International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I10P127

Volume 72 Issue 10, 282-305, October 2024

Review Article

Dissecting Instrumental Acoustics by Comparing Traditional and Avant-Garde Techniques

S.P Sakthidevi[1], C. Divya[2], V. Kowsalya[3]

[1,2,3]*Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India.*

[1]*Corresponding Author : sakthidevisp@gmail.com*

Received: 04 June 2024 Revised: 07 October 2024 Accepted:14 October 2024 Published: 25 Ocotber 2024

Abstract *- An in-depth study of audio separation, delving into avant-garde and conventional methodologies for isolating musical tones. The exploration aims to investigate various techniques for isolating musical tones and extracting individual components of sound. By comparing traditional and advanced approaches, the study seeks to offer insights beneficial to researchers, educators, musicians, and composers. This paper briefly investigates conventional approaches like Non-Negative Matrix Factorization (NMF), Independent Deeply Learned Matrix Analysis (IDLMA), Independent Low Rank Matrix Analysis (ILRMA), Independent Component Analysis (ICA), and Principal Component Analysis (PCA), detailing their working principles and advantages. Also, Analysis of the numerous forms of machine learning techniques. Afterwards, it explores how the models are used to dissect the instrumental acoustics when deep learning techniques like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are applied. Additionally, coupled deep learning frameworks that include High-Resolution Long Short-Term Memory (HR-LSTM), Dense-U-Net, Wave-U-Net, Conv-tasnet, Res-U-Net and Long-term Recurrent Convolutional Network (LRCN) are analyzed. DenseLSTM and Audio Spectrogram Transformer are evaluated because the combined architecture is more efficient than the individual architecture. This Paper bridges avant-garde and conventional audio separation methodologies, offering valuable insights for various stakeholders and indicating a path towards enhanced practical applications in the field of audio separation.*

Keywords *- Conventional approaches, Machine learning, Deep learning, Coupled deep learning framework, Avant-Garde.*

1. Introduction

The extraction and analysis of instrumental acoustics from musical pieces represent a vital endeavour within the realm of audio processing. This comprehensive exploration delves into the methodologies employed, ranging from avant-garde to conventional techniques, to dissect and isolate instrumental tones from the complex audio landscape. By comparing traditional methods with advanced approaches, this study aims to offer valuable insights for researchers, educators, musicians, and composers alike. The research field of instrumental acoustics provides a means of accessing the complex interaction between human creativity and sound creation in the context of music. Customs, steeped in centuries of musical history, serve as a basis for the creation of compositions and the execution of performances. Conversely, avant-garde methods defy expectations by expanding the realm of auditory inquiry and rethinking the fundamental qualities of music. At its core, instrumental acoustics examines how sound is produced inside different instruments by taking into account things like material composition, resonance chambers, and playing style. Traditional techniques demonstrate a deep awareness of instrument artistry and performance history. They are frequently rooted in historical practices and cultural settings. From the rich tapestry of musical expression refined over decades, these approaches embody the resonant timbres of classical string instruments to the percussive brightness of brass and woodwinds. The fundamental goal of this study is to shed light on the diverse approaches utilised in the separation of instrumental acoustics. Conventional methodologies such as NMF [1-6], IDLMA [7-9], ILRMA [10,11,6], ICA [12-14], and PCA [15-17] form the bedrock of analysis, each offering distinct advantages and modes of operation. Furthermore, this investigation extends its purview to encompass various machine learning techniques, including reinforcement learning [24,81], unsupervised learning [20,21], semi-supervised learning [22,23], and supervised learning [18,19]. By dissecting these methodologies, researchers can gain a nuanced understanding of their efficacy in isolating and analysing instrumental acoustics. As the study progresses, it delves into the application of deep learning techniques, such as CNNs and RNNs, in the realm of audio processing. The exploration of coupled deep learning frameworks, such as HR-LSTM [47], Dense-U-Net [55-57], Wave-U-Net [50-54], Conv-Tasnet [62-65], Res-U-Net [58,59], and LRCN [49], provides a holistic view of the cutting-edge approaches employed in this

field. Moreover, the evaluation of combined architectures, such as DenseLSTM [60,61] and Audio Spectrogram Transformer [77], underscores the potential for increased efficiency and efficacy in separating instrumental acoustics. The requirements of instrumental performance are tested by unusual inquiries, on the other hand, which embrace nontraditional methods and experimental soundscapes. Here, the focus switches from following tradition to pursuing innovation as musicians work to stretch the bounds of what is technically possible. Avant-garde musicians explore unknown ground and create new aural environments that resist easy classification by employing extended approaches, electronic manipulation, and unorthodox instrumentation. Conventional approaches, rooted in respect for history and artistry, provide an insight into the enduring beauty of classical music customs. A new age of sound inquiry, where the limits of musical expression are constantly stretched and redefined, is heralded by avant-garde experimentation. Figure 1 illustrates that it is beneficial to musicians, composers, educators, researchers, and audiences alike to compare classic and avant-garde methodologies in the study of instrumental acoustics. By learning about a variety of acoustic possibilities, musicians are able to experiment and broaden their creative boundaries. Both well-established practices and novel strategies serve as sources of inspiration for composers, encouraging innovation and expanding the possibilities for musical expression. By using these parallels, teachers may enhance their courses in music education and provide students with a thorough grasp of musical history and technique. Scholars utilize the understanding gained from this examination to propel the discipline of acoustics forward and create novel methods for producing and modifying sound. A wide range of musical experiences is available to audiences, ranging from the cozy confines of tradition to the thrilling boundaries of avant-garde innovation. By taking a comprehensive approach to analyzing instrumental acoustics, one may enhance cultural heritage and encourage artistic creativity in society while also fostering awareness of the richness and diversity of musical expression.

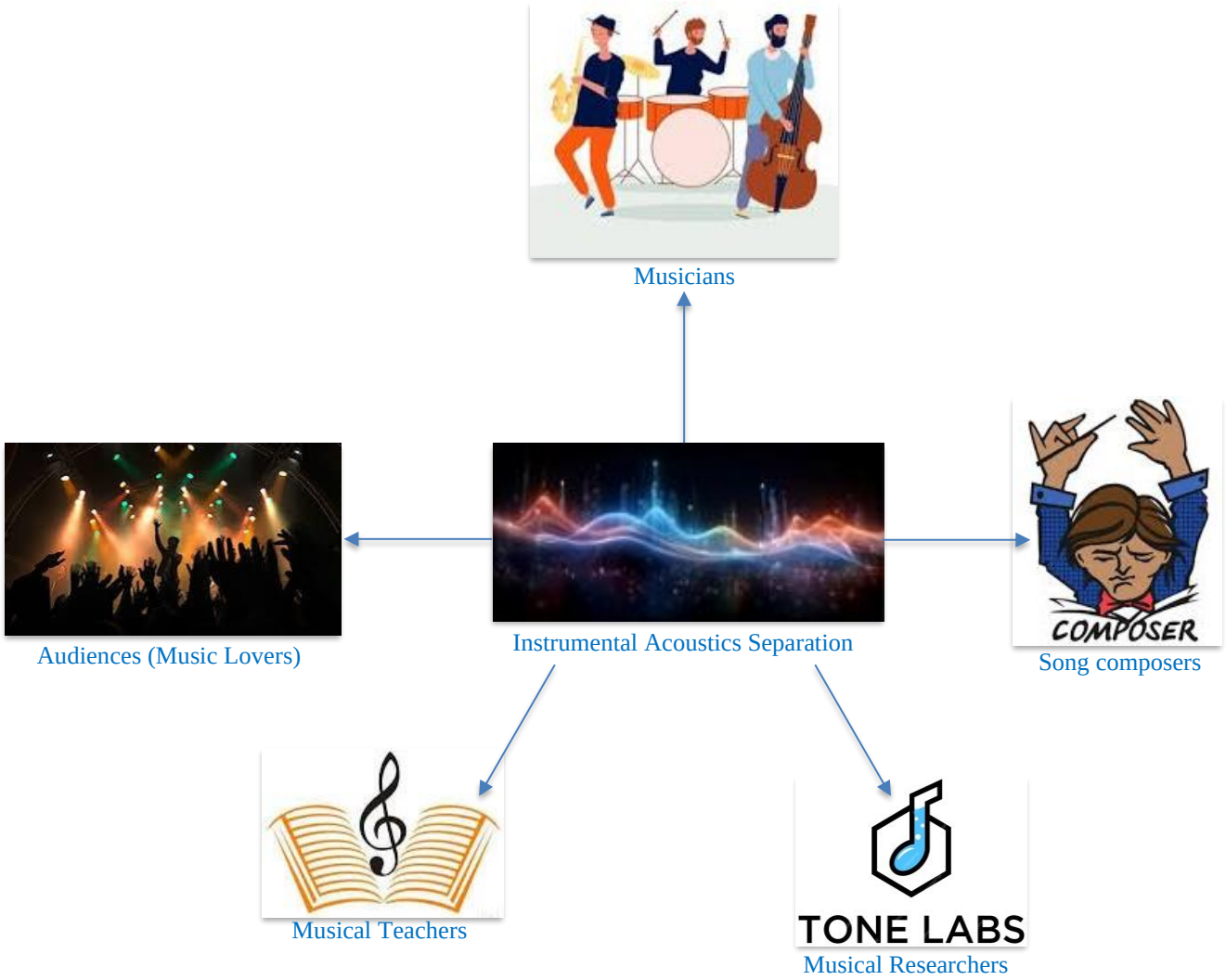

Fig. 1 Societal benefits of instrumental acoustics

Instrumental acoustics serves as a fascinating gateway into understanding the intricate relationship between human creativity and the production of sound within the realm of music. As musicians delve into this field, they navigate through centuries-old customs and traditions, drawing inspiration from the rich tapestry of musical history to inform their compositions and performances. However, alongside these time-honored practices, there exists a realm of avant-garde exploration where artists push the boundaries of sonic experimentation. Here, conventional norms are challenged, and new frontiers of musical expression are forged through innovative techniques and unconventional approaches. Through the lens of instrumental acoustics, we embark on a journey that traverses the spectrum between tradition and innovation, unveiling the diverse array of ideologies and methodologies that shape the ever-evolving landscape of music creation and performance. Our goal in doing this comparison analysis is to shed light on how musical expression and creativity are always growing while also unravelling the complexities of instrumental acoustics. Through an analysis of the methods, ideologies, and auditory environments of conventional and innovative methods, we acquire a more profound understanding of the multiplicity and energy that are intrinsic to the realm of instrumental music. We cordially encourage readers to accompany us on this voyage of inquiry and revelation as we work to solve the puzzles of sound, tradition, and innovation.

2. Conventional Voice Extraction Techniques

The conventional voice extraction techniques are discussed in this section, about primary methods like NMF, IDLMA, ILRMA, ICA, and PCA are included. These techniques provide a basis for the extraction of voice components from intricate audio signals, and they all have unique benefits and uses in the domains of source separation and audio processing. To clarify their function in improving the quality and clarity of extracted vocal content, hence advancing audio processing technologies, by analysing their fundamental ideas and working methods.

2.1. Non-Negative Matrix Factorization

The manner of splitting down a non-negative matrix into two lower-dimensional matrices is known as NMF. NMF is particularly helpful for studying data containing intrinsic non-negative characteristics, such as pictures or audio signals, because it limits the generated matrices to only contain non-negative values, in contrast to other matrix factorization techniques.

Table 1 offers a succinct overview of the main source separation strategies employed in instrumental acoustics research, outlining the procedures, advantages, and disadvantages of each strategy. In order to help academics and practitioners make well-informed decisions while searching for efficient source separation solutions, that are available for identifying distinct sound sources inside complicated audio recordings.

2.1.1. Independent Deeply Learned Matrix Analysis

IDLMA diverges from matrix factorization, employing deep learning architectures to glean an array of independent factors from datasets. It aids in tasks such as feature extraction and dimensionality reduction across various domains. By harnessing Deep Neural Networks (DNN), IDLMA extracts pertinent representations from data, capturing intricate relationships within. This approach circumvents the limitations of traditional matrix factorization techniques. Its utilization extends to diverse fields, leveraging the power of deep learning to unravel complex data structures. IDLMA stands as a versatile tool, offering insights into intricate data patterns through sophisticated neural network architectures. With its ability to unveil meaningful features, IDLMA reshapes how data is analyzed, transcending conventional methods.

Table 1. Comparative analysis of source separation NMF techniques for instrumental acoustics

Author Name & Year	Methods (NMF)	Justification	Benefits	Drawbacks
Schmidt, Mikkel N., and Morten Mørup [1] 2006	Non-negative Matrix Factorization (NMF), 2-D deconvolution	Utilizing a 2-dimensional convolution model to factorize spectrogram representations into time-frequency and time-pitch signatures, enabling effective instrument separation and analysis	Blindly separating instruments in single-channel polyphonic music and employing a convolutional model that operates across both time and frequency for factorization are central to this approach. It finds applications in automatic music transcription, music information retrieval, and computational auditory scene analysis, offering versatile solutions across these domains.	Implicitly addressing the challenge of grouping notes, this method may encounter limitations if the assumption of identical pitch-shifted time-frequency signatures for all notes is not met.
Lee, Daniel, and H.	Non-negative Matrix	NMF involves decomposing multivariate	This technique offers effective decomposition for multivariate	Sensitivity to choosing a step size in gradient-

Sebastian Seung [2] 2000	Factorization (NMF)	data into non-negative matrices, optimizing for approximation quality via iterative update rules, and ensuring convergence to locally optimal solutions.	data, with its monotonic convergence empirically proven. Its interpretation as diagonally rescaled gradient descent provides a clear framework for understanding its optimization dynamics.	based methods.
Smaragdis, Paris [3] 2004	Non-negative Matrix Factorization deconvolution (NMFD)	NMFD effectively captures complex temporal structures in audio data. NMF with temporal shift (NMF-TS), the model considers the temporal positions of spectral components, enhancing the representation of audio signals by incorporating temporal information.	This method extends its capabilities to identify components with temporal structure, making it suitable for applications such as isolating and removing different sound elements from an auditory context, even when working with a single channel input.	Lack of a useful and intuitive measure to describe separation quality.
Ozerov, Alexey, and Cedric Févotte [4] 2009	Multichannel Nonnegative Matrix Factorization (MNMF)	MNMF is a data-driven approach for multichannel audio source separation, utilizing a model inspired by NMF.	This method offers inference for audio source separation within convolutive blends, employing two distinct approaches for estimating both the mixing and source parameters: the EM algorithm and the multiplicative update algorithm.	The potential for Heightened computational complexity hinges on the precise parameterization of the model.
Lee, Seokjin, Koeng-Mo Sung and Sang Ha Park [5] 2011	Beamspace-domain & Multichannel NMF (MC-NMF)	Transforming input signals into the beamspace domain. MC-NMF aims to capture underlying patterns or features present in MC data.	This method proves more effective in multichannel source separation compared to traditional NMF techniques, leveraging the beamspace transform for enhancement purposes.	Specificity to multichannel real-world recording data.
Wang, Jianyu, and Shanzheng Guan [6] 2024	Separating sources of multichannel blind speech using a disjoint constraint source model	s-MNMF was tailored for specific optimization objectives and constraints. Disjoint constraint Penalizes components between vocal and non-vocal parts of the mixed audio signal.	Enhancing separation performance is achieved by incorporating the sparseness properties of speech signals, which involves integrating a disjoint constraint regularizer into both MNMF and ILRMA algorithms.	There is a potential for increased computational complexity, which is contingent upon accurate model parameterization.

Table 2. Audio source separation methods based on IDLMA

Author Name & Year	Dataset Applied	Method (IDLMA)	Justification	Functionalities
Makishima, Naoki [7] 2019	Music signals	IDLMA	The use of DNNs in IDLMA offers a significant advantage by allowing for the integration of pretrained DNN source models and statistical independence-based multichannel audio source separation.	Combines statistical independence between sources and DNN for separation. A heavy-tailed distribution is introduced in order to achieve better results. Uses the right data augmentation to handle a semi-supervised scenario.

Hasumi, Takuya [8] 2021	DSD100	PoSM-based IDLMA	The utilization of DNNs within IDLMA enhances its ability to model complex relationships within audio signals effectively, adaptively represent sources, efficiently estimate parameters, and address timbral mismatches, ultimately resulting in more accurate and robust source separation compared to conventional methods like NMF.	The concept of the product of source models is expanded to incorporate the DNN-based source paradigm as well as the NMF-based source paradigm. Develops a parameter estimation approach that is computationally efficient. efficient at addressing timbral mismatches.
Hasumi, Takuya [9] 2023	DSD100	PoP-IDLMA	Compared to traditional techniques such as NMF, source separation performance is improved when DNNs are used in IDLMA to learn flexible source representations, model intricate relationships within audio signals, incorporate expert knowledge, and effectively resolve timbral mismatches.	Provides source models that are built on both DNN and NMF models present the source power spectrogram's previous distribution, which was based on an expert's notion. Efficient in resolving timbral mismatches without compromising DNNs' expressive capabilities.

The main objective of these studies is to use the most traditional technique for audio source separation, IDLMA, as presented in Table 2, with modifications and improvements to handle problems such as timbral mismatches and semi-supervised situations. They introduce techniques like PoSM-based IDLMA and PoP-IDLMA, which extend IDLMA's capabilities for enhanced performance in many settings, and they evaluate their findings using datasets such as music signals and DSD100.

2.2. Independent Low Rank Matrix Analysis

Consisting of basis vectors that can be combined linearly to represent any source, ILRMA is a technique for dissecting a mixture of sources into their component elements. ILRMA is very helpful for applications like blind source separation in audio signal processing as it seeks to separate the sources while reducing interference between them.

Table 3 presents a succinct overview of the latest developments in ILRMA-based blind source separation techniques, revealing information about the approaches, datasets, and new features brought forward by every study.

2.3. Independent Component Analysis

A statistical method called Independent Component Analysis (ICA) is used to divide a multivariate signal into additive, statistically independent components. Through the utilization of the sources' non-Gaussian characteristics, ICA may reveal latent variables that lie beneath the observable data, rendering it advantageous for applications including feature extraction, picture denoising, and blind source separation across diverse domains.

Table 4 compares three studies that use Independent Component Analysis (ICA) for audio processing. A succinct synopsis of the authors' methodologies and conclusions is given, together with information on the publishing years, techniques used, advantages, and disadvantages. This systematic comparison helps to clarify the many uses and consequences of ICA in the audio processing domain.

Table 3. Blind source separation techniques based on ILRMA

Author & Year	Dataset Applied	Methods (ILRMA)	Description	Functionalities
Mogami, Shinichi, et al. (2017) [10]	Music and Speech	Independent Low-Rank Matrix Analysis (ILRMA)	Assuming a time-varying distribution for every source, ILRMA is a blind source separation approach. By minimising a negative log-likelihood function, it estimates source spectrograms from mixes using a generative model with NMF parameters.	Performance and stability in blind audio source separation are improved by a generalized source-generating model that uses a complicated Student's t-distribution.
Kitamura, Daichi, and Kohei Yatabe (2020) [11]	SiSEC 2011 and RWCP-SSD	Consistent ILRMA (Consistent	Consistent ILRMA improves blind source separation by integrating consistency in the	Optimizes isolation efficiency during estimated blind source separation by utilizing the

		ILRMA)	spectrogram during parameter updates. It guarantees coherence between time-frequency components by projecting the spectrogram of each separated signal onto a collection of uniform spectrograms, which helps solve permutation problems.	general structure of spectrograms to address permutation issues.
Wang, Jianyu, and Shanzheng Guan (2024) [6]	Wall Street Journal (WSJ0) corpus	ILRMA and Multichannel Nonnegative Matrix Factorization (MNMF)	Advanced algorithms for blind source separation, such as ILRMA and MNMF, efficiently extract speech signals from multichannel mixes by modelling the sources' spectral characteristics and boosting separation efficiency with sparse prior data.	The sparse character of speech signals is taken into account by using Bingham and Laplace distributions, which improves separation efficiency for multichannel blind voice source separation.

Table 4. Comparative overview of ICA-based techniques in audio processing

Author & Year	Methods (ICA)	Illustration	Benefits	Drawbacks
Uddin, Zahoor, et al. (2021) [12]	Sensor fault diagnosis using ICA and SOT–EST	By improving recorded signals, comparing amplitude factors with thresholds to identify problematic sensors, and then modifying mixed signals appropriately to boost ICA performance, sensor fault diagnosis using ICA and SOT-EST is accomplished.	Enhanced separation performance and efficient problem diagnostics with the use of expanded sensor techniques	Limited robustness to variance in signals.
Ezilarasan, M. R., et al. (2023) [13]	Blind source separation with ICA-FFT algorithms	Overcoming ICA's inefficiency in the presence of additive noise, blind source separation with ICA-FFT algorithms successfully denoises mixed signals and separates sources approximating the originals.	Denoising mixed signals effectively and generating isolated signals that resemble the originals	The ICA method is ineffective in the presence of additive noise.
Shihab, Ammar I. (2023) [14]	ICA	ICA is a voice and source separation signal processing approach that uses iterative optimisation to minimise mutual information between components by breaking down mixed signals into statistically independent components.	Improved robustness and accuracy in voice isolation, as well as insightful knowledge of speech processing	Permutation ambiguity complicates speech separation.

Table 5. Analysis of PCA-based techniques in singing voice separation

Author & Year	Dataset	Method (PCA)	Benefits	Drawbacks
Burute et al. (2015) [15]	MIR-1K	Robust Principal Component Analysis (RPCA)	PCA benefits by separating signals based on rank and sparsity.	Struggle with complex music compositions.
Watanabe et al. (2016) [16]	MIR-1K	Improved RPCA, Post-processing	PCA minimizes dimensionality, effectively isolating crucial information for study.	Post-processing adds computational overhead.
Li et al. (2023) [17]	ccMixter, DSD100	Weighted RPCA with Gammatone Auditory Filterbank, Vocal Activity Detection	Dimensionality reduction and feature extraction for data analysis.	Potential challenges with complex music mixtures.

2.4. Principal Component Analysis

Principal Component Analysis (PCA) serves as a dimensionality reduction method, striving to condense high-dimensional data into a lower-dimensional space while retaining the bulk of the original data's variability. This process is facilitated by identifying principal components—orthogonal vectors delineating the directions of maximal variance within the dataset. PCA's versatility finds application across diverse domains, including signal processing, image analysis, and machine learning. It proves invaluable for tasks like exploratory data analysis, aiding in the visualization of complex datasets, and mitigating noise interference. Its efficacy lies in its capacity to distill essential information while minimizing information loss. Through PCA, data analysts and researchers gain insights into the underlying structures of their datasets, facilitating informed decision-making processes. Its widespread adoption underscores its utility in extracting meaningful patterns from high-dimensional data, enhancing understanding and interpretation across various fields.Table 5 focuses on vocal activity detection, post-processing, and weighting based on a gamma tone filter bank in addition to various forms of Robust Principal Component Analysis (RPCA) to separate the singing voice from the musical backdrop. Though RPCA-based techniques are useful for distinguishing singing voices, managing intricate musical compositions still presents difficulties, and further processing stages may result in computational costs.

3. ML Approaches for Vocal Extraction

The goal of machine learning is to create algorithms as well as models that permit systems to utilize data as well as make predictions without requiring them to be manually programmed. Distinct machine learning techniques are appropriate for different kinds of challenges. Figure 2 articulates the machine learning techniques. Supervised learning makes predictions on unseen data by mapping inputs to outputs based on identified information.

To improve learning accuracy, semi-supervised learning makes use of unidentified as well as identified information. Unsupervised learning finds latent structures and patterns in unidentified information without the need for supervision. Through trial-and-error interactions with their surroundings, agents are trained to maximize rewards through reinforcement learning. Such prominent approaches are discussed below:

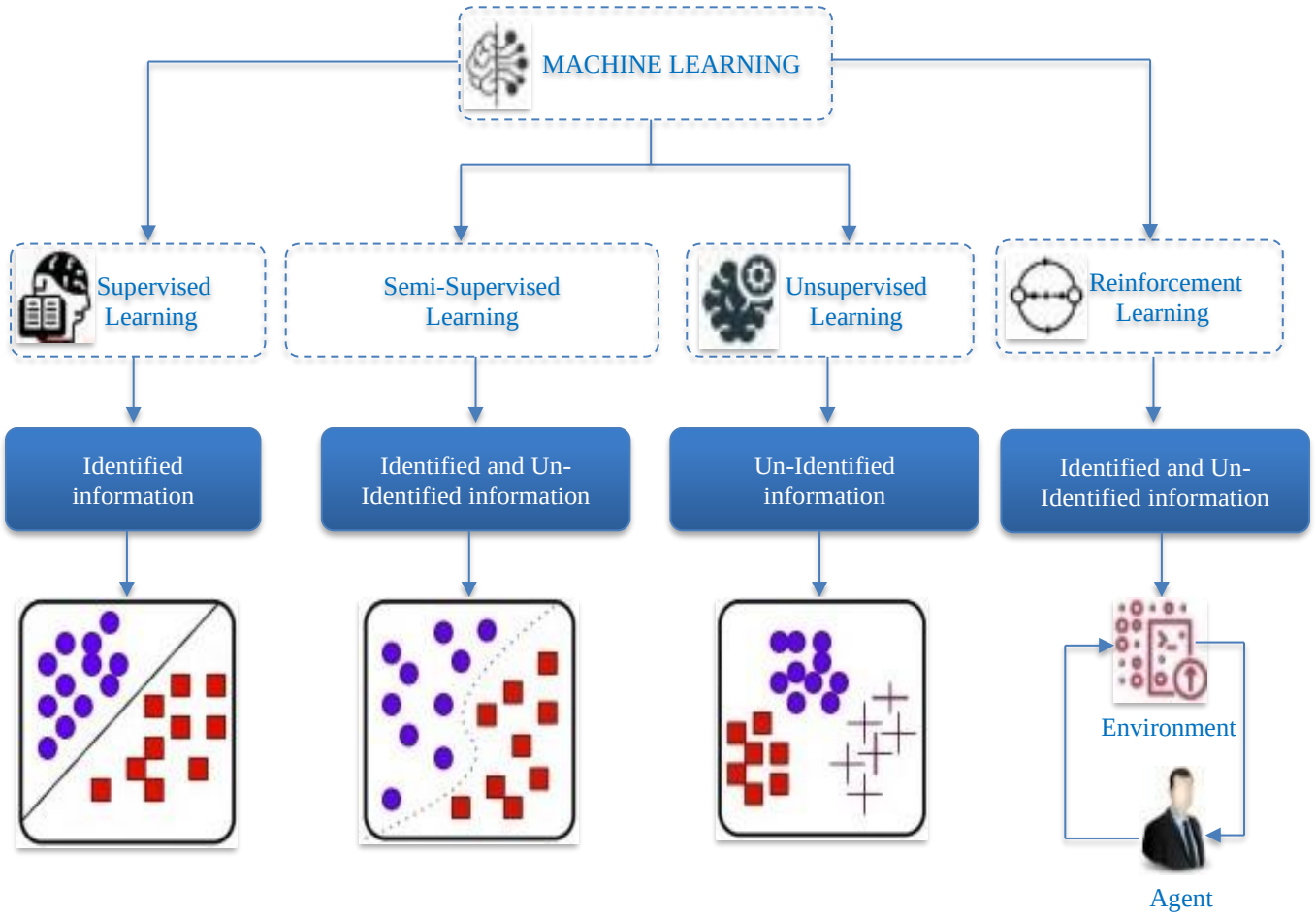

Fig. 2 Approaches of machine learning

3.1. Performance of Supervised Learning

In supervised learning, the models learn to differentiate among sources based on known inputs (mixed audio signals) and their matching labels (specific source frequencies) through the process of learning from data with labels. The two-phase method of separating sources via end-to-end Neural Autoregressive Networks (NAEs) [18] was employed, involving training and inference. During training, appropriate end-to-end NAE models for predicted sources were discovered. Pre-trained models were utilized to construct a deductive network in the inference step, which split sources from unknown mixtures. Two inference frameworks were analysed: one that utilized only decoders and the other that utilized both encoders and decoders. Either waveform spaces or activation spaces were used in the inference optimization process. The approach was illustrated using the Device and Produced Speech (DAPS) dataset. K-Nearest Neighbours (KNN) [19] had been utilized as a component of an approach that involved breaking down data into spectrograms and then processing audio recordings using soft-masking. KNN was used to classify or predict the classification of each data point depending on its proximity to other points in the dataset. Additionally associated with the supervised learning principles was the classification task, which involved using a pre-trained model from the Librosa library. The model mentioned above demonstrates supervised learning by using labelled samples to train the model to make predictions on unseen data. In order to extract voices, NAEs were used in a two-phase process: first, models were trained on labels and known inputs in order to forecast sources; next, the models were used to build a deductive network that would separate sources from unknown mixes. Additionally, spectrogram data points were classified using KNN, and supervised learning-based vocal extraction was carried out using a pre-trained model from the Librosa package.

3.2. Role of Unsupervised Learning

Unsupervised learning is intended to take relevant details from unlabelled data so the model can learn from the dataset's intrinsic structure. Differentiable parametric approaches, like as the voice production source-filter structure, were used in the DNN-based Source Separation with Source-Filter Model [20], which indicates an unsupervised approach, as the algorithm did not require independent source data during training. The training reconstruction loss was developed as a multiple-scale spectral loss by comparing the input mixture's magnitude spectral images with their estimated values. This loss function was ideal for unsupervised learning because it does not depend on labelled information. The DNN characteristics were set at test time, and each source's soft mask was created by dividing the generated source signal's magnitude spectrogram element-by-element. The fact that these soft masks were created without clear supervision suggests that source isolation was done in an unsupervised manner. The Unsupervised Multi-Source Separation (UMSS) model [20] did not require isolated source signals during training, indicating an unsupervised separation of sources. Alternatively, the model [21] learned to differentiate only between sources and mixed signals in the absence of direct supervision. The UMSS model was trained with datasets including BC1Song, BCBSQ, and the Choral Singing Dataset when individual sources were not available for training. Unsupervised learning techniques, such as the UMSS model [20], which did not require isolated source signals during training and learned to distinguish between sources and mixed signals without direct supervision, and the DNN-based Source Separation with Source-Filter Model [20], which used differentiable parametric methods like the voice production source-filter structure, were used to extract vocal information.

3.3. Behaviour of Semi-Supervised Learning

Semi-supervised learning makes use of unlabelled as well as labelled information to improve the accuracy of models, using techniques like generative models, self-training, co-training, and graph-based techniques. To enable the student model to make predictions on unlabeled data reliably, a semi-supervised learning architecture (Teacher-Student model) [22] pre-trained the instructor model with labeled data. As part of an iterative process, the student model was trained after the teacher model. Labeled datasets from RWC, MedleyDB, and iKala were used in the training process, in addition to unlabeled datasets from YouTube and FMA. Filtering data samples, training a student network with the labelled and filtered data, training a teacher separator structure on a little labelled dataset, and using the teacher model to allocate fake labels to a massive unlabeled dataset comprised the noisy self-training approach [23] for Singing Voice Separation (SVS). Using the student framework as the fresh instructor, the process was repeated until no performance increase was seen.

The fully-convolutional two-dimensional U-Net with DenseNet and focused attention blocks discovered in the PoCoNet architecture were employed by both the student and teacher models [66]. The DAMP dataset served as the unlabeled training dataset, and examples of labelled datasets were MUSDB, MIR-1K and ccMixter training split. Using semi-supervised learning techniques, such as the Teacher-Student model [22], vocal extraction was accomplished by having the student model make iterative predictions on unlabeled data with the assistance of the teacher model, which had been pre-trained with labelled data. A student network was trained using labelled and filtered data in a noisy self-training manner [23]. A teacher model was then used to provide fictitious labels to a sizable amount of unlabeled datasets, and the procedure was repeated until performance improvement reached a plateau.

3.4. Functions of Reinforcement Learning

Reinforcement Learning (RL) incorporates an agent acquiring decision-making abilities through interactions with its environment, wherein it is given feedback about its behaviour in the form of incentives or fines. To manage

complicated contexts, Deep Reinforcement Learning (DRL) was frequently used in conjunction with DNNs to optimize expected discounted cumulative rewards. DRL algorithms facilitated a variety of applications, including audio-based activities, by varying in characteristics such as policy-based or value-based approaches, off-policy or on-policy, and model-based or model-free [81]. The RL model consisted of reviewers calculating predicted rewards, actors producing vocalization instances, and an intrinsic reward. The pitch of syllables or calls created by a single-actor model could be affected by mean frequency and contextual changes.

A bunch of extended typical distributions were used to simulate the produced frequency and target, which determined the intrinsic reward. By updating the mean frequency in the direction of the goal, learning used stochastic gradient descent to reduce the mean square error. Hebbian-like learning methods were used to update the critics of many actors, who tracked the expected reward linked with vocalization instances [24].

Through the use of DRL and RL approaches, vocal extraction was accomplished. In this process, actors produced vocalisation events based on predicted rewards determined by reviewers. The RL model tracked the expected reward connected with vocalisation events by incorporating Hebbian-like learning approaches to update criticisms of various actors and random gradient descent to change average frequencies towards the target.

4. Deep Learning Techniques for Voice Separation

Neural network architectures are used in deep learning approaches for vocal separation to extract individual sound sources from mixed audio inputs automatically. Deep learning often relies on RNNs and CNNs for learning representations of audio signals and extracting individual voice components from complex audio mixtures. By investigating the core concepts and operational procedures, we can understand the role of enhancing the quality and clarity of captured vocal information and upgrading the processing of audio.

4.1. Various Convolutional Neural Network Models

CNNs are designed especially to handle data that is organized into grids, such as spectrograms or images. They are made of various layers, such as fully connected, pooling, and convolutional layers, which can effectively capture complex relationships in audio data, boosting the efficiency of separation compared to conventional extraction methods. In-depth analysis of several CNN Models, including U-net, WaveNet, denseNet, Transformer, HRNet, ResNet, and TasNet, are discussed below:

4.1.1. U-Net

U-Net aims to categorize every pixel in an image into pre-established groups or classes. Then, U-Net was adopted for manipulating audio separation tasks. It is appropriate for hiring with less training data since it promotes the effective utilisation of labelled data. While DL strategies for musical separation of sources typically performed satisfactorily in the instrument classes they were instructed on, they encountered difficulty extending and isolating instruments that were not utilised in the training phase.

To tackle that problem, suggested conditions included a U-Net separation model with few target instrument audio examples via few-shot learning. It integrated conditioning vectors at the bottleneck layer using Feature-wise Linear Modulation (FiLM) [67], which enabled tuning for various instruments. In contrast to other methods, this approach went from one-shot to a few-shot by using a more straightforward conditioning technique that was only performed at the layer of bottlenecks for better performance [25].

The U-Net architecture [68] utilised soft masks to identify speakers from audio input periphrastically. Its layers were composed of Strided 2D convolutions, batch normalisation, and the activation of Rectified Linear Units (ReLUs). Masks were implemented to change and retain the stage while handling the magnitudes of spectrograms, contributing to isolating vocals amid mixtures as well [26].

A contracting path with convolutions and Downsampling (DS), as well as an expansive path with Upsampling (US), were employed in the U-Net architecture to extract features from mixed coefficients. The training involved using Huber loss, which struck a balance between linear and quadratic rates. The integration of both actual and fictive components improved stability and accuracy in various acoustic situations, enabling real-time processing of sound [27]. Vocal extraction was achieved using U-Net architecture [68], which employed soft masks to identify speakers from the audio input and utilized conditioning vectors via FiLM for instrument separation [25]. The architecture's layers, consisting of Strided 2D convolutions, batch normalization, and ReLUs, facilitated the manipulation of spectrogram magnitudes to isolate vocals within mixtures [26].

Figure 3 reveals the typical U-Net, which earns this name because it looks like the U-shaped. It is made up of a decoder path (right side) for a precise mask and an encoder path (left side) to collect information. U-Net's skip connections (orange arrow) between corresponding layers in the decoder and encoder enable gathering small details by facilitating the flow of high-resolution information.

It is comprised of a pooling layer (red arrow), convolutional layers (blue arrow), and final convolutional layers (yellow arrow). Usually, in order to convert an extracted mask back to an image, each class in the mask is given a distinct color, which is then superimposed over an empty background.

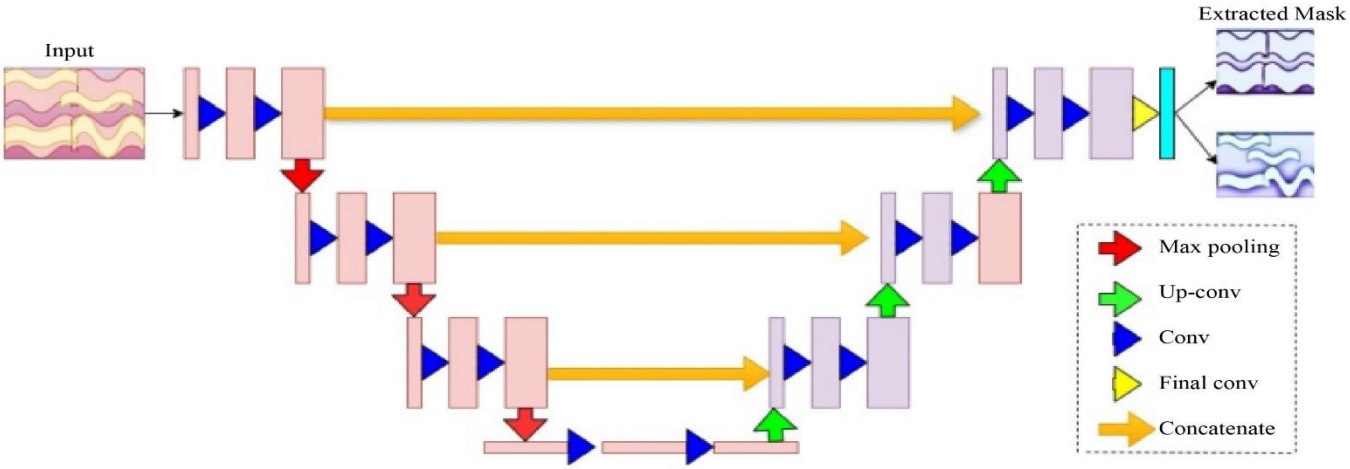

Fig. 3 Schematic structure of U-Net

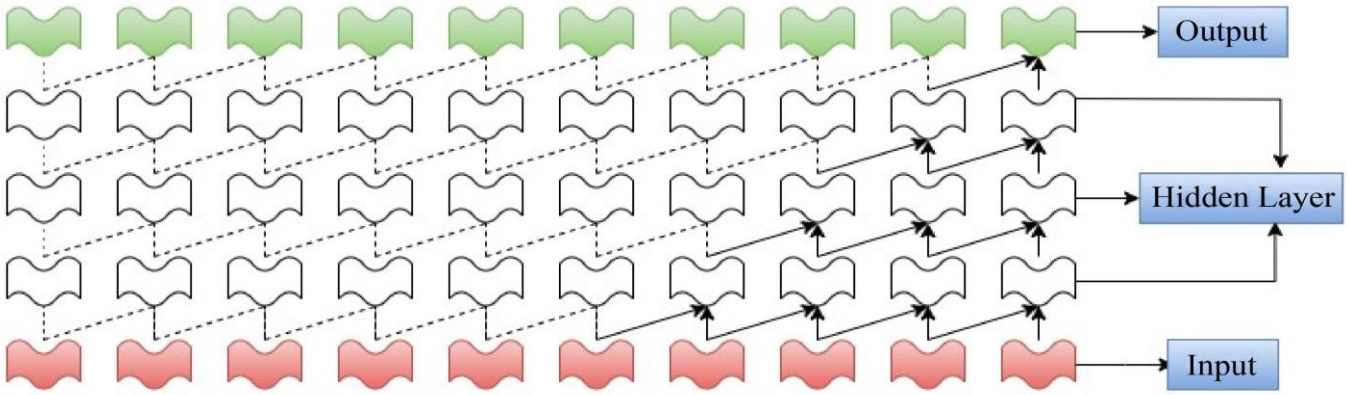

Fig. 4 Casual convolutional design of WaveNet

4.1.2. WaveNet

The primary purpose of WaveNet is to take the raw audio signal and derive realistic, high-quality waveforms from it. Its ability to generate high-fidelity audio waveforms with fine-grained details. Fast Wavenet [28] was similar to one layer of a multiple-layer RNN in that it produced audio diligently by utilizing stored recurrent phases. This procedure was made easier by its two primary parts, the convolution queues and generation model. After initializing the model and queues, repeated pop and push cycles were performed for every result, and recurrent states were estimated and altered. Wavenet [29] enforced time continuity by predicting the likelihood distribution of the subsequent data based on the preceding ones, resulting in the creation of a natural-sounding voice. It included gated units for activation control and causal, dilated convolutions for receptive field expansion, tailoring PixelCNN characteristics for audio. It also made use of skip connections for deep model training and feature inclusion and μ-law quantization for computational tractability. Context stacks reduced the receptive field length without significantly deepening the network, but they came with a large time-complexity cost in the form of sequential sample production. Vocal extraction was facilitated by WaveNet [29], which enforced time continuity by predicting the likelihood distribution of subsequent data based on preceding ones, creating a natural-sounding voice. Fast WaveNet [28] utilized stored recurrent phases, akin to one layer of a multiple-layer RNN, to diligently produce audio with fine-grained details.

Figure 4 indicates WaveNet's Casual Convolutional Design, where each convolutional layer solely considered past and present inputs due to the model's lack of access to future data. This method protected against conflicting with the natural data order, preserving fidelity to historical data without anticipation. Employing causal convolutions circumvented recurrent connections, resulting in expedited training, which is notably beneficial for lengthy sequences compared to RNNs. Nevertheless, causal convolutions faced the challenge of necessitating numerous filters or layers to expand their receptive field. The receptive field in Figure 4 remained limited to 5, calculated as the sum of layers and filter length minus one. To address this constraint, dilated convolutions were recommended. Operating akin to widening the filter by zero-padding dilation, dilated convolutions augmented efficiency by enlarging the effective filter area through selective input omission [69].

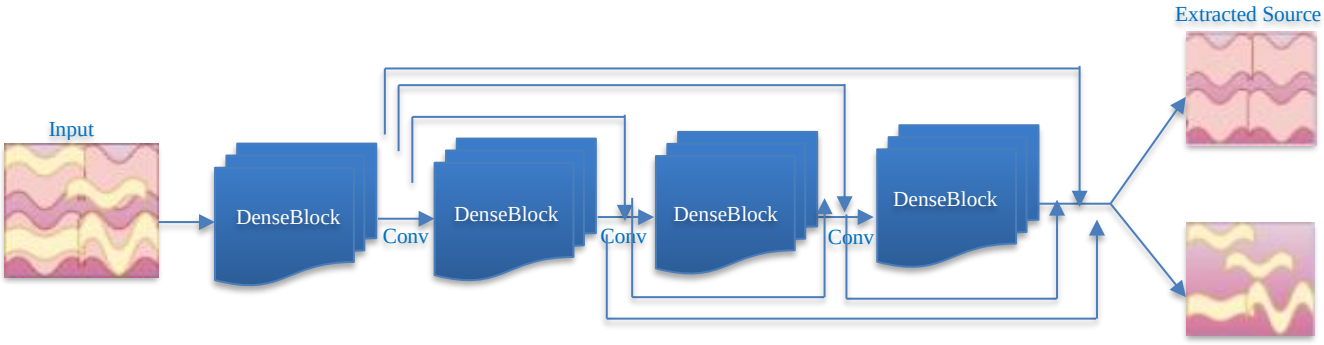

Fig. 5 Workflow of DenseNet

4.1.3. DenseNet

In the working of DenseNet, all previous layers' features are accessible to each layer due to their high interconnectedness. In addition to helping to capture complex patterns and relationships contained in the audio signals, this speeds up information propagation throughout the network. DenseNets had the potential to increase computational complexity, potentially resulting in greater processing and memory requirements. Methods such as model optimization and compression were utilized to alleviate this issue. One approach involved incorporating transition layers into the DenseNet architecture, which reduced computational load by decreasing the number of feature maps or compressing feature representations before forwarding them to higher layers. In addition, methods like pruning, quantization, and the use of effective designs were used to minimise the overall size and processing requirements of the model without sacrificing its functionality [30].

The operation of DenseNet is demonstrated in Figure 5, where the mixed audio spectrogram passes through four dense blocks. The convolution layers can alter the feature map sizes and, lastly, receive an extracted source. Recently developed approaches utilise massive neural networks to extract harmonic spectra using mixed audio, tackling the complex and ambiguous nature of audio source separation. In order to reduce computational cost and capture longer contexts and wider frequency relationships, the MMDenseNet framework makes use of dense blocks and down-sampling layers using inter-block skip connections and up-sampling layers to provide compression-free signal flow. For effective modelling of fine and global structures, input is divided into frequency bands, multi-scale DenseNet is applied to each, and outputs are then concatenated [31].

The blended song's spectrogram was represented in its Short-Time Fourier Transform (STFT) format. A DenseNet-based model was utilized, operating on magnitude spectrograms with distinct networks for each source. US and skip connections were incorporated for signal reconstruction and flow. The spectrogram was divided into frequency bands, which were then individually handled by DenseNet autoencoders and concatenated again. VGG-based feature losses and composite spectrogram losses with pixel-level L2 were employed during inference. The musdb18 dataset was used, which includes 150 professionally recorded tracks in various genres with isolated voices, drums, bass, and other sounds available for analysis, and was made available by SiSEC [32].

The source separation architecture for media and musical background recognition utilised a dilated time-frequency DenseNet. The goal of the suggested architecture was to efficiently increase the receptive field by introducing a multiband, multiscale, dilated time-frequency DenseNet. It employed both up- and down-sampling, similar to MDenseNet. Drawing inspiration from its performance in semantic segmentation tasks, dilated convolution was used to improve the receptive field. To avoid overfitting, dropout was utilized following each dilated dense block convolution layer. The DSD100 dataset was used in the studies [33].

More effective convolutional designs, such as DenseDsc, were suggested to mitigate the problem of parameter redundancy in CNNs. Depthwise separable convolutions were established, which were used to create DenseDsc, a more efficient structure, in place of regular convolutions in DenseNet. This method saved parameters without compromising functionality. Additionally, Dense2Net, inspired by Res2Net, was introduced. It built progressively larger scales inside a single convolution module to enhance multi-scale representation capabilities. To improve parameter efficiency, input feature maps were rearranged so that every convolution group received information from a different previous block. DenseDsc and Dense2Net were assessed using the CIFAR and ImageNet datasets. Tested on CIFAR and ImageNet datasets, these models optimized parameter utilization while retaining performance [34].

Vocal extraction leveraged DenseNet-based models, which operated on magnitude spectrograms with distinct networks for each source, utilizing skip connections for signal reconstruction. Additionally, a dilated time-frequency DenseNet architecture was utilized, aiming to efficiently increase the receptive field and improve source separation performance through multiband, multiscale processing.

4.1.4. Transformer

Transformer architecture has been utilised for vocal separation, which entails modifying the model to recognise and isolate the vocals from an audio file that has a variety of sound sources, including background noise, instruments, and vocals. The model learns to distinguish the vocals from the other sounds in the mixture by identifying patterns in the audio spectrogram that correspond to the vocals. It consists of an encoder and a decoder. Figure 6 illustrates how the sequence is created by the decoder and analysed by the encoder.

SepFormer was an advanced neural model for voice separation. It was designated as an RNN-free design and used a masking network made entirely of transformers. The model made use of two different kinds of transformer blocks, IntraT and InterT, to represent long-term and short-term dependencies, respectively. Several transformer layers were used, each including a Feed-Forward Network (FFW), Multi-Head Attention (MHA), and layer normalization.

The datasets WSJ0-3mix and WSJ0-2mix were utilised to assess the model's performance for source separation [35]. An adaptation of the original Hybrid Demucs model was the Hybrid Transformer Demucs model, which included dual U-Nets, one in the spectrogram domain and another in the time domain. The Cross-domain Transformer Encoder, which processed spectral and temporal data using self-attention and cross-attention, replaced the deepest convolutional layers in the Hybrid Transformer Demucs. To increase memory usage and attention performance, the model made use of Locally Sensitive Hashing (LSH) and sparse attention kernels. As a result, the model could scale to longer sequences [36].

For source separation tasks, the Single-Input-Multi-Output (SIMO) [74] paradigm employed multiple outcomes that corresponded to the target source spectrograms and just one input that represented the input mixing. Drawing inspiration from the U-Net topology, the model integrated skip connections within an encoder-decoder design. Convolutional layers were used to downsample spectral representations of features, followed by the inclusion of residual CNN blocks. These blocks facilitated the recovery of high-resolution information by aggregating details from neighboring regions. At the core of the architecture lay the Stripe-Transformer block, a crucial component responsible for capturing dependencies between vertical and horizontal stripes in representations of features with multiple scales. The Stripe-Transformer block allowed for efficient modeling of complex relationships, comprising three modules: a Squeeze-and-Excitation (SE) [73] module, a Mixed-scale Convolutional Feed-Forward Network (MixCFN) [72], and a Stripe-wise Self-Attention (SiSA) module. The Musdb18 dataset was utilized for both training and evaluation purposes [37]. Vocal extraction was achieved through the SepFormer model, utilizing a transformer-based masking network. In contrast, the Hybrid Transformer Demucs model employed a cross-domain transformer encoder for spectral and temporal data processing, enabling source separation. Additionally, the SIMO paradigm incorporated skip connections and a Stripe-Transformer block for efficient modeling of complex relationships in the source separation task.

4.1.5. High-Resolution Network

High-Resolution Networks (HRNet) might be useful with tasks like identifying overlapping sources with similar spectral characteristics by maintaining fine-grained features in audio signals. This could involve integrating techniques for attending to both short-term and long-term temporal features, as well as recurrent or convolutional layers designed specifically to interpret audio data. Figure 7 Exploring an outline of HRNet, which eliminates fixed-resolution DS by simultaneously extracting features at numerous resolutions from an input image. It processes information at different resolutions simultaneously while capturing minute details and the larger context. To preserve high resolution, HRNet began with a magnitude spectrogram input and progressed through two convolutional layers without DS. The process comprised four steps featuring residual units akin to ResNet-50 [75], gradually refining the representation. Multi-resolution convolutions and fusion were employed in each step to exchange data between branches with varying resolutions. Each step doubled the number of channels while halving the resolution.

The final output comprised four feature maps, which were processed and concatenated to form the target mask. Selected datasets, including MIR1k, MUSDB18, MedleyDB, iKala, RWC Popular Music, MIREX05, and ADC2004, were utilized to train HRNet for tasks such as source separation and Vocal Melody Extraction (VME) [38]. To retain resolution, HRNet processes the input image first through an image stem made up of two stride convolutions. Streams with varying resolutions are gradually included in the main body by repeated multi-resolution fusions and parallel multi-resolution convolutions. Ultimately, the Representation Head, similar to HRNetV1, HRNetV2, and HRNetV1p, integrates low-resolution representations by concatenating and US them, or it utilises high-resolution streams alone to process the outputs for various tasks [39].

Using convolutional layers to analyse magnitude spectrogram inputs without DS, multi-resolution convolutions and fusion to refine representations gradually, and concatenated feature maps to produce target masks were the methods used by HRNet to perform vocal extraction. By using two stride convolutions to process the image stem and progressively adding streams with different resolutions via multi-resolution fusions and parallel convolutions, the HRNet architecture was able to maintain input resolution. In the end, low-resolution representations were integrated for voice utilisation tasks.

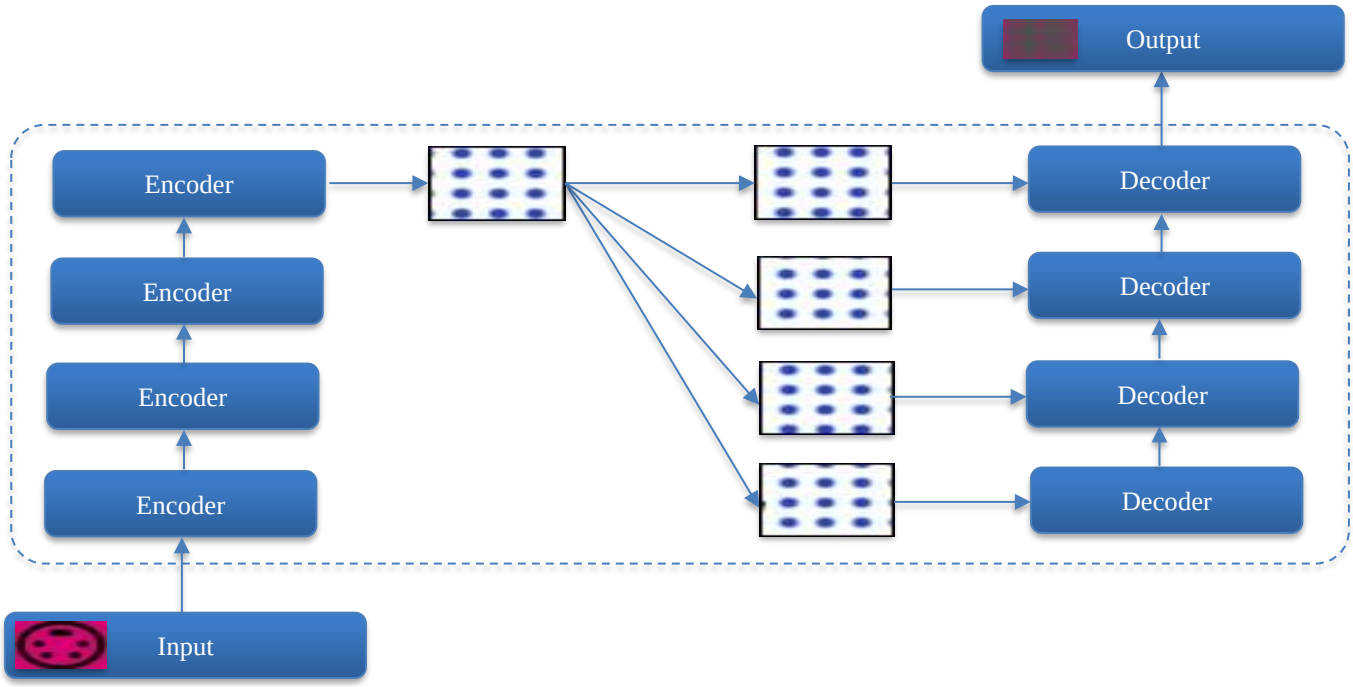

Fig. 6 Unveiling the transformer model

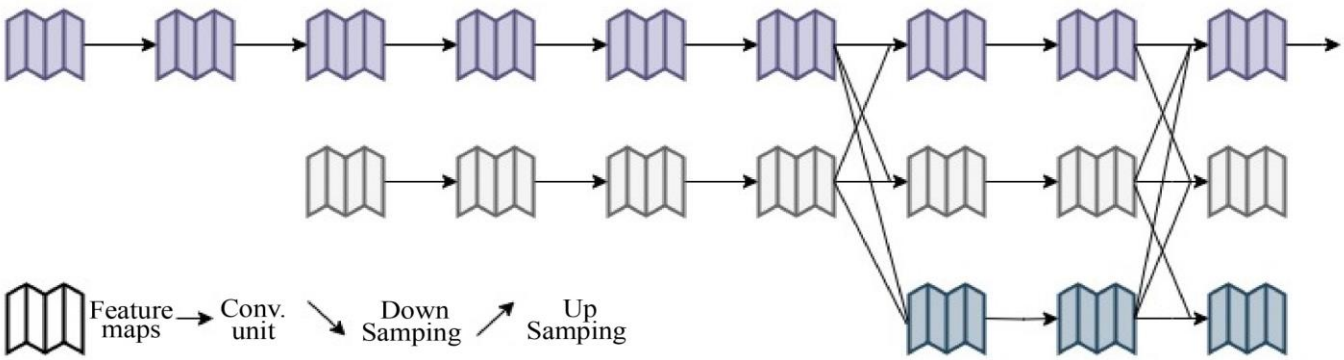

Fig. 7 Overview of HRNet

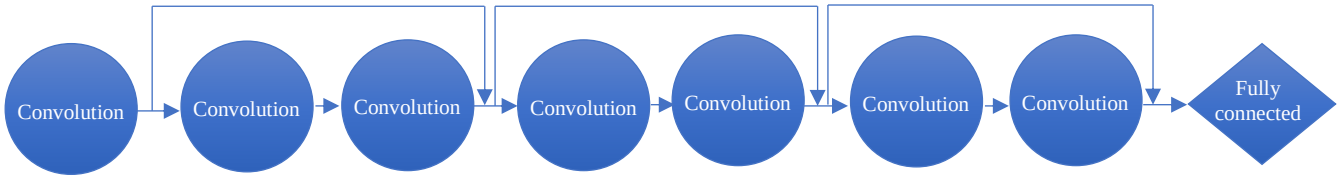

Fig. 8 Primary structure for ResNet

4.1.6. Residual Networks

A residual network, commonly known as ResNet, allows the network to attempt to learn residual mappings rather than the intended underlying mapping directly. Figure 3 reflects the framework of ResNet, where the residual block has more than one convolutional layer. A sequence of convolutional layers processes the input and extracts features. The features are flattened and then sent through one or more fully connected layers after the convolutional layers. These layers map the features to the intended output and further alter them. Nonlinear processing without shortcut connections is made possible by the addition of parallel streams for residual and transient information flow in the ResNet extended residual block. It enables variable-depth processing by merging these streams and utilising batch normalisation and ReLU to bridge standard CNNs and ResNet blocks. This method, when used repeatedly, creates the generalised residual architecture, which can be adjusted for different processing depths. RiR is a derivative architecture that improves expressivity and flexibility by substituting generalised residual blocks for the

convolutional layers of ResNet [40]. An autoencoder with a latent separation network based on ResNet was trained. To create distinct latent representations, the architecture mapped inputs to a latent space using an encoder and a residual network. After that, a common decoder received these representations in order to reassemble them. The input mixture was divided into length segments that each comprised frequency bands along the time axis. The first latent representation was produced by the encoder, a 3-layer CNN, and fresh latent representations were then produced by a residual network. A decoder evaluated each source. It was assessed using the DSD100 and MUSDB18 datasets [41]. It incorporates a residual network and an encoder to map inputs to a latent space for unique representations, which makes voice extraction easier. Splits the input mixture along the time axis into length segments that contain frequency bands. The first latent representation is produced by the encoder (3-layer CNN), and then new representations are produced by the residual network. The decoder assesses every source.

4.1.7. Time-domain Audio Separation Network

Unlike conventional techniques that require frequency-domain representations like spectrograms, Time-domain Audio Separation Networks or TasNet for short acts directly in the time domain. TasNet can comprehend intricate temporal correlations in the audio signal. Figure 9 discloses the TasNet Encoder-Decoder Architecture, which receives a mixed waveform as input, which is split up into smaller units called frames and comes from several sources. Usually, each section has an amount of time.

The encoder neural network changes the input mixture into a latent representation that has the necessary information to distinguish between different sources. After that, a separation module receives the latent representations in an effort to separate the mixed waveforms into their sources. For every source, the discrete latent representations are decoded back into individual audio signals. Finally, acquire an extracted waveform. Using TasNet and Dual-path RNN (DPRNN) in the time domain, it attempted to create a multi-talker, real-time, speaker-independent separation of speech structure. Architectural changes included adding skip connections between stacked RNNs in the separator, swapping out the encoder's 1-D convolutional layers for complete layers and 50% overlap in voice.

The model used the Acoustic-Phonetic Continuous Speech Corpus dataset to conduct experiments [43]. The suggested Beam-TasNet integrated frequency-domain beamformers, such as Minimum Variance Distortion-less Response (MVDR) [76], with TasNet. A frequency domain blend enhanced the observed Short Time Fourier Transform (STFT) coefficients of the mixture. TasNet outputs were used to compute spatial covariance (SC) matrices for voice and noisy signals, allowing for improved time-domain waveforms and beamforming filter coefficient construction. A cross-correlation function-based inter-channel permutation solver was developed to resolve the inter-channel permutation problem and guarantee the correct alignment of TasNet outputs for every channel [42]. This study investigated the difficulties in modifying current demixing models to meet the demand for current time minimal latency audio processing applications, like live concerts and hearing aids. The suggested HS-TasNet, drawing inspiration from the Hybrid Demucs architecture, utilized both the spectral and waveform domains to enhance performance. The architecture of HS-TasNet comprised LSTM-based memory blocks, a learned convolution encoder, and a spectrogram encoder, concatenating 1000 hidden units. To achieve comparable performance with fewer parameters, HS-TasNet-Small, a computationally less expensive alternative, used summation instead of concatenation and single LSTMs in memory blocks. The MusDB-HQ dataset was used to train the model [44].

Time-domain processing is used in these experiments to separate speech in real-time, independent of the speaker. Complete layers in the encoder are also integrated, and a 50% overlap in voice is implemented. An alternative method combines frequency-domain beamformers with Beam-TasNet to improve STFT coefficients and uses TasNet outputs to compute spatial covariance matrices for noisy and spoken data. They also include waveform and spectral domains for vocal extraction, including 1000 hidden unit spectrogram encoder, convolution encoder, and LSTM-based memory blocks.

4.2. Vocal Separation on Recurrent Neural Network

This section explores methods for voice separation using RNNs. RNN-based methods take use of the sequential structure of audio data to extract voice components from mixed signals and capture temporal relationships. These models offer a viable solution for audio source separation problems by employing frameworks such as Gated Recurrent Units (GRUs) or Long Short-Term Memory (LSTM), which show promising results in identifying vocalists from background noise or music.

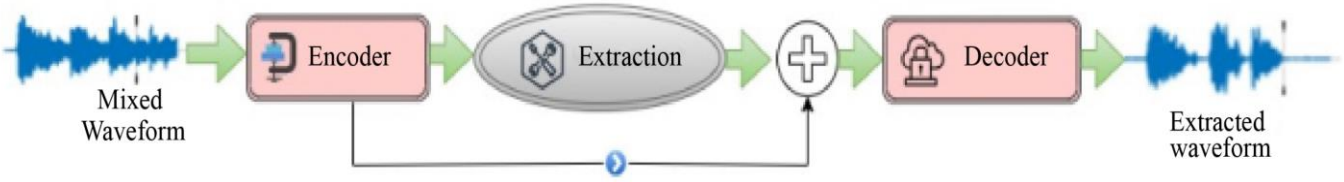

Fig. 9 TasNet Encoder-Decoder paradigm

4.2.1. LSTM and GRU based Models

Several research works have investigated the effectiveness of RNN designs, including GRU and LSTM models, within the domain of audio source separation. To accomplish multi-talker speaker-independent speech separation tasks, a Speech separation model in real-time [44] makes use of the TasNet and DPRNN. They experimented with different batch sizes, optimizers, and RNN topologies to examine how hyperparameters affected the performance of the model. Using the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus dataset, their study showed that their model was more accurate and had less latency than earlier models. In order to extract voices from musical compositions, developed algorithms that use neural networks to comprehend harmonic overtones and the mathematical patterns of spoken language. Their research [45] concentrated on two different model architectures: a GRU-based model that captured speech temporal relationships and a CNN-based model that employed semantic segmentation techniques. The CNN-based model also produced encouraging results, but the GRU-based system performed better and demonstrated the viability of quick inference with fewer datasets. Bidirectional GRU (BGRU) and Gaussian Mixture Models (GMM) were added to the improved Deep Attractor Network (DANet) [46] for voice separation in order to lower model complexity and improve learning speed and accuracy. Their approach outperformed the original DANet model concerning the Perceptual Evaluation of Speech Quality (PESQ) and Signal-to-Distortion Ratio (SDR) scores when tested on speaker mixture datasets from the TIMIT corpus, demonstrating improvements in speech separation methods.

Combining an LSTM module with a High-Resolution Learning (HR-Net) system [47], known as HR-LSTM, solved the drawbacks of low-resolution representation in music source separation. The goal of this method was to enhance the separation of music sources by preserving high-resolution feature maps and capturing temporal dynamics. When the HR-LSTM system was tested on many datasets, such as DSD100 and MIR-1K, it performed better than earlier techniques and showed improved accuracy in differentiating between singing voices. The band-split RNN (BSRNN) [48] model was expanded to facilitate stereo signal analysis in music source separation. Through the adaptation of BSRNN to a stereo and Single-Input-Multiple-Output (SIMO) mode, their goal was to lower the costs associated with inference and training while maintaining overall system performance. Wherein the temporal dimension on T is successively applied to two residual bidirectional LSTM (BLSTM) layers. The results of the experiment demonstrated the efficacy of SIMO stereo BSRNN in improving the separation of music tracks, providing improvements in stereo signal demonstrating for source separation tasks. The utilization of RNN-based models for audio source separation has advanced significantly, as seen by this research; nonetheless, issues with model architecture optimization, computational complexity, and reliable performance on a variety of datasets and real-world applications still need to be resolved. Novel ways to enhance the effectiveness, precision, and scalability of RNN-based methods in audio source separation tasks should be investigated further.

5. Integrated Deep Learning Techniques

When multiple deep learning architectures and methodologies are utilised to extract discrete sound sources from a mixture of audio signals, this is referred to as integrated deep learning techniques. Incorporating domain-specific knowledge into the network design or training process or merging various architectures into hybrid models are common integration tasks for these techniques. It tends to rely on HR-LSTM, Dense-U-net, Wave-U-net, Res-U-Net, DenseLSTM, Conv-TasNet and long-term Recurrent Convolutional Network (LRCN). The following section discusses each of the strategies mentioned above.

5.1. HR-LSTM

HR-LSTM [47] improved SVS by integrating LSTM and HRNet. It had four branches that processed mixed spectrograms concurrently. These branches upheld resolutions. Features were concatenated following LSTM processing to provide multi-scale feature maps that depicted both local and global structures. A fusion layer balances time and frequency resolutions by combining data from high-, medium-, and low-resolution feature maps. Each branch's output feature maps were produced by adding the features that had been up- and down-sampled. The architecture improved separation accuracy by effectively capturing global structures and minute local features. The Nepal Idol SVS (NISVS) and DSD100 datasets were utilised.

5.2. Long-Term Recurrent Convolutional Network

For deep feature extraction, LRCN, a deep topological temporal model, combined spatial information and contextual interactions between time series. Consecutive frames' audio features created two-dimensional figures that aided LSTM layers and convolutional layers with deep feature encoding and spatial extraction, respectively. Input, hidden, and output layers made up the LRCN architecture, which also included LSTM cells that learned temporal dynamics and convolutional filters. To ensure robustness and diversity in model evaluation, data from publicly accessible datasets like Jamendo, RWC, MIR1k, iKala, and MedleyDB were used for both training and validation [49].

5.3. Dense-U-Net

A DL architecture for the separation of sources and speech augmentation was called Dense-U-Net. It consisted of a mask estimation network, a decoder, and an encoder. The encoder created speech and noise masks for the input combination by utilizing STFT. To estimate noise and clean speech, these masks were applied to the input. To improve voice features, the mask estimation network used a U-Net

variation with Channel Attention (CA) units for beamforming-like operations. Learned attention weights that reflected the Signal-to-Noise Ratio (SNR) demonstrated how the CA mechanism dynamically adjusted attention to optimize signal extraction by capturing global dependencies. Practice with datasets such as CHiME-3 improved robustness to changing SNRs [57].

A deep learning model named Dense-U-Net was designed for source separation in musical blends. It predicted voice and accompaniment masks by utilizing the magnitude of the STFT of a musical combination as input. After then, these masks were added to the mixture in order to recreate the original origins. The model employed self-attention techniques to capture long-range dependencies in music structures, utilizing DenseNet blocks for feature extraction. It was constructed on a modified version of the UNet architecture. Dense blocks, DS, US, and self-attention subnets were all incorporated into the architecture to facilitate effective learning and the capture of intricate relationships. Once trained on datasets like DSD100, MedleyDB, and CCMixter using the ADAM optimizer, the model's performance was assessed using metrics such as the Source-to-Artifact Ratio (SAR), Source-to-Distortion Ratio (SDR), and Source-to-Interference Ratio (SIR) [55].

Traditional methods, such as CNNs, encountered performance degradation when time-frequency characteristics rapidly changed. The widely adopted Dense U-Net employed a bottleneck structure and dense block for feature extraction, yet it did not consider the spectrogram variations among different instruments. This issue was tackled by deformable convolution (deform-conv), which adapted the convolutional receptive field to the characteristics of each instrument. By introducing variable receptive fields and modifying sample sites in response to input features, deform-conv enhanced separation performance [56].

The technique generates speech and noise masks using a mask estimation network with a U-Net variation and CA units, improving voice attributes by means of dynamic attention changes. It also integrates self-attention methods to capture long-range dependencies, uses STFT magnitude to estimate voice and accompaniment masks, and integrates DenseNet blocks into a modified u-Net architecture. Additionally, it enhances separation performance by modifying sample sites according to input features through the use of deform-conv to alter convolutional receptive fields.

5.4. Wave-U-Net

A one-dimensional modification of the U-Net architecture intended for Audio Source Separation (ASS) was called the Wave-U-Net [50]. It avoided preset spectral transformations and worked directly in the time domain while maintaining phase information. Through iteratively resampling feature maps at various time scales, long-range temporal correlations that were essential for high-quality separation were captured. It computed multi-scale features using blocks of DS and US, guaranteeing precise predictions. In order to prevent unlikely outputs, it was essential to impose source additivity via a different output layer. Transposed convolutions were replaced for the US with linear interpolation and convolution, which minimized artifacts. This architecture offered promising improvements in ASS and performed comparably to spectrogram-based techniques.

An improvement on the U-Net design, the Wave-U-Net [51] processed audio signals directly, reducing problems with reconstruction. Constantly resampling feature maps allowed it to capture long-range temporal correlations. Accurate predictions were guaranteed by the computation of multi-scale features by DS and US blocks. In order to prevent improbable outputs, it was essential to impose source additivity using a different output layer. Transposed convolutions were replaced by interpolation using linearity and convolution for the US, reducing artifacts. The musdb18 dataset was processed independently for voice, drum, bass, and accompaniment in order to do data augmentation. Pitch shifting, temporal stretching, and harmonic envelope modifications were examples of operations. Pitch and formant shifting in singing voice transformation were continuously controlled by F0.

In order to facilitate the implementation of Acoustic Echo Cancellation (AEC), Wave-U-Net integrated an attention network and an auxiliary encoder. It estimated near-end speech by using as inputs a mixed signal and far-end speech. Far-end speech features were extracted by the auxiliary encoder, and pertinent features were emphasized by the attention network. The encoder of Wave-U-Net processed concatenated features in order to extract pertinent characteristics and recover clear near-end speech. Similar features were found in the latent space of far-end speech, and this architecture made use of attention mechanisms to enhance performance [52].

The suggested model addressed problems akin to Multi-Resolution Analysis (MRA) by integrating Discrete Wavelet Transform (DWT) and inverse DWT layers into Wave-U-Net. It maintained the same encoder-decoder architecture with DS blocks and US blocks as Wave-U-Net. With the DWT layer, every DS block reduced the time resolution by half, and every US block increased it by double using the inverse DWT layer. A reflection padding layer preceded every DWT layer, and the final time elements of feature maps produced by inverse DWT layers were removed. This model utilized DWT and inverse DWT layers to provide a more dependable source separation technique. The framework was reduced to Wave-U-Net if linear US and decimation layers were substituted and if reflection padding layers were omitted [53].

For multi-channel speech enhancement, the TC-Wave-U-Net used an encoder, bottleneck, and decoder topology.

Dilated casual convolution, batch normalisation, non-linearity, dropout, and convolution were the steps in each block. A Parametric Rectified Linear Unit (PReLU) came next [79]. Self-attention mechanisms captured global dependencies, enhancing convergence and relevant feature extraction. Down-operation halved the time dimension after each Temporal Convolution (TC) block. Linear interpolation was used for up-sampling in the decoder. Streaming inference incorporated hidden state reuse to extend history context length without dramatic accuracy reduction. The Chinese and English datasets from VCTK, Librispeech, AISHELL-1, and AISHELL-3 were used to train the model [54].

To extract vocals, input the audio mixture into the Wave-U-Net model, generating masks to separate vocals from other sources. Apply these masks to isolate vocals and perform post-processing for refinement, ensuring clarity and fidelity of the extracted vocals. Evaluate the quality of the separation to confirm effectiveness in separating vocals from accompanying music or noise.

5.5. Res-U-Net

The Res-U-Net architecture addressed music source separation using real-valued Short-Time Discrete Cosine Transform (STDCT) data. It employed MultiRes blocks, residual skip connections, and an attention system. MultiRes blocks with progressively larger filter sizes and Res routes were integrated into the encoder-decoder architecture for effective feature extraction and transmission. A self-attention module aided in preserving important traits while suppressing unimportant ones. Mean Square Error (MSE) served as the loss function for regression, with continuous application of batch normalization and Exponential Linear Units (ELU) activation [78]. Each network was trained separately for every stem, producing a mono STDCT representation of a single desired stem [58].

The three-dimensional (3D) Inception-ResUNet structure [59] utilized spectral and spatial information from spectrograms to improve SVS on robots. It implemented magnitude correlation consistency loss, phase consistency loss, and consistency loss as multi-objectives for training. The architecture comprised six encoder or decoder layers, fractionally stridden convolutions, and directional Inception-ResNet blocks. Each layer integrated Inception-ResNet blocks to handle multichannel spectrograms, enabling alignment of singing voice and accompaniment sources and management of sound propagation delays during recording. To prevent representational bottlenecks, the reduction blocks in the encoder layers used simultaneous 3×3 convolutions with stride 2.

5.6. Conv-TasNet

Conv-TasNet employed a time-domain fully convolutional method, revolutionizing single-channel speech separation. It addressed the shortcomings of conventional time-frequency approaches by directly operating on waveform representations. Temporal convolutional networks estimated masks for individual speakers at phases of encoder, separation, and decoder. Superior performance in applications that operated in real-time was ensured by employing normalization techniques and depth wise separable convolutions, which improved efficiency and adaptability [62].

Conv-TasNet, a waveform-based neural network system, achieved cutting-edge performance in speech separation. For expressing more intricate signal transformations, this article suggested improving the encoder or decoder by using deep, nonlinear variations. The deep encoder used a stack of small filters with nonlinear activations to transform the waveform into a nonlinear space known as latent. The deep decoder processed masked encodings to generate time-domain estimated source signals, mirroring the architecture of the encoder. Dilated convolutional layers were used as a variation to enhance the temporal receptive field, and Gated Linear Units (GLUs) [80] were used in place of PReLU activations to represent the relative kernel importance [63].

Demucs and Conv-TasNet were two architectures utilized for separating music sources. Conv-TasNet employed masked separation through stacked residual blocks, then transformed the signal into a high-dimensional representation, and finally reconstructed the sources. Demucs, however, estimated stereo sources directly from the input mixture waveform using an encoder-decoder structure with skip connections, bidirectional LSTM, and convolutional layers. Conv-TasNet adapted its architecture for stereophonic music separation, while Demucs improved performance with skip connections and convolutional operations. Both models aimed to isolate individual sources from mixed audio [64].

The suggested framework, VAT-SNet [65], addressed single-channel music source separation by optimizing Conv-TasNet's structure. Convolution at the sample level was utilized in both the encoder and decoder to preserve deep audio data. An auxiliary network's voice and accompaniment embeddings had an impact on mask accuracy. VAT-SNet's separator combined these embeddings to create masks, enhancing separation accuracy. Waveform music was processed directly by the encoder, which fragmented it and applied deep convolution layers to extract information. The decoder reconstructed the separated sources using a symmetrical deconvolution method. During mask production, data from the main network was fused with voice and accompaniment embeddings produced by an auxiliary network that extracted deep acoustic properties. Consistent feature representation was ensured by the common weights throughout the main and auxiliary networks, facilitating fusion during mask construction. The music extractor then used 1D convolution and ResNet to extract features from the initial music sequence normalized it, and projected it onto a fixed-dimensional embedded space.

Table 6. Evaluation of performance

Author and Year	Model	Dataset	Vocal			Accompaniment		
			SDR	SIR	SAR	SDR	SIR	SAR
Takahashi et al., 2018 [60]	MMDenseLSTM	MUSDB18	4.77	53.40	19.62	15.80	37.74	21.32
Gong, Yuanet al., 2021 [77]	Audio Spectrogram Transformer	AudioSet	10.79	159.14	70.28	-	-	-

To isolate vocals from mixed audio, employ Conv-TasNet's time-domain fully convolutional method, which directly operates on waveform representations, estimating masks for individual speakers at various stages. Enhance the system by utilizing deep, nonlinear variations in both the encoder and decoder, incorporating dilated convolutional layers and GLU for improved performance. Additionally, consider VAT-SNet, which optimizes Conv-TasNet's structure for single-channel music source separation, utilizing convolution at the sample level, auxiliary networks for mask accuracy, and symmetrical deconvolution in the decoder, ultimately enhancing separation accuracy by fusing embeddings and ensuring consistent feature representation across the network.

5.7. DenseLSTM

The suggested dilated time-frequency multi-scale multi-band DenseLSTM (MMDilDenseLSTM) [61] combined dilated blocks [33] with LSTM and DenseNet for audio source separation. Dilated blocks addressed independent effects on both frequency and time axes, improving the field of reception in spectrograms. MMDilDenseLSTM used parallel MDilDenseLSTM architectures and separated spectrograms into frequency bands. Through convolutions, compression blocks [30], and dilated dense blocks, each MDilDenseLSTM generated a mask. Each band was given a different set of hyperparameters, with low-frequency bands receiving more attention. Speech signals were combined with noise and music signals from multiple databases, including 115 sounds from database (DB) [82], ESC-50 [70], NOISEX-92 [71], WSJ1, and MUSDB.

6. Performance Analysis of MMDenseLSTM and Audio Spectrogram Transformer

Performance analysis is an empirical investigation of different metrics, like SDR, SIR, and SAR, and components associated with the task, and efficiency of a certain process. Performance analysis seeks to deliver practical insights for decision-making, optimization, and improvement through data gathering, statistical analysis, and interpretation.

DenseLSTM utilises the MUSDB18 dataset, which includes 50 songs for testing and 100 songs for training. The analysis focused on the prevailing methodology, where the four sources of each song, vocals, drums, bass, and others, were recorded in stereo at 44.1 kHz and made available as a blend. The network was trained to estimate the source spectrograms using STFT magnitude frames of the mixture as inputs. The LSTM layers played a critical role in capturing global modulations, particularly when positioned at lower scales in the upscaling process [60].

Additionally, the Audio Spectrogram Transformer (AST) was assessed using the challenging AudioSet dataset, which is a standard for the categorization of audio events with weak labels. AudioSet consists of over 2 million 10-second audio snippets organized into 527 sound categories. The results indicated that when both models are trained from scratch, splitting the audio spectrogram into 128x2 rectangular patches performs better than the conventional 16x16 square patches. However, 16x16 patches remain the best option because pretrained models for 128x2 patches are not available. The superior performance of the AST, an attention-based model, demonstrates that CNNs are not necessary for audio differentiation [77].

Table 6 presents the performance evaluations of different music separation models. Takahashi's DenseLSTM model achieved an SDR of 4.77 for vocals and 15.80 for accompaniment, along with SIR and SAR metrics. Gong's Audio Spectrogram Transformer model, evaluated on the AudioSet dataset, achieved significantly higher SDR, SIR and SAR scores for vocals, showcasing advancements in music separation technology.

This significant enhancement in performance can be attributed to the AST's attention-based architecture, which excels in capturing and distinguishing intricate patterns in audio data compared to the convolutional approach used in MMDenseLSTM. The values provided above have been assessed by us and are not directly copied from the referenced papers.

7. Rigorous Methodologies for Audio Separation

To ensure that the comparison between conventional and avant-garde techniques for audio separation is scientifically rigorous and reproducible, a detailed explanation of the experimental setup and data used is crucial. The setup begins with the selection of appropriate datasets that contain multi-instrumental recordings where different musical instruments are played simultaneously. Commonly used datasets for this task include MUSDB18, which provides 150 tracks with individual stems (vocals, bass, drums, etc.), and DSD100, which offers 100 full-length tracks with ground truth isolated

tracks. These datasets ensure diversity in audio signals, making them suitable for testing the capabilities of both traditional and advanced models. Clearly, specifying the version of the dataset and any preprocessing steps applied is vital for reproducibility.

Before applying the separation techniques, preprocessing steps transform the raw audio data into a format suitable for analysis. This typically involves resampling all audio files to a consistent sample rate, such as 44.1 kHz, and, if necessary, converting stereo audio to mono. A crucial step for most models is applying the STFT, which converts the time-domain signal into a time-frequency representation (spectrogram). The parameters for STFT, such as window length, hop size, and frequency bin resolution, must be specified, as they directly impact the input to both traditional and deep learning models. Ensuring consistency in preprocessing allows for a fair comparison across different techniques.

To evaluate the performance of the audio separation methods, several widely accepted evaluation metrics are used. The most common is SDR, which measures the overall quality of the separated signal. SIR evaluates how well each sound source is separated from interfering sources, and SAR quantifies the presence of artifacts introduced by the separation process. These metrics are computed using the BSS Eval toolkit. For deep learning models, additional metrics such as MSE or MAE between the predicted and ground truth spectrograms can be used during training to track model performance.

The implementation of conventional methods in the experiment involves techniques like NMF, which decomposes the magnitude spectrogram into non-negative components that represent individual sound sources. Important parameters like the rank (number of components) and initialization method should be detailed. Similarly, for ICA, which separates sources based on statistical independence, details of the optimization method (e.g., FastICA) and the assumed number of sources should be provided. Methods like PCA, ILRMA, and IDLMA also need to be clearly described in terms of how they were implemented and tuned.

For the avant-garde techniques, deep learning models are applied to handle the complexities of audio separation. CNN and RNN are typically used to process spectrograms and extract spatial and temporal features. The architecture details, including the number of layers, kernel sizes, and pooling strategies, should be provided. The study also examines more sophisticated models like Dense-U-Net, Wave-U-Net, and HR-LSTM, which combine convolutional and recurrent layers to enhance feature extraction. Describing the model architectures, such as the depth of layers, activation functions, and any regularization techniques (e.g., dropout), is crucial for replicating the results. Finally, the training setup for the deep learning models must be thoroughly described. This includes specifying the loss function, which could be MSE or a spectrogram-based loss, and the optimizer (e.g., Adam, RMSprop) with associated hyperparameters like learning rate and momentum. Details about the batch size, the number of epochs, and the computational resources (e.g., GPU type, memory) used for training are essential to ensure that others can replicate the deep learning models' performance. Additionally, the testing protocol, including the cross-validation method or train-test split ratio, should be clearly documented. This ensures that the results of both conventional and avant-garde techniques are evaluated consistently.

By including these detailed steps, the study would not only offer a robust comparison between conventional and avant-garde approaches for audio separation but also ensure that future researchers can replicate the experimental results. The combination of well-documented datasets, preprocessing, model training, and evaluation procedures strengthens the scientific validity of the study, allowing for deeper insights into the performance of each method.

7.1. Discussion

This study focuses on comparing conventional and avant-garde techniques for audio separation, aiming to isolate and extract individual musical tones. Conventional methods, such as NMF, IDLMA, ILRMA, ICA, and PCA, are well-established in signal processing. These approaches rely on mathematical and statistical principles like matrix factorization, independence, and dimensionality reduction. They have been widely used in tasks involving audio decomposition, providing a solid foundation for separating mixed signals into their respective components.

In contrast, avant-garde techniques leverage recent advancements in deep learning and artificial intelligence to address the complexities of audio separation. Models like CNN and RNN, which are adept at handling spatial and temporal data, have been adapted to analyze audio spectrograms and extract individual sound sources more efficiently. The study also highlights advanced frameworks such as HR-LSTM, Dense-U-Net, and Wave-U-Net, which combine different architectures to enhance feature extraction and achieve better separation results.

By comparing these two sets of methodologies, the study offers insights into the strengths and limitations of each approach. Conventional methods provide a clear, interpretable foundation but may struggle with complex audio signals, while avant-garde models are more adaptable and powerful, though they often require more computational resources and data. The paper suggests that coupled architectures, which blend multiple deep learning techniques, offer a path toward more efficient and effective audio separation. Overall, this exploration benefits a range of audiences, including researchers, educators, musicians, and composers. It provides a clearer understanding of how traditional and advanced

techniques can be used to separate instrumental sounds. It also emphasizes the potential of deep learning frameworks to push the boundaries of audio processing, offering enhanced practical applications in the field. Instrumental acoustics separation techniques face significant challenges when applied to real-world scenarios due to the complexity of acoustic environments, computational demands, and the variability in musical structures. One key issue is the mismatch between pre-trained bases and real-world sounds, particularly affecting traditional supervised methods like NMF, which struggle with reduced accuracy when applied outside controlled environments. Moreover, techniques like BSS suffer from high computational complexity, making real-time processing difficult, especially when estimating numerous acoustic parameters. Real-world environments also introduce challenges such as reverberation and background noise, complicating the separation process. Additionally, the availability and quality of labeled training data are often insufficient for data-intensive approaches like deep learning, which further hinders performance. Real-time processing requirements and the variability in the number and type of sound sources, particularly in polyphonic music, add to the difficulty of designing universally applicable solutions. Evaluating the quality of separation remains complex due to a lack of standardized metrics, making it hard to compare different methods. Finally, incorporating additional information, such as instrument labels or visual data, can improve separation but introduce further complexity.

Addressing these challenges requires advancements in algorithm efficiency, more robust evaluation metrics, and innovative methods capable of handling the variability and intricacies of real-world audio scenarios.

8. Conclusion and Future work

This comprehensive exploration of audio separation methodologies bridges avant-garde and conventional approaches, providing valuable insights for researchers, educators, musicians, and composers. By comparing traditional methods like NMF, IDLMA, ILRMA, ICA, and PCA with advanced machine learning techniques, including reinforcement, supervised, semi-supervised, and unsupervised Learning, the study illuminates the landscape of instrumental acoustics separation. Deep learning models, particularly CNNs and RNNs, showcase promising capabilities in dissecting instrumental acoustics, as evidenced by the analysis of coupled frameworks such as HR-LSTM, Dense-U-Net, Wave-U-Net, conv-tasnet, Res-U-Net, and LRCN. Furthermore, the evaluation of combined architectures like DenseLSTM and Audio Spectrogram Transformer underscores their efficiency over individual models. As a comprehensive resource, this paper not only enriches audio processing research but also offers practical applications. Future endeavors could focus on refining hybrid architectures, exploring novel deep learning techniques, and addressing real-world implementation challenges, thereby advancing the field of audio separation towards greater precision and usability.

References

[1] Mikkel N. Schmidt, and Morten Mørup, "Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation," *Independent Component Analysis and Blind Signal Separation, Lecture Notes in Computer Science*, vol. 3889, pp. 700-707, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[2] Daniel Lee, and H. Sebastian Seung, "Algorithms for Non-Negative Matrix Factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 1-7, 2000. [Google Scholar] [Publisher Link]

[3] Paris Smaragdis, "Non-Negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs," *Independent Component Analysis and Blind Signal Separation, Lecture Notes in Computer Science*, vol. 3195, pp. 494-499, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[4] Alexey Ozerov, and Cédric Fevotte, "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550-563, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[5] Seokjin Lee, Sang Ha Park, and Koeng-Mo Sung, "Beamspace-Domain Multichannel Nonnegative Matrix Factorization for Audio Source Separation," *IEEE Signal Processing Letters*, vol. 19, no. 1, pp. 43-46, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[6] Jianyu Wang, and Shanzheng Guan, "Multichannel Blind Speech Source Separation with a Disjoint Constraint Source Model," *Arxiv*, pp. 1-5, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Naoki Makishima et al., "Independent Deeply Learned Matrix Analysis for Determined Audio Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1601-1615, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Takuya Hasumi et al., "Multichannel Audio Source Separation with Independent Deeply Learned Matrix Analysis Using Product of Source Models," *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Tokyo, Japan, pp. 1226-1233, 2021. [Google Scholar] [Publisher Link]

[9] Takuya Hasumi et al., "PoP-IDLMA: Product-of-Prior Independent Deeply Learned Matrix Analysis for Multichannel Music Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2680-2694, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Shinichi Mogami et al., "Independent Low-Rank Matrix Analysis Based on Complex Student's t-Distribution for Blind Audio Source Separation," *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing*, Tokyo, Japan, pp. 1-6, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[11] Daichi Kitamura, and Kohei Yatabe, "Consistent Independent Low-Rank Matrix Analysis for Determined Blind Source Separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2020, pp. 1-35, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Zahoor Uddin, Aamir Qamar, and Farooq Alam, "ICA Based Sensors Fault Diagnosis: An Audio Separation Application," *Wireless Personal Communications*, vol. 118, pp. 3369-3384, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] M.R. Ezilarasan et al., *Blind Source Separation in the Presence of AWGN Using ICA-FFT Algorithms a Machine Learning Process*, 1st ed., Recent Trends in Computational Intelligence and its Application, CRC Press, pp. 1-9, 2023. [Google Scholar] [Publisher Link]

[14] Zaineb H. ibrahemm, and Ammar I. Shihab, "Voice Separation and Recognition Using Machine Learning and Deep Learning a Review Paper," *Journal of Al-Qadisiyah for Computer Science and Mathematics*, vol. 15, no. 3, pp. 11-34, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Harshada Burute, and P.B. Mane, "Separation of Singing Voice from Music Background," *International Journal of Computer Applications*, vol. 129, no. 4, pp. 22-26, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[16] Tomohiro Watanabe, Takanori Fujisawa, and Masaaki Ikehara, "Vocal Separation Using Improved Robust Principal Component Analysis and Post-Processing," *2016 IEEE 59th International Midwest Symposium on Circuits and Systems*, Abu Dhabi, United Arab Emirates, pp. 1-4, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[17] Feng Li, Yujun Hu, and Lingling Wang, "Unsupervised Single-Channel Singing Voice Separation with Weighted Robust Principal Component Analysis Based on Gammatone Auditory Filterbank and Vocal Activity Detection," *Sensors*, vol. 23, no. 6, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Shrikant Venkataramani, Efthymios Tzinis, and Paris Smaragdis, "End-To-End Non-Negative Autoencoders for Sound Source Separation," *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, pp. 116-120, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[19] Pankaj Ramakant Kunekar et al., "Audio Feature Extraction: Foreground and Background Audio Separation Using KNN Algorithm," *International Journal of Science and Research Archive*, vol. 9, no. 1, pp. 269-276, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Kilian Schulze-Forster et al., "Unsupervised Music Source Separation Using Differentiable Parametric Source Models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1276-1289, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Gaël Richard, Pierre Chouteau, and Bernardo Torres, "A Fully Differentiable Model for Unsupervised Singing Voice Separation," *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, Korea, pp. 946-950, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Sangeun Kum et al., "Semi-Supervised Learning Using Teacher-Student Models for Vocal Melody Extraction," *Arxiv*, pp. 1-8, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Zhepei Wang et al., "Semi-Supervised Singing Voice Separation With Noisy Self-Training," *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, ON, Canada, pp. 31-35, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[24] Hazem Toutounji et al., "Learning the Sound Inventory of a Complex Vocal Skill via an Intrinsic Reward," *Science Advances*, vol. 10, no. 13, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[25] Yu Wang et al., "Few-Shot Musical Source Separation," *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, pp. 121-125, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[26] Aakanksha Desai et al., "Targeted Voice Separation," *International Journal of Innovative Science and Research Technology*, vol. 7, no. 10, pp. 947-950, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[27] Samiul Basir et al., "U-NET: A Supervised Approach for Monaural Source Separation," *Arabian Journal for Science and Engineering*, vol. 49, pp. 12679-12691, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Tom Le Paine et al., "Fast Wavenet Generation Algorithm," *Arxiv*, pp. 1-6, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[29] Dario Rethage, Jordi Pons, and Xavier Serra, "A Wavenet for Speech Denoising," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, pp. 5069-5073, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[30] Gao Huang et al., "Densely Connected Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognitio*, Honolulu, HI, USA, pp. 2261-2269, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[31] Naoya Takahashi, and Yuki Mitsufuji, "Multi-Scale Multi-Band Densenets for Audio Source Separation," *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, pp. 21-25, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[32] Abhimanyu Sahai, Romann Weber, and Brian McWilliams, "Spectrogram Feature Losses for Music Source Separation," *2019 27th European Signal Processing Conference*, A Coruna, Spain, pp. 1-5, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[33] Woon-Haeng Heo, Hyemi Kim, and Oh-Wook Kwon, "Source Separation Using Dilated Time-Frequency DenseNet for Music Identification in Broadcast Contents," *Applied Sciences*, vol. 10, no. 5, pp. 1-18, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[34] Guoqing Li et al., "Efficient Densely Connected Convolutional Neural Networks," *Pattern Recognition*, vol. 109, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[35] Cem Subakan et al., "Attention is All You Need in Speech Separation," *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, ON, Canada, pp. 21-25, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[36] Simon Rouard, Francisco Massa, and Alexandre Défossez, "Hybrid Transformers for Music Source Separation," *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, pp. 1-5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[37] Jiale Qian et al., "Stripe-Transformer: Deep Stripe Feature Learning for Music Source Separation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[38] Yongwei Gao, Xulong Zhang, and Wei Li, "Vocal Melody Extraction via HRNet-Based Singing Voice Separation and Encoder-Decoder-Based F0 Estimation," *Electronics*, vol. 10, no. 3, pp. 1-14, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[39] Jingdong Wang et al., "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349-3364, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[40] Sasha Targ, Diogo Almeida, and Kevin Lyman, "Resnet in Resnet: Generalizing Residual Architectures," *Arxiv*, pp. 1-7, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[41] Gino Brunner et al., "Monaural Music Source Separation Using a ResNet Latent Separator Network," *2019 IEEE 31st International Conference on Tools with Artificial Intelligence*, Portland, OR, USA, pp. 1124-1131, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[42] Tsubasa Ochiai et al., "Beam-TasNet: Time-domain Audio Separation Network Meets Frequency-Domain Beamformer," *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, pp. 6384-6388, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[43] Alfian Wijayakusuma et al., "Implementation of Real-Time Speech Separation Model Using Time-Domain Audio Separation Network (TasNet) and Dual-Path Recurrent Neural Network (DPRNN)," *Procedia Computer Science*, vol. 179, pp. 762-772, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[44] Satvik Venkatesh et al., "Real-Time Low-Latency Music Source Separation Using Hybrid Spectrogram-Tasnet," *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, Korea, pp. 611-615, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[45] Vanshaj Agrawal, and Sunil Karamchandani, "Audio Source Separation as Applied to Vocals-Accompaniment Extraction," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 5, pp. 1-8, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[46] Rawad Melhem, Assef Jafar, and Riad Hamadeh, "Improving Deep Attractor Network by BGRU and GMM for Speech Separation," *Journal of Harbin Institue of Technology*, vol. 28, no. 3, pp. 90-96, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[47] Bhuwan Bhattarai et al., "High-Resolution Representation Learning and Recurrent Neural Network for Singing Voice Separation," *Circuits, Systems, and Signal Processing*, vol. 42, pp. 1083-1104, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[48] Yi Luo, and Rongzhi Gu, "Improving Music Source Separation with Simo Stereo Band-Split Rnn," *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, Korea, pp. 426-430, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[49] Xulong Zhang et al., "Research on Singing Voice Detection Based on a Long-Term Recurrent Convolutional Network with Vocal Separation and Temporal Smoothing," *Electronics*, vol. 9, no. 9, pp. 1-23, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[50] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," *Arxiv*, pp. 1-7, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[51] Alice Cohen-Hadria, Axel Roebel, and Geoffroy Peeters, "Improving Singing Voice Separation Using Deep U-Net and Wave-U-Net with Data Augmentation," *2019 27th European Signal Processing Conference*, A Coruna, Spain, pp. 1-5, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[52] Jung-Hee Kim, and Joon-Hyuk Chang, "Attention Wave-U-Net for Acoustic Echo Cancellation," *Interspeech*, Shanghai, China, pp. 3969-3973, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[53] Tomohiko Nakamura, and Hiroshi Saruwatari, "Time-Domain Audio Source Separation Based on Wave-U-Net Combined with Discrete Wavelet Transform," *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, pp. 386-390, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[54] Vasiliy Kuzmin et al., "Real-time Streaming Wave-U-Net with Temporal Convolutions for Multichannel Speech Enhancement," *Arxiv*, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[55] Yuzhou Liu et al., "Voice and Accompaniment Separation in Music Using Self-Attention Convolutional Neural Network," *Arxiv*, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[56] Shun Takeda, and Shuichi Arai, "Music Source Separation Using Deform-Conv Dense U-Net," *2021 3rd International Conference on Cybernetics and Intelligent System*, Makasar, Indonesia, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[57] Bahareh Tolooshams et al., "Channel-Attention Dense U-Net for Multichannel Speech Enhancement," *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, pp. 836-840, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[58] Thomas Sgouros, Angelos Bousis, and Nikolaos Mitianoudis, "An Efficient Short-Time Discrete Cosine Transform and Attentive MultiResUNet Framework for Music Source Separation," *IEEE Access*, vol. 10, pp. 119448-119459, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[59] DaDong Wang, Jie Wang, and MingChen Sun, "3 Directional Inception-ResUNet: Deep Spatial Feature Learning for Multichannel Singing Voice Separation with Distortion," *Plos One*, vol. 19, no. 1, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[60] Naoya Takahashi, Nabarun Goswami, and Yuki Mitsufuji, "Mmdenselstm: An Efficient Combination of Convolutional and Recurrent Neural Networks for Audio Source Separation," *2018 16th International Workshop on Acoustic Signal Enhancement*, Tokyo, Japan, pp. 106-110, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[61] Woon-Haeng Heo, Hyemi Kim, and Oh-Wook Kwon, "Integrating Dilated Convolution into DenseLSTM for Audio Source Separation," *Applied Sciences*, vol. 11, no. 2, pp. 1-19, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[62] Yi Luo, and Nima Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256-1266, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[63] Berkan Kadıoğlu et al., "An Empirical Study of Conv-Tasnet," *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, pp. 7264-7268, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[64] Alexandre Défossez et al., "Music Source Separation in the Waveform Domain," *Arxiv*, pp. 1-16, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[65] Xiaoman Qiao et al., "VAT-SNet: A Convolutional Music-Separation Network Based on Vocal and Accompaniment Time-Domain Features," *Electronics*, vol. 11, no. 24, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[66] Umut Isik et al., "PoCoNet: Better Speech Enhancement with Frequency-Positional Embeddings, Semi-Supervised Conversational Data, and Biased Loss," *Arxiv*, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[67] Florian Strub et al., "FiLM: Visual Reasoning with a General Conditioning Layer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, pp. 3942-3951, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[68] Andreas Jansson et al., "Singing Voice Separation with Deep U-Net Convolutional Networks," *18th International Society for Music Information Retrieval Conference*, Suzhou, China, pp. 745-751, 2017. [Google Scholar] [Publisher Link]

[69] Aäron van den Oord et al., "WaveNet: A Generative Model for Raw Audio," *Arxiv*, pp. 1-15, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[70] Karol J. Piczak, "ESC: Dataset for Environmental Sound Classification," *Proceedings of the 23rd ACM international Conference on Multimedia*, New York, United States, pp. 1015-1018, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[71] Andrew Varga, and Herman J.M. Steeneken, "Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Speech Communication*, vol. 12, no. 3, pp. 247-251, 1993. [CrossRef] [Google Scholar] [Publisher Link]

[72] Jiaqi Gu et al., "Multi-Scale High-Resolution Vision Transformer for Semantic Segmentation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 12084-12093, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[73] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-Excitation Networks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City*, UT, USA, pp. 7132-7141, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[74] Yi Luo et al., "Rethinking The Separation Layers In Speech Separation Networks," *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, ON, Canada, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[75] Kaiming He et al., "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 770-778, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[76] Mehrez Souden, Jacob Benesty, and SofiÈne Affes, "On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260-276, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[77] Yuan Gong, Yu-An Chung, and James Glass, "AST: Audio Spectrogram Transformer," *Arxiv*, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[78] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," *Arxiv*, pp. 1-14, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[79] Kaiming He et al., "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," *2015 IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 1026-1034, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[80] Ke Tan, Jitong Chen, and DeLiang Wang, "Gated Residual Networks with Dilated Convolutions for Supervised Speech Separation," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, pp. 21-25, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[81] Siddique Latif et al., "A Survey on Deep Reinforcement Learning for Audio-Based Applications," *Artificial Intelligence Review*, vol. 56, pp. 2193-2240, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[82] Yong Xu et al., "Multi-Objective Learning and Mask-Based Post-Processing for Deep Neural Network Based Speech Enhancement," *Arxiv*, pp. 1-5, 2017. [CrossRef] [Google Scholar] [Publisher Link]