, NY, USA, 2025; pp. 1–5.
125. Michelsanti, D.; Tan, Z.H.; Zhang, S.X.; Xu, Y.; Yu, M.; Yu, D.; Jensen, J. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1368–1396. [CrossRef]
126. Afouras, T.; Chung, J.S.; Zisserman, A. The conversation: Deep audio-visual speech enhancement. *arXiv* **2018**, arXiv:1804.04121. [CrossRef]
127. Gao, R.; Grauman, K. Visualvoice: Audio-visual speech separation with cross-modal consistency. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: New York, NY, USA, 2021; pp. 15490–15500.
128. Wu, J.; Xu, Y.; Zhang, S.X.; Chen, L.W.; Yu, M.; Xie, L.; Yu, D. Time domain audio visual speech separation. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; IEEE: New York, NY, USA, 2019; pp. 667–673.
129. Kalkhorani, V.A.; Kumar, A.; Tan, K.; Xu, B.; Wang, D. Time-domain transformer-based audiovisual speaker separation. In Proceedings of the Interspeech, Dublin, Ireland, 20–24 August 2023; pp. 3472–3476.
130. Kalkhorani, V.A.; Kumar, A.; Tan, K.; Xu, B.; Wang, D. Audiovisual speaker separation with full-and sub-band modeling in the time–frequency domain. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: New York, NY, USA, 2024; pp. 12001–12005.
131. Pan, Z.; Wichern, G.; Masuyama, Y.; Germain, F.G.; Khurana, S.; Hori, C.; Le Roux, J. Scenario-aware audio-visual TF-Gridnet for target speech extraction. In Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Taipei, Taiwan, 16–20 December 2023; IEEE: New York, NY, USA, 2023; pp. 1–8.
132. Lee, S.; Jung, C.; Jang, Y.; Kim, J.; Chung, J.S. Seeing through the conversation: Audio-visual speech separation based on diffusion model. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: New York, NY, USA, 2024; pp. 12632–12636.
133. Xu, X.; Tu, W.; Yang, Y. Efficient audio–visual information fusion using encoding pace synchronization for Audio–Visual Speech Separation. *Inf. Fusion* **2025**, *115*, 102749. [CrossRef]
134. Lee, C.H.; Yang, C.; Saidutta, Y.M.; Srinivasa, R.S.; Shen, Y.; Jin, H. Better Exploiting Spatial Separability in Multichannel Speech Enhancement with an Align-and-Filter Network. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; IEEE: New York, NY, USA, 2025; pp. 1–5.
135. Qian, Y.; Li, C.; Zhang, W.; Lin, S. Contextual understanding with contextual embeddings for multi-talker speech separation and recognition in a cocktail party. *npj Acoust.* **2025**, *1*, 3. [CrossRef]
136. Wang, Q.; Du, J.; Dai, L.R.; Lee, C.H. Joint noise and mask aware training for DNN-based speech enhancement with sub-band features. In Proceedings of the 2017 Hands-Free Speech Communications and Microphone Arrays (HSCMA), San Francisco, CA, USA, 1–3 March 2017; IEEE: New York, NY, USA, 2017; pp. 101–105.
137. Luo, Y.; Han, C.; Mesgarani, N. Ultra-lightweight speech separation via group communication. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: New York, NY, USA, 2021; pp. 16–20.
138. Tan, H.M.; Vu, D.Q.; Lee, C.T.; Li, Y.H.; Wang, J.C. Selective mutual learning: An efficient approach for single channel speech separation. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: New York, NY, USA, 2022; pp. 3678–3682.
139. Elminshawi, M.; Chetupalli, S.R.; Habets, E.A. Slim-Tasnet: A slimmable neural network for speech separation. In Proceedings of the 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 22–25 October 2023; IEEE: New York, NY, USA, 2023; pp. 1–5.

140. Rong, X.; Sun, T.; Zhang, X.; Hu, Y.; Zhu, C.; Lu, J. GTCRN: A speech enhancement model requiring ultralow computational resources. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: New York, NY, USA, 2024; pp. 971–975.
141. Yang, L.; Liu, W.; Meng, R.; Lee, G.; Baek, S.; Moon, H.G. FSPEN: An ultra-lightweight network for real time speech enahncment. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: New York, NY, USA, 2024; pp. 10671–10675.
142. Li, Y.; Li, K.; Yin, X.; Yang, Z.; Dong, J.; Dong, Z.; Yang, C.; Tian, Y.; Lu, Y. Sepprune: Structured pruning for efficient deep speech separation. *arXiv* **2025**, arXiv:2505.12079. [CrossRef]