International Journal of Communication Networks and Information Security
ISSN: 2073-607X, 2076-0930
Volume 15 Issue 02 Year 2023

Enhancing the Performance of Single-Channel Blind Source Separation by Using ConvTransFormer

Santosh Kumar S*
Research Scholar, School of ECE, REVA University, Bengaluru, Department of ECE, Sri Venkateshwara College of Engineering, Bengaluru, India
reachsun@gmail.com
Bharathi S H
Professor, School of ECE, REVA University, Bengaluru, India
bharathish@reva.edu.in

Article History

Received: 15 September 2023
Revised: 13 October 2023
Accepted: 11 November 2023

Abstract

In the specialized field of audio signal processing, this study introduces a pioneering ConvTransFormer architecture aimed at enhancing the performance of single-channel blind source separation (SCBSS). This innovative architecture ingeniously combines the strengths of a multiple simple-weak attention mechanism with the triple-gating feature of a Gated Attention Unit (GAU) within the ConvTransFormer. This combination allows for a more focused and effective targeting of specific segments within the input sequence. The efficacy of this ConvTransFormer architecture is rigorously evaluated using the WSJ0-2mix dataset, a standard benchmark in the field. The results of this evaluation are significant, demonstrating substantial improvements in key performance metrics. Notably, there is an increase in the Signal-to-Interference (SI)-Signal-to-Noise Ratio improvement (SNRi) by 16.5 and in the Signal-to-Distortion Ratio improvement (SDRi)-Signal-to-Interference (SDRi) by 16.8. These improvements are crucial indicators of the quality of source separation in SCBSS. The findings of this research are groundbreaking, indicating that the proposed ConvTransFormer architecture surpasses existing methods in both SI-SNRi and SDRi performance metrics. This advancement marks a significant step forward in the field of SCBSS, offering new avenues for more effective and precise audio signal processing, especially in scenarios where isolating individual sound sources from a single-channel input is essential.

Keywords: ConvTransFormer, Speech Separation, Gating Mechanism,Single Channel Source Separation

1. Introduction

Single-Channel Blind Source Separation (SCBSS) represents a technique used for isolating distinct signals within a signal mixture, all accomplished without requiring any prior knowledge about the original sources. We are working on the assumption that the observed signal is a linear mixture of several independent sources; the unmixing matrix that can distinguish between them is what we are looking for. A Single-channel recording typically includes several separate speech signals coming from different places, such as a conversation between numerous people or a single

person speaking in an environment with a lot of background noise. The purpose of speech separation is to isolate and improve upon intended speech signals while cancelling out or reducing noise. The time-domain method for speech separation requires working directly with the audio signal's waveform. This stands in contrast to frequency-domain methodologies like the Short-Time Fourier Transform (STFT), which operates based on the spectral representation of the signal. Due to their capacity to better manage non-stationary settings and preserve the temporal properties of the speech signal, time-domain approaches have become increasingly prominent.

The concept of source-filter models is the foundation of one of the more widespread approaches to time-domain single-channel speech separation. These hypotheses propose that the observed speech mixture can be represented by a model that combines the original signal with the output of a time-varying filter. The source signal can be approximated by first estimating the filter and then deconvolving the mixture. Time-Frequency domain (T-F) representation of the mixture signal, obtained by using the Short-Time Fourier Transform (STFT) [1] has been the predominant format for previous speech separation methods. This type of representation is also known as a spectrogram. In the T-F domain, speech separation techniques attempt to reconstruct the original spectra of the isolated sources from the resulting spectral mixture. The spectrogram representation of each source in the mixture can be approximated directly by using the non-linear regression techniques, these spectrograms of the clean sources serve as the training target. This is something that can be accomplished.

Deep clustering (DC) [2] introduces an innovative approach to address the complex permutation problem in audio signal processing. Utilizing a hybrid of Recurrent Neural Network (RNN) and Bi-Directional Long Short-Term Memory (BLSTM) architectures, DC assigns a unique embedding vector to each Time-Frequency (T-F) unit in a mixed spectrogram. These embeddings are precisely generated using the dot product method. During training, DC employs the concept of the ideal binary mask, representing the perfect speaker assignment, to measure the correlation between this theoretical assignment and the embedding vectors' affinity matrices, utilizing the Frobenius norm for assessment. A key strength of DC lies in its robustness against permutation challenges, attributed to the permutation invariance characteristic of the affinity matrices. The training process is designed to align embedding vectors of T-F units from the same source closer together, while distancing those from different sources. This strategic arrangement significantly boosts the model's ability to distinguish between various sources.

In parallel, the integration of Convolutional Neural Networks (CNNs) [1] has garnered considerable attention within the audio signal processing research community, particularly in Single-Channel audio source separation. Employing advanced deep learning techniques, this methodology facilitates the deconstruction of audio captured by a single microphone into its constituent components. Given that multiple instruments or voices are often amalgamated in a single-channel recording, the primary goal of source separation is to disentangle these sources from the inherent noise. Challenges arise in scenarios characterized by time and frequency domain overlap between sources, intensifying the complexity of the separation task.

A subset of deep neural networks, specifically convolutional neural networks (CNNs), has demonstrated encouraging outcomes in the realm of audio source separation tasks. CNNs excel in extracting valuable features from audio signals due to their adeptness in capturing local patterns and data dependencies. The fundamental concept underlying the utilization of CNNs in single audio channel source separation revolves around training the network to establish a mapping between the input mixture signal and various sources. This involves leveraging a substantial dataset of training examples, where the mixture waveform serves as input and the desired source waveform serves as the output. The CNN is intricately designed to learn the optimal prediction of sources with minimal error.

In the context of audio signal processing, the Time-Frequency (T-F) representation known as Short-Time Fourier Transform (STFT) serves as a prevalent input for CNNs tasked with separating distinct audio sources. The network undergoes a learning process to disentangle sources by scrutinizing the mixture's spectrogram and estimating individual spectrograms for each source. To obtain the separated waveforms, the spectrograms undergo inversion through the inverse Short-Time Fourier Transform (STFT). Diverse CNN architectures and training methodologies have been introduced for single audio channel source separation. U-Net architectures, encompassing both encoder and decoder components, have gained considerable traction for capturing both low-level

and high-level information in the audio stream. Additionally, techniques such as data augmentation, regularization, and adversarial training have been employed to enhance the network's separation performance and overall generalization.

The contemporary domain of single-channel blind source separation (SCBSS) research has made notable strides, with advancements introduced by methodologies like Deep Clustering (DC) and Convolutional Neural Networks (CNNs). Nevertheless, a discernible research gap persists in effectively tackling the inherent challenges of precisely isolating individual sources within a single-channel audio recording. Current approaches, encompassing DC and CNNs, might encounter limitations in optimizing Signal-to-Interference (SI), Signal-to-Noise (SNR), and Signal-to-Distortion (SDR) ratios, particularly in intricate audio settings characterized by source overlap and varying background noise levels. Efforts to bridge this research gap are essential for enhancing the robustness and effectiveness of SCBSS methods in diverse and challenging audio scenarios.

Moreover, the need for improved generalization across diverse datasets and real-world scenarios poses a challenge for existing SCBSS techniques. The proposed ConvTransFormer architecture seeks to fill this gap by introducing a novel combination of simple-weak attention mechanisms and a triple-gating structure in the ConvTransFormer's gated attention unit (GAU). This innovative approach aims to provide a more selective and focused means of separating sources in challenging audio recordings.

2. Related Works

Convolutional Neural Networks (CNNs) have emerged as a powerful tool in various audio processing domains, proving effective in tasks such as voice recognition, speech enhancement, audio tagging, and a broad spectrum of music-related applications [3] [4]. Within the CNN family, Convolutional Denoising Autoencoders (CDAEs) represent a subtype designed to identify robust, low-dimensional recurrent patterns in input signals. An advantage of CDAEs lies in their utilization of shared parameters, resulting in a lower total parameter count compared to traditional Denoising Autoencoders (DAEs) [5]. The inherent capability of CDAEs to recognize recurrent patterns makes them particularly well-suited for applications like extracting speech signals from noisy environments and enhancing music signals to facilitate speech augmentation and recognition.

While CNNs have demonstrated prowess in diverse audio tasks, recent investigations have delved into the application of deep learning for time-domain audio separation [6]. Various systems explored share a fundamental approach, leveraging a data-driven representation that is jointly optimized through an end-to-end training paradigm. This optimization plays a crucial role in replicating the functionality of Short-Time Fourier Transform (STFT) during the feature extraction process. As an alternative to STFT and its inverse, explicit generation of these representations has been proposed. One notable approach involves the explicit inclusion of feature extraction and separation within the network design, exemplified by end-to-end Convolutional Neural Networks (CNNs) [7]. Each of these methods employs a distinct set of waveform properties and building blocks within the separation module.

In a specific study [6], front-end processing is facilitated by a convolutional encoder modelled after the discrete cosine transform (DCT). Subsequently, the encoded features are input into a Multilayer Perceptron (MLP), responsible for separating the input signals. By reversing the encoder's action, the original waveforms can be accurately reconstructed. Conversely, another approach, as demonstrated in [5], integrates the separation process into a U-Net 1-D CNN architecture without explicitly transforming the input into a representation analogous to a spectrogram. These diverse methodologies showcase the evolving landscape of time-domain audio separation, with a focus on leveraging the capabilities of CNNs for improved signal processing. The exploration of explicit representations, shared parameters, and end-to-end optimization serves to push the boundaries of audio processing techniques, opening avenues for enhanced speech separation, noise reduction, and overall improved audio quality. The intricate interplay between waveform properties and specialized building blocks in these approaches underscores the nuanced nature of effective single-channel blind source separation in complex audio environments.

The challenge of speaker separation has seen recent advancements through deep learning methodologies. Deep Neural Networks (DNNs) play a pivotal role in predicting Time-Frequency (T-F) masks or spectra for two speakers within a mixture [4]. Typically, a DNN designed for this purpose has dual output layers, each dedicated to one of the speakers. These analyses assume a constant pairing of speakers during both training and testing phases. Notably, individuals with

hearing impairments derive significant benefits from talker-dependent training [8]. However, the limitation arises when this type of training, contingent on specific speakers, fails to generalize to untrained voices, presenting what is commonly known as the permutation problem [2]. Resolving this problem is crucial for achieving talker-independent speaker separation.

On the flip side, the extent to which these techniques can be applied to extensive speech corpora, exemplified by the benchmark in [2], remains uncertain. Another noteworthy approach in this domain is the time-domain audio separation network, colloquially known as TasNet [9]. TasNet adopts a convolutional encoder-decoder architecture to reconstruct the mixed waveform. This architecture includes a restricted-output encoder and an inverted-output linear decoder. The framework draws parallels with the Independent Component Analysis (ICA) method [10] and the semi-nonnegative matrix factorization (semi-NMF) method [4]. In scenarios featuring a non-negative mixing matrix, the decoder's parameters serve as the basis signals. TasNet's source separation process involves determining a weighting function for each source at every time step of the encoder's output, resembling a time-frequency masking approach. Studies consistently showcase TasNet's superior or comparable performance in comparison to preceding time-frequency domain systems. This underscores its efficacy in advancing the state-of-the-art in audio source separation techniques.

The research landscape reveals two critical gaps in prior investigations. Initially, prevailing studies predominantly emphasize speaker separation using deep neural networks (DNNs) with a focus on talker-dependent training, especially beneficial for those with hearing impairments [8]. However, these models encounter challenges in extending their effectiveness to untrained voices due to the inherent permutation problem [2]. Thus, there is a discernible need for innovative methodologies that surmount the limitations associated with talker-dependent training, fostering more adaptable and universally applicable talker-independent speaker separation.

Secondly, while strides have been made in time-domain audio separation networks like TasNet [9], their potential across extensive speech corpora, as outlined by the benchmark in [2], remains uncertain. The research acknowledges this uncertainty, prompting further exploration into the scalability and adaptability of these techniques across diverse datasets. Addressing this gap entails scrutinizing the real-world performance of proposed methodologies under varying conditions, encompassing different speakers, acoustic environments, and speech contexts. This pursuit aims to offer a more nuanced comprehension of the practical efficacy and limitations of advanced time-domain audio separation networks on a larger scale. In essence, the research gap involves fortifying the talker-independent speaker separation models and elucidating the broader application spectrum of sophisticated time-domain audio separation networks.

In this paper, we have presented speech separation utilizing a single channel by employing the ConvTransFormer model, which is displayed in Figure 1. You should suggest adding a gated attention unit (GAU) to the ConvTransFormer so that the model can concentrate its attention on certain elements of the input sequence. Convolution modules, scale and offset operations, gating activities, and a multiple simple-weak attention mechanism are all part of the ConvTransFormer block. To complete the remaining processing, we make use of the convolution module (Convo), which can be seen in Figure 4.

3. Methodology

The Convolutional Time-Domain Single Channel Source Separation Network (CTD-SCSSN) represents a deep learning architecture grounded in convolutional neural networks (CNNs). It is designed to perform blind source separation on single-channel speech inputs. Employing convolutional layers, the CTD-SCSSN extracts features from the time-domain input signal. Additionally, a separation module is incorporated to discern the distinct sources. Encoder, Separation Module, and Decoder are the three main components that make up the CTD-SCSSN paradigm, which may be broken down into their functions as in Figure 1. Multiple convolutional layers make up the encoder, and these layers are responsible for extracting features from the input signal. As input, the encoder stage receives brief portions of the mixing waveform, which it then converts into the equivalent representations in an intermediate feature space.

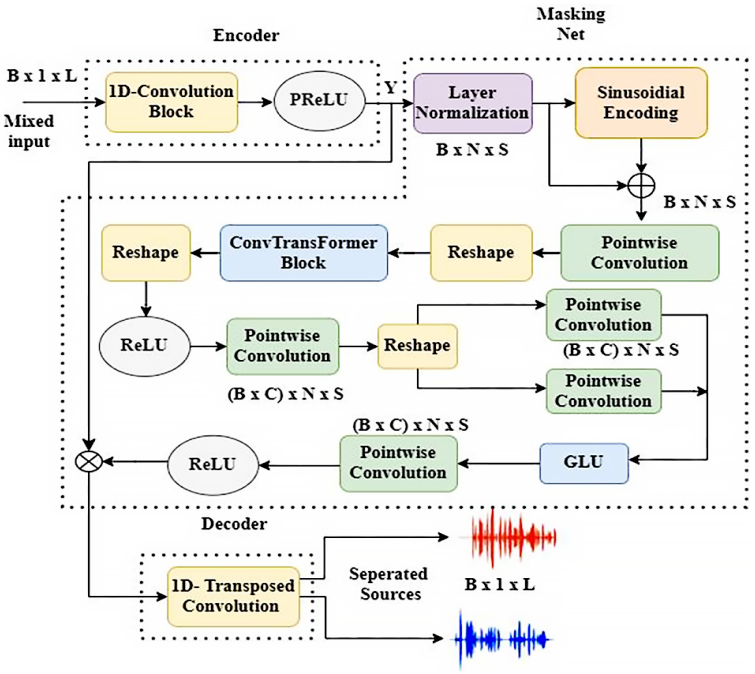

Figure 1. ConvTransFormer Architecture

To distinguish the different sources from the combined input signal, the separation module is put into operation. The intermediate features that were produced by the encoder stage are then used by the separation stage, which estimates a multiplicative function (mask) for each source at each time step using those intermediate features. In most cases, this is accomplished by utilizing yet another neural network, specifically a convolutional network. The mask is essentially a filter that specifies how much of each source should be present at each time step. It does this by indicating how much of each source should be present. One way to think about the separation stage is as a means of physically dividing the sources up into different parts of the feature space.

The decoder is responsible for recreating the individual sources based on the features that have been separated. The masked encoder characteristics that were produced by the separation step are brought into the decoder stage, where they are transformed back into the time domain so that the source waveforms can be reconstructed. This is often accomplished through the utilization of a network architecture that is analogous to that of the encoder stage but in reverse. The decoder step contributes to the transformation of the sources that have been separated back into the time domain, which ultimately results in the production of the final output waveforms. Figure 2 presents an overall block diagram of the encoder and decoder for your viewing pleasure.

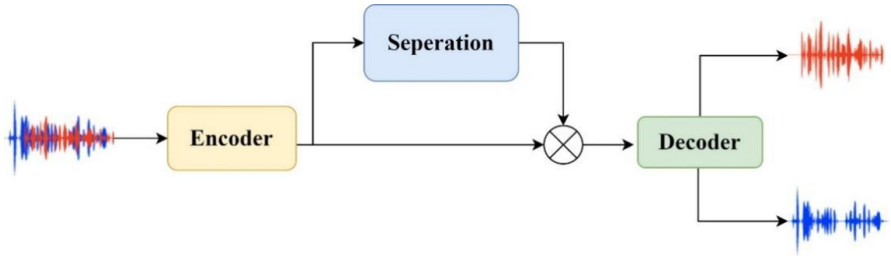

Figure 2. Encoder-decoder for the Speech Separation

3.1 Speech Separation in Time-domain

The challenge of separating speech on a single channel can, in fact, be framed in terms of determining where C sources are coming from $s_1(t), s_2(t), \ldots, s_c(t) \in R^{1xT}$ from a single mixed audio signal x(t). This problem is challenging because the mixed signal contains a combination of all the source signals, which can be difficult to separate without any additional information.

In this formulation, the goal is to estimate the source signals sources $s_1(t), s_2(t), \ldots, s_c(t)$ such that they add up to the observed mixed signal x(t):

$$x(t) = \sum_{i=0}^{C} S_i(t) \quad (1)$$

3.2 Convolutional Encoder-Decoder

In the source separation, an encoder consists of the 1-dimensional Convolution layer and Parametric rectified layer unit(PReLU) for the extraction of the signal features and converts the encoded data to the Non-Negative values. The input mixture signal is divided into the overlapping segments of length N, represented by $x_i \in R^{1*N}$, where i = 1…… L denotes the segment index and L denotes the total number of segments in the input x_i is transformed into a N- dimension signal, $w \in R^{1*N}$ by a convolution operation, N is the total number of segments, followed by a PReLU, the encoded output for the input sequence $x \in R^{1*N}$ is Y given by:

$$Y = PReLU\left(convb(x)\right) \quad (2)$$

Where the mathematical representation of the PReLU is given by:

$$f(y_i) = \begin{cases} y_i, & if\ y_i > 0 \\ \alpha_i y_i, & if\ y_i \leq 0 \end{cases} \quad (3)$$

After subjecting the sequence Y to element-wise multiplication by individual signal mask, the resulting extracted feature signal is then decoded. In order to recover the original waveform from this representation, the decoder performs a transposed convolution operation in dimension one, as expressed by the matrix multiplication :

$$\hat{x} = wV \quad (4)$$

3.3 Masking Separation Net

The Masking Net takes as input which is the encoded sequence and performs a non-linear mapping to produce a set of masks that indicate which parts of the input sequence should be masked or ignored. These masks are then used during the decoding process to prevent the decoder from attending to certain parts of the input sequence that may be irrelevant or misleading.

Normalization is performed on the output of the encoder Y, after which the positional encodings are added. The resultant is then subjected to pointwise convolution, reshaped, and then sequentially processed. Following the processing done by the convolution modules, the sequence goes via the gates. In the Convolutional modules, first, the signal is subjected to the linear projections, maps the input features to a dimensionality of BxSx2N then followed by Depth Wise Convolution in which the input channels are processed independently by separate filters on the sequence as part of the processing. The attentive gating mechanism joins local attention (U, V) and Global Self-Attention followed by the gating operation. The ConvTransFormer block is responsible for simply learning residual parameters, and it uses a skip connection that was provided by the input to make training easier. The currently active block output is used as an input for the ConvTransFormer block that comes after. The operation that takes place is repeated R times within the ConvTransFormer.

The ConvTransFormer module output is subjected to processing by a PReLU, then it is processed by the pointwise linear convolution, where the sequence dimension is increased. After that, data is processed by another type of convolution called parallel pointwise convolution and a Gated Linear Units(GLu) process. Finally, the sequence is forwarded through point-wise convolution again, and then it is followed by a PReLU to generate the masking sequences $M \in R^{C*N*S}$. This sequence M_i is reconstructed for every unique channel M_i, and further forwarded to the decoder to process in its own distinct fashion.

3.4 ConvTransFormer Module

The sequence modelling deep learning model that makes use of our suggested ConvTransFormer block is seen in Figure 3 below. Built upon the gated attention unit (GAU), this approach enables the model to concentrate its attention on particular segments of the input sequence. The ConvTransFormer block integrates convolution modules, scale and offset operations, a simplistic multiple-weak attention mechanism, and gating operations. Within the ConvTransFormer block, the convolution modules are harnessed to detect local patterns within the input sequence, contributing to an improved grasp of the sequence's structure and interconnections among its segments. The scale

and offset operations are used to normalize the input before it is passed through the convolution modules.

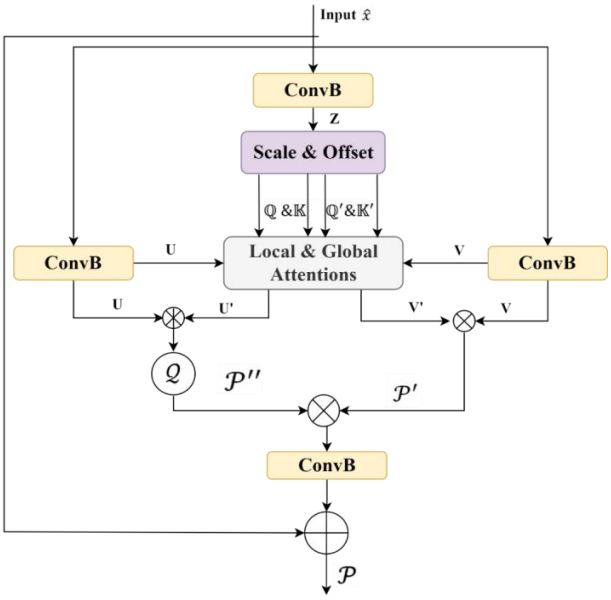

Figure 3. ConvTransFormer Block

The utilization of a multiple simple-weak attention mechanism facilitates targeted concentration on distinct segments of the input sequence. By leveraging this mechanism, the model gains the capacity to emphasize pertinent portions of the sequence and disregard extraneous ones. The integration of gating operations within the ConvTransFormer block regulates the information flow. Within the ConvTransFormer block's triple-gating structure, three gating mechanisms oversee information flow at various model stages. This augmentation enhances the model's efficacy in comprehending intricate interconnections among different elements of the sequence.

The triple-gating process in the ConvTransFormer block combines attention with gating to enhance the model's capability to capture complex relationships between different parts of the input sequence. The input sequence to the current ConvTransFormer block is denoted as $Y \in R^{S*N}$, where N is the length of the input sequence. The sequence is processed by the convolution module to obtain two sets of values: $U \in R^{S*2N}$ and $V \in R^{S*2N}$ where U and V are the output of the convolution block denoted as ConvB(), given by:

$$U = ConvB(Y),\ V = ConvB(Y) \qquad (5)$$

In a ConvB (Convolutional Block) layer with an expansion factor of 2, the feature dimensions would be expanded from N to 2N in the linear layer. This means that if the input to the layer has N feature dimensions, the output of the layer will have 2N feature dimensions. This is accomplished using a linear layer that has N input units and 2N output units. The weights of this linear layer are learned during training, and they determine how each input feature is mapped to its corresponding output features.

The attention matrix is denoted as $A_t \in \mathbb{R}^{S*S}$, and the output sequence is $\mathcal{P} \in \mathbb{R}^{S*N}$ of the ConvTransFormer are expressed as follows:

$$\mathcal{P}^{'} = \mathcal{Q}\left(U \odot V^{'}\right) \quad \text{where } V^{'} = A_t V \qquad (6)$$

$$\mathcal{P}^{''} = \mathcal{Q}\left(U^{'} \odot V\right) \quad \text{where } U^{'} = A_t U \qquad (7)$$

The output of the ConvTransFormer block is given by

$$\mathcal{P} = Y + ConvB\left(\mathcal{P}^{'} \odot \mathcal{P}^{''}\right) \qquad (8)$$

The feature dimension is decreased from N to 2N thanks to the linear layer of the ConvB, and the element-wise activation function is $\mathcal{Q}$.

3.5 Convolution Module

The convolution module is a component of the ConvTransFormer block that is designed to extract fine-grained local characteristic patterns in the input sequence. This is accomplished by replacing the dense layers in the gated attention unit (GAU) with a 1D depthwise convolution. Figure 3 presents a diagrammatic representation of the convolution module's underlying architecture. It begins with the normalization and projection of the input sequence by a linear layer, which is then followed by a Derivative of Sigmoid-weighted Linear Unit (dSiLU). The dSiLU activation function is a variant of the Sigmoid function that is designed to better handle vanishing gradients during training. The convolution block is shown in Figure 4.

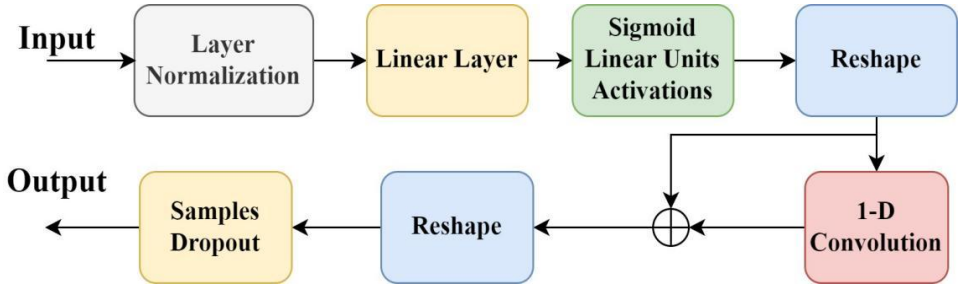

Figure 4. Convolution Module

After normalization and projection, the input sequence is convolved by a 1D depthwise convolution. This type of convolution applies a separate filter to each input channel, allowing the model to extract different features from different parts of the input sequence. The resulting feature maps are then passed through a skip connection, which helps to preserve information from the original input sequence. To help prevent overfitting, the convolution module also includes a dropout layer, which randomly drops out a portion of the feature maps during training.

3.6 Joint Local and Global Self-Attention

The attention mechanism allows the model to learn which parts of the input data are most relevant to the task at hand. The gating functions, on the other hand, allow the model to control the flow of information by selectively filtering out or amplifying the input data based on the attention weights. Local attention mechanisms focus only on a small window of the input data at a time, while global attention mechanisms consider the entire input sequence. By combining both mechanisms, Joint Local and Global Self Attention allows for the model to attend to both local and global information simultaneously. The input sequence Y is first converted to a shared representation through the convolution module: $\mathcal{L} = ConvB(Y)$. A low-cost per-dim scalar and offset are then applied to get the queries and keys $\mathbb{Q}, \mathbb{Q}^{'}, \mathbb{K}, \mathbb{K}^{'}$ for both the local followed by the global attentions respectively, which are then shared. To represent long-range global interactions for both sequences V and U, we use the following cost-effective linearized form:

$$V^{'}_{global} = \mathbb{Q}^{'}\left(\rho \mathbb{K}^{'T} V\right) \quad (9)$$

$$U^{'}_{global} = \mathbb{Q}^{'}\left(\rho \mathbb{K}^{'T} U\right) \quad (10)$$

Whereas, $\rho = {}^{1}/_{S}$ is a scaling aspect. The local attentions are calculated by dividing the V, U, Q, K into H non-overlapping components of size P, zeros are padded when S < H x P. The local attention is applied independently to each block as shown:

$$V^{'}_{local,\,h} = ReLU^2\left(\lambda \mathbb{Q}_h \mathbb{K}_h^T\right) V_h \quad (11)$$

$$U^{'}_{local,\,h} = ReLU^2\left(\lambda \mathbb{Q}_h \mathbb{K}_h^T\right) U_h \quad (12)$$

Where λ= 1/p is a scaling factor. The performance optimisation is performed by using ReLU. The factor $\mathbb{Q}_h \mathbb{K}_h^T$ is computed once for the calculation of the $V^{'}_{local,\,h}$ and $U^{'}_{local,\,h}$ these are combined to derive the local and global variables i.e $V^{'}_{local} = \left[V^{'}_{local,\,1}, \ldots\ldots V^{'}_{local,\,h}\right]$ and $U^{'}_{local} = \left[U^{'}_{local,\,1}, \ldots\ldots\ldots U^{'}_{local,\,h}\right]$. Local and global attention are combined to obtain the final joint attention, given by:

$$V^{'} = V^{'}_{local} + V^{'}_{global} \quad (13)$$

$$U^{'} = U^{'}_{local} + U^{'}_{global} \quad (14)$$

4. Experiment

4.1 Datasets

We have taken into account the WSJ0-2mix and WSJ0-3mix datasets, as well as WHAM!/WHAMR! [11], where random mixing of utterances from the WSJ0 corpus is used to construct mixtures with two and three speakers, respectively. The 8kHz data version is what we use. During the training and validation phases, the utterances are arbitrarily cut into four-second chunks. We also consider dynamic mixing alongside the static copies of the data. The speakers for the training and testing sets are randomly selected. We use the 'min' version of the dataset, the additional waveforms are clipped to the shorter signal's length, and the dataset is converted to an 8kHz version of the original one. Source separation methods are typically tested on these datasets.

4.2 Performance Evaluation Metrics

The core objective of training the end-to-end system lies in the meticulous optimization of the Scale-Invariant Source-to-Noise Ratio (SI-SNR) and Signal Distortion Ratio (SDR), metrics that have gained prevalence as the primary evaluation criterion for source separation, supplanting the conventional source-to-distortion ratio (SDR). The SI-SNR metric, chosen for its robustness and applicability in diverse audio processing scenarios, plays a pivotal role in assessing the effectiveness of the source separation process.

SI-SNR is defined as the ratio of the power of the clean source signal to the power of the residual noise, capturing the quality of the separated source while considering the impact of background noise. This metric's scale invariance ensures that the evaluation remains consistent across different scales, enhancing its utility in various real-world applications. By prioritizing the maximization of SI-SNR during the training process, the end-to-end system endeavours to enhance the system's ability to disentangle source signals from noise, ultimately contributing to improved performance and fidelity in source separation tasks. The SI-SNR metric is characterized by its definition as follows:

$$S_{target} = \frac{\langle \hat{x},\ x \rangle x}{||x||^2} \quad (15)$$

$$e_{noise} = \hat{x} - S_{target} \quad (16)$$

$$SI - SNR = 10log_{10}\frac{||S_{target}||^2}{||e_{noise}||^2} \quad (17)$$

4.3 Architecture and Training Setup of ConvTransFormer

In selecting models, we consider optimal parameters and constraints related to network size, training resources, model dimensions, convolution kernel sizes, chunk sizes, attention dimensions, and the gating activation function. Our training setup includes a single GPU with 16 GB RAM. The ConvTransFormer is equipped with a 256-convolutional-filter encoder, utilizing a PReLU gating activation function, a kernel size of 16 samples, and a stride factor of 8 samples. The decoder's kernel size mirrors that of the encoder. We utilize the Adam optimizer [12] for training our models over a maximum of 200 epochs, starting with a learning rate of 14e-5 and a batch size of 1. The learning rate is halved after 85 epochs, with a patience setting of 2. To maintain training stability, we apply gradient clipping with a threshold of 5 for the l2 norm of the training gradients.

5. Results and Discussion

The WSJ0-2/3 Mix datasets are commonly utilized as standard benchmarks in source separation research. In our study, we assess the performance of the newly developed ConvTransFormer against leading-edge methods using the WSJ0-2mix dataset. The results of this evaluation are detailed in Table 1. A notable outcome of the ConvTransFormer's application on the test set is a significant enhancement in audio quality metrics. Specifically, there is a 22.4 dB increase in Signal-to-Interference (SI)-Signal-to-Noise Ratio improvement (SNRi) and a 22.6 dB rise in Signal-to-Distortion Ratio improvement (SDRi)-Signal-to-Interference (SDRi). State-of-the-art performance is attained by employing dynamic mixing in the suggested design. In comparison to other systems,

ConvTransFormer performs better without employing Dynamic Mixing (DM). However, Wave Split achieves its outstanding performance by using speaker identification as additional information.

Table 1. WSJ0-2mix Dataset Results

Model	SI-SNRi	SDRi	# Parameter	Stride
Tasnet	10.8	11.1	-	20
SignPredictionNet	15.3	15.6	55.2M	8
ConTasnet	15.3	15.6	5.1M	10
Two-Step CTN	16.1	--	8.6M	10
MGST 19	17.0	17.3	-	-
DeepCASA	17.7	18.0	12.8M	1
DualPathRNN	18.8	19.0	2.6M	1
DPTNet*	20.2	20.6	2.6M	1
ConvTransFormer	22.4	22.6	26.8M	8

5.1 Results on WHAM! and WHAMR! Datasets

The WSJ0-2Mix dataset was modified to include environmental noise in WHAM! And reverberation in WHAMR!. The models employed by WHAM! And WHAMR! Are identical to those utilized by the WSJ0-2/3Mix dataset. We train the model to perform source separation and voice augmentation simultaneously across both datasets. A de-noising and separating model is trained on the WHAM! Dataset, while a WHAMR! The model is trained on the WHAM! And WHAMR! Datasets [6], [13]. When constructing new mixtures in WHAMR!, we randomly select a room-impulse response from the WHAMR! Dataset's training set and combine it with the results of the speed augment and/or dynamic mixing.

In Table 2 and 3, we present the findings that we obtained using ConvTransFormer on the WHAM! And WHAMR! Datasets and compare them to the results that were obtained using the other approaches described in the literature. On the WHAM! And WHAMR! Datasets, we find that ConvTransFormer outperforms the previously described approaches despite the fact that it does not make use of DM, which further increases the performance of the algorithm.

Table 2. WHAM! Dataset Results

Model	SI-SNRi	SDRi
MGST 19	13.1	
Wavesplit + DM	16.0	16.7
ConvTasnet	12.7	-
Learnable fbank	12.9	--
ConvTransFormer	16.5	16.8

Table 3. WHAMR Dataset Results

Model	SI-SNRi	SDRi
Wavesplit + DM	13.2	12.2
BiLSTM Tasnet	9.2	--
ConvTasnet	8.3	-
ConvTransFormer	14.2	13.3

6. Conclusion

In this paper, our ConvTransFormer architecture stands out as a robust solution for advancing Single-Channel Blind Source Separation (SCBSS). By incorporating multiple simple-weak attention mechanisms and a triple-gating structure within the gated attention unit (GAU), the model demonstrates enhanced selectivity in capturing diverse elements within input sequences. Rigorous experimentation on the WSJ0-2mix dataset validates the ConvTransFormer's exceptional performance, marked by substantial improvements in critical metrics such as Signal-to-Interference (SI)-Signal-to-Noise (SNRi) and Signal-to-Distortion (SDRi)-Signal-to-Interference (SDRi). Our findings highlight the ConvTransFormer's potential to significantly impact audio signal processing, offering an effective solution to the challenges posed by SCBSS. The nuanced attention mechanisms and triple-gating structure contribute to its efficacy, positioning it as a valuable tool for isolating distinct sources within single-channel recordings. Beyond affirming its real-world impact, these results underscore the ConvTransFormer's versatility across various audio processing scenarios. In pushing the boundaries of source separation methodologies, our study positions the ConvTransFormer as a compelling contribution to the evolving landscape of audio signal processing, promising advancements in both research and practical applications.

7. Acknowledgement

The authors acknowledge the support from REVA University for the facilities provided to carry out the research.

References

[1] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702-1726, Oct. 2018.

[2] Y. Luo, Z. Chen and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 46-50). IEEE, 2020, May.

[3] A. Gulati, J. Qin, C.C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[4] J. Kim, M. El-Khamy and J. Lee, "T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement," In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6649-6653). IEEE, 2020, May.

[5] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[6] M. Maciejewski, G. Wichern, E. McQuinn and J. Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 696-700). IEEE, 2020, May.

[7] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1-6). IEEE, 2020, September.

[8] Y. Liu, and D. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2092-2102, pp. 2092-2102, 2019.

[9] L. Dong, S. Xu and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5884-5888). IEEE, 2018, April.

[10] Q. Zhang, H. Zhu, Q. Song, X. Qian, Z. Ni and H. Li, "Ripple sparse self-attention for monaural speech enhancement," In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE, 2023, June.

[11] Z.Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 71-75). IEEE, 2019, May.

[12] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan and P. Smaragdis, "Two-step sound source separation: Training on learned latent targets," In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 31-35). IEEE, 2020, May.

[13] G. Wichern, J. Antognini, M. Flynn, L.R. Zhu, E. McQuinn, D. Crow, E. Manilow and J.L. Roux, "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.