electronics

MDPI

Article

Improving Top-Down Attention Network in Speech Separation by Employing Hand-Crafted Filterbank and Parameter-Sharing Transformer

Aye Nyein Aung and Jeih-weih Hung *

Department of Electrical Engineering, National Chi Nan University, No. 301, University Road, Puli Township, Nantou 54561, Taiwan; s111356509@mail1.ncnu.edu.tw

* Correspondence: jwhung@ncnu.edu.tw

Abstract: The "cocktail party problem", the challenge of isolating individual speech signals from a noisy mixture, has traditionally been addressed using statistical methods. However, deep neural networks (DNNs), with their ability to learn complex patterns, have emerged as superior solutions. DNNs excel at capturing intricate relationships between mixed audio signals and their respective speech sources, enabling them to effectively separate overlapping speech signals in challenging acoustic environments. Recent advances in speech separation systems have drawn inspiration from the brain's hierarchical sensory information processing, incorporating top-down attention mechanisms. The top-down attention network (TDANet) employs an encoder–decoder architecture with top-down attention to enhance feature modulation and separation performance. By leveraging attention signals from multi-scale input features, TDANet effectively modifies features across different scales using a global attention (GA) module in the encoder–decoder design. Local attention (LA) layers then convert these modulated signals into high-resolution auditory characteristics. In this study, we propose two key modifications to TDANet. First, we substitute the fully trainable convolutional encoder with a deterministic hand-crafted multi-phase gammatone filterbank (MP-GTF), which mimics human hearing. Experimental results demonstrated that this substitution yielded comparable or even slightly superior performance to the original TDANet with a trainable encoder. Second, we replace the single multi-head self-attention (MHSA) layer in the global attention module with a transformer encoder block consisting of multiple MHSA layers. To optimize GPU memory utilization, we introduce a parameter sharing mechanism, dubbed "Reverse Cycle", across layers in the transformer-based encoder. Our experimental findings indicated that these proposed modifications enabled TDANet to achieve competitive separation performance, rivaling state-of-the-art techniques, while maintaining superior computational efficiency.

Keywords: speech separation; TDANet; multi-phase gammatone filterbank; parameter-sharing transformer

check for updates

Citation: Aung A.N.; Hung, J.-w. Improving Top-Down Attention Network in Speech Separation by Employing Hand-Crafted Filterbank and Parameter-Sharing Transformer. *Electronics* **2024**, *13*, 4174. https://doi.org/10.3390/electronics13214174

Academic Editor: Andreas Mauthe

Received: 15 September 2024
Revised: 17 October 2024
Accepted: 23 October 2024
Published: 24 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Infants can recognize their names even at a young age of five months [1]. Behavioral research has demonstrated that the capacity to focus on relevant stimuli while disregarding irrelevant stimuli gradually improves as children grow older [2]. This highlights a key brain function: the ability to focus on a single auditory stimulus in a complex auditory environment. This is a difficult problem, but the brain solves it quickly and precisely. This impressive ability unfolds through three key stages: first, the precise processing of low-level auditory attributes; second, the segregation of sounds into distinct, coherent sources; and third, the selective focus on one source while filtering out others, enabling higher-level cognitive processing [3]. As soon as a multi-source, multi-acoustic signal reaches the human auditory system, it begins to decompose into its component acoustic waves at different stages of the auditory pathway. A wide range of spectro-temporal features reach the

auditory cortex. By selectively integrating relevant features and discarding irrelevant ones, a distinct auditory object is perceived [4]. A notable example is the cocktail party effect [5], where a listener can distinguish one speaker from the crowd, focusing on his speech.

Recent neuroscience studies have suggested that the human brain uses top-down attention to solve the cocktail party problem [6]. This function helps people concentrate by removing unnecessary interruptions. In particular, top-down attention alters the cortical response to many stimuli [7]. This enables the brain to focus on the speech it wants to hear while shutting out other speakers in multi-speaker scenarios [8]. In computational terms, a spectro-temporal receptive field is a filter that is employed to focus on a specific acoustic signal. Known as a mask, this filter facilitates the isolation of the desired signal components. Element-wise multiplication of the input signal and the mask produces the focused attention response. This process of superimposing a filter on top of an input signal is usually known as top-down attention [9].

Traditional statistical methods like computational auditory scene analysis (CASA) [10–14] and non-negative matrix factorization (NMF) [15–17] face limitations in solving the cocktail party problem, due to their reliance on pre-defined acoustic features, making them less adaptable to unseen conditions, especially when encountering new speakers. These methods also struggle with speaker permutation ambiguity, hindering their ability to separate individual speakers in complex real-world scenarios. Deep neural networks (DNNs) [18–20] address these issues by leveraging large-scale data and advanced representation learning. Techniques such as deep clustering (DC) [21] and permutation invariant training (PIT) [22] effectively resolve speaker permutation ambiguity by learning embeddings that distinguish speakers. Additionally, DNNs demonstrate strong generalization to unseen speakers, resulting in superior performance in dynamic, real-world auditory environments [23–26].

Recent progress in DNN-based speech separation systems, such as SuDoRM-RF (successive downsampling and resampling of multi-resolution features) [27,28] and A-FRCNN (asynchronous fully recurrent convolutional neural network) [29], has mirrored the diverse types of connections seen in the stacked structure of the brain's sensory information processing. These architectures feature top-down, bottom-up, and lateral connections, mirroring the hierarchical structure of the human brain used for processing sensory information [30]. While these models primarily simulate interactions between the lower (e.g., A1) and higher (e.g., A2) sensory areas in primates, they often overlook the contributions of higher cortical regions, such as the frontal and occipital cortices, in performing complex auditory tasks. Nevertheless, these frameworks present promising opportunities for implementing top-down attention mechanisms. Partially inspired by the aforementioned framework, the top-down attention network (TDANet) [31] introduces an encoder–decoder architecture with top-down attention to enhance feature modulation. By incorporating a global attention (GA) module into the encoder–decoder design, TDANet modulates features across scales by exploiting attention signals that are obtained from multi-scale input features. These modulated features are then further transformed into high-definition auditory features by means of local attention (LA) layers.

Our work in this paper aimed to improve the separation capabilities of TDANet, while ensuring a favorable balance between performance and resource efficiency. Two proposed modifications for TDANet are succinctly outlined, as follows:

- We suggest utilizing a fixed, manually designed filterbank called the multi-phase gammatone filterbank (MP-GTF) as a substitute for the fully trainable convolutional encoder in TDANet. This substitution aims to mimic the auditory perception of the human ear. The experimental results indicated that this strategy produced a moderately superior performance compared to the original TDANet with a trainable encoder.
- The other modification is to expand the global attention (GA) module of the original TDANet by replacing the single multi-head self-attention (MHSA) layer with a whole transformer encoder block that includes multiple MHSA layers. In order to maximize the efficiency of GPU memory usage, we propose employing a parameter

sharing approach called "Reverse Cycle" that is implemented across layers in the transformer-based encoder. The experimental results showed that this modification could further improve TDANet for obtaining excellent separation performance, allowing it to perform alongside the state-of-the-art (SOTA) approaches, while preserving a higher computational efficiency.

2. TDANet

2.1. Overall Pipeline

Here, we introduce TDANet [31], which will act as the archetypal framework for the methods described later. Like Conv-TasNet [32], TDANet uses a three-part pipeline consisting of an audio encoder, a separation network, and an audio decoder. The pipeline's general layout is depicted in Figure 1. The input y is transformed by the audio encoder into a frame-based, N-dimensional embedding sequence with a length of T, called $E \in \mathbb{R}^{N \times T}$. With a kernel size of L and a stride size of $\lfloor L/4 \rfloor$, a 1-D convolutional layer is used to generate this sequence, which is known as mixed audio embedding. TDANet adopts a masking-based approach [33–35] and employs a DNN-based separation network $f(E;\theta)$ to generate masks $M = \{M_i \in \mathbb{R}^{N \times T} \mid i = 1, \ldots, C\}$ for each target speaker, thereby estimating the target speech embeddings. Each target speech embedding $\bar{E}_i \in \mathbb{R}^{N \times T}$ is derived by applying a corresponding mask M_i to the mixture audio embedding E, expressed as $\bar{E}_i = E \odot M_i$, where $\odot$ denotes the element-wise product. An audio decoder, typically implemented with a 1-D transposed convolutional layer having kernel and stride sizes identical to those of the audio encoder, then reconstructs the target waveform $\bar{x}_i$ from the target speech embedding $\bar{E}_i$.

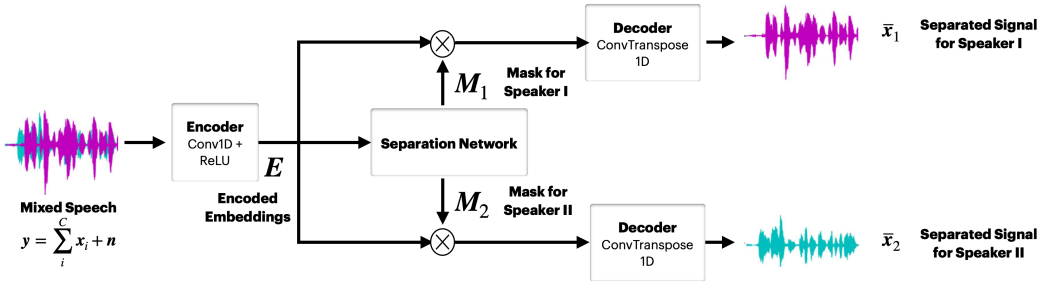


Figure 1. Schematic representation of a speech separation framework. The system processes an input mixture signal to produce two separated source signals for a two-speaker scenario.

2.2. Separation Network

The exceptional performance of TDANet can be attributed to its three main components: (a) encoder, (b) global attention (GA) module, and (c) decoder within the separation network. The separation network of TDANet is depicted in Figure 2. The step-by-step process inside the separation network is as follows: First, the encoder captures multi-scale features through top-down connections. Next, the GA module extracts global features from the top layer of these hierarchical features, using them as attention features to modulate features from different scales via bottom-up connections. The features from adjacent layers are then input into the decoder to extract acoustic features through local attention (LA) layers, thereby completing one processing cycle in TDANet. This approach is similar to that of A-FRCNN [29]; the separation network of TDANet is repeated for several cycles, with the output features from the final cycle serving as the final separation network output.

The key differences between the local attention (LA) and global attention (GA) modules in TDANet are based on their operational focus and input processing. The GA module processes multi-scale acoustic features to extract a global attention signal, which then modulates features across different scales through direct top-down connections. This allows the model to capture broader contextual information, enhancing the overall feature representation. In contrast, the LA layers utilize features from adjacent layers to extract local attention signals, which modulate the input features in a more localized manner. This

focused attention helps improve the detail of the features passed to the decoder. Together, these attention mechanisms significantly contribute to TDANet's performance, by enabling effective modulation of features, thereby improving the quality of separated audio outputs, while maintaining computational efficiency. The GA module enhances the model's ability to utilize global context, while the LA layers refine the focus on local details, leading to superior speech separation outcomes.

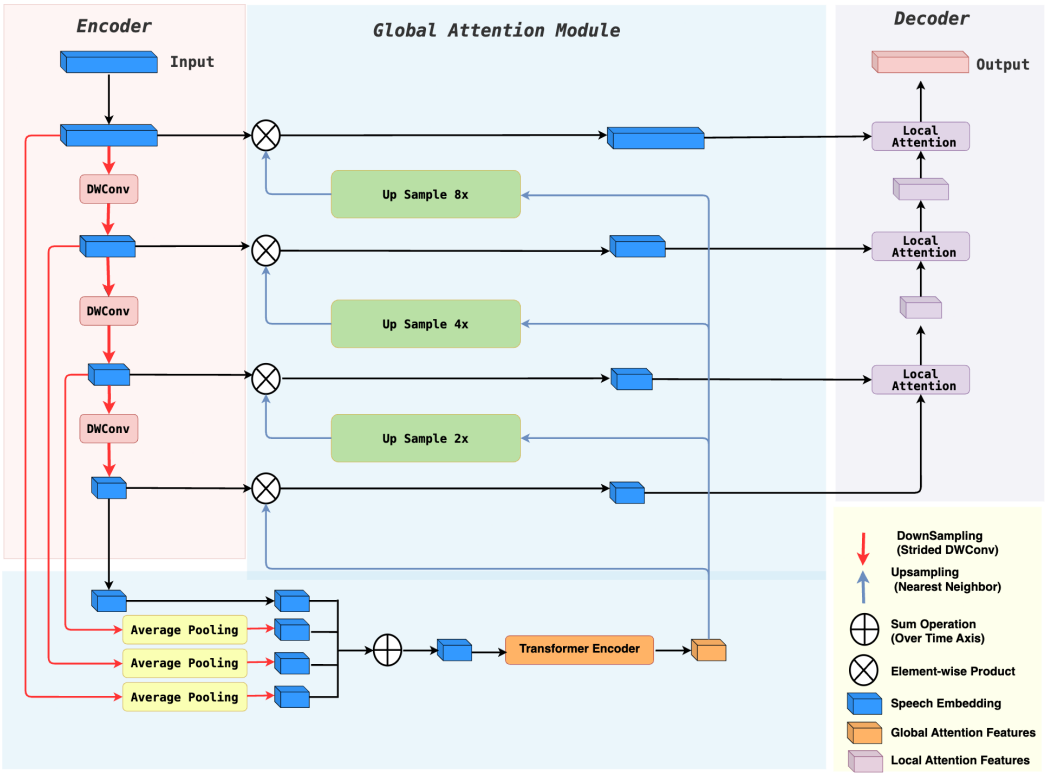


Figure 2. Overall architecture of separation network within TDANet (redrawn from to [31]).

3. Proposed Method A: Muliti-Phase Gammatone Filterbank as Encoder in TDANet

3.1. Introduction

Early approaches to monaural speech separation using deep neural networks, such as deep clustering [21] and Chimera++ [36], tackled the separation task by transforming mixed speech into a short-time Fourier transform (STFT) representation. These methods train a network to generate an optimal multiplicative mask for each speaker within the domain. Although these STFT magnitude-based techniques achieve reasonable performance, their effectiveness is constrained by the limitation of ideal mask calculation, which generally omits phase information.

End-to-end speech separation systems such as TasNet [37] and Conv-TasNet [32] overcome these limitations by introducing several key changes. First, they define the training loss in the time domain instead of the STFT domain. This allows the replacement of the STFT analysis and synthesis stages with learned convolutional analysis and synthesis filterbanks, referred to as the encoder and decoder, respectively. Second, the encoder and decoder use significantly smaller frame sizes compared to the STFT analysis–synthesis windows used in traditional approaches. This reduction in frame size significantly decreases the algorithmic latency when the separation network is developed in a causal manner.

As first proposed in TasNet, TDANet also uses a fully learnable convolutional encoder–decoder structure. Therefore, TDANet employs data-driven encoding features created using a convolutional network. However, this raises a significant issue: should we instead use a well-known, deterministic encoder (analysis filterbank) established in signal processing and perceptual theories, in order to achieve better results for TDANet?

Based on this motivation, we proposed replacing the learned encoder of the TDANet with a pre-defined gammatone filterbank for the purpose of investigation.

3.2. Multi-Phase Gammatone Filterbank (MP-GTF)

An auditory gammatone filterbank (A-GTF) [38] is a type of deterministic filterbank that closely imitates the way human auditory signals are encoded. The multi-phase gammatone filterbank (MP-GTF) [39] that we selected for investigation is an improved version of the ordinary A-GTF. The complete construction of the MP-GTF may be found in [39]. A MP-GTF models the movement patterns of the basilar membrane in the human auditory system. This filterbank consists of narrow-band filters that are spaced non-linearly and have increasing bandwidths corresponding to the filter's center frequency. This design accurately represents the masking effect of the human auditory system, and it is very likely to generate suitable features for speech separation.

Referring to [38,39], we introduce the process of creating a multi-phase gammatone filter-bank in the following. First, the impulse response of all filters in the filter-bank is a gamma probability distribution function multiplied with a sinusoid:

$$\gamma(t) = \alpha t^{n-1} \exp(-2\pi bt) \cos(2\pi f_c t + \phi), \tag{1}$$

where n, b, f_c, α, and ϕ are the parameters for the order, bandwidth, center frequency, amplitude, and phase shift, respectively. The details of the impulse response and parameter settings are described as follows:

- The length of the impulse response for each filter is set to 2 ms, equating to 16 samples at an 8 kHz sampling rate. This maintains a low system latency by means of a short filter length.
- The order n is set to 2, which enables the impulse response to move the amplitude peaks inside the time span and towards the time origin.
- The number of center frequencies is set to 24, and the center frequencies are evenly spaced on the scale of equivalent rectangular bandwidth (ERB). Note that the relationship between the acoustic frequency f_{Hz} (in Hz) and the ERB scale is described by a nonlinear logarithmic equation:

$$\text{ERB}_{\text{scale}}(f_{\text{Hz}}) = 9.265 \log\left(1 + \frac{f_{\text{Hz}}}{24.7 \times 9.265}\right). \tag{2}$$

  The center frequencies for any two adjacent filters are spaced by 1 unit on the ERB scale, and thus we have the following equation to determine the center frequencies iteratively:

$$f_c^{(m+1)} = \text{ERB}_{\text{scale}}^{-1}(\text{ERB}_{\text{scale}}\,(f_c^{(m)}) + 1), \tag{3}$$

  where $\text{ERB}_{\text{scale}}^{-1}(\cdot)$ is the inverse of $\text{ERB}_{\text{scale}}(\cdot)$, and the lowest center frequency $f_c^{(0)}$ is set to 100 Hz. This also allows the highest center frequency $f_c^{(23)}$ to be less than 4000 Hz, which is the Nyquist frequency for the 8000 Hz sample rate.
- For a filter with a center frequency f_c, the bandwidth parameter b is set to $\text{ERB}_{\text{scale}}(f_c)$.
- The amplitude parameter α is determined by normalizing the peak value of all filters in the frequency domain.
- A negative version $-\gamma(t)$ is included in the filterbank for each filter $\gamma(t)$ by applying a phase shift π as $\phi_{\text{inv}} = \phi + \pi$. In addition, multiple filters with varying phase shifts for the same center frequency are contained in the filterbank. The number of filters N is 48 minimum, with 24 center frequencies and their phase-inverted variants. Choosing several phase values for every center frequency generates extra filters for $N > 48$. These phase shifts are equally spaced over the interval $[0, \pi)$ and their inverted counterparts $[\pi, 2\pi)$.

In Figure 3, the time and frequency domain representations of the MP-GTF with $N = 128$ filters are plotted.

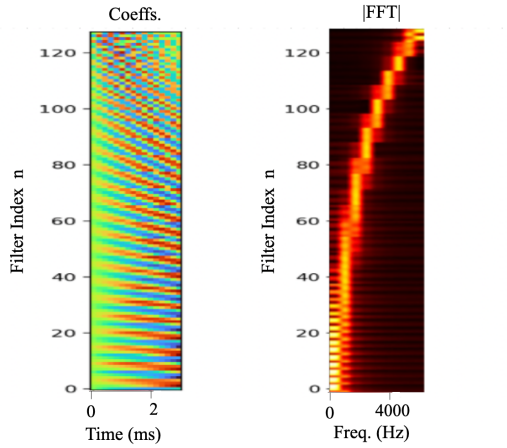


Figure 3. Temporal and spectral domain representations of all filters in the MP-GTF, with the number of filters N set to 128.

The modified TDANet that incorporates MP-GTF is shown in Figure 4. MP-GTF replaces the fully learnable Conv1D audio encoder of TDANet in this architecture. The purpose of this modification is to improve the separation performance by utilizing the auditory-inspired features of the MP-GTF. Nevertheless, the audio decoder remains a completely learnable element. We chose not to replace it with the pseudo-inverse of the MP-GTF coefficient matrix. This decision was based on the comparison findings reported in [40], which demonstrated that the fully learnable decoder outperformed the pseudo-inverse method when deterministic hand-crafted filterbanks were used.

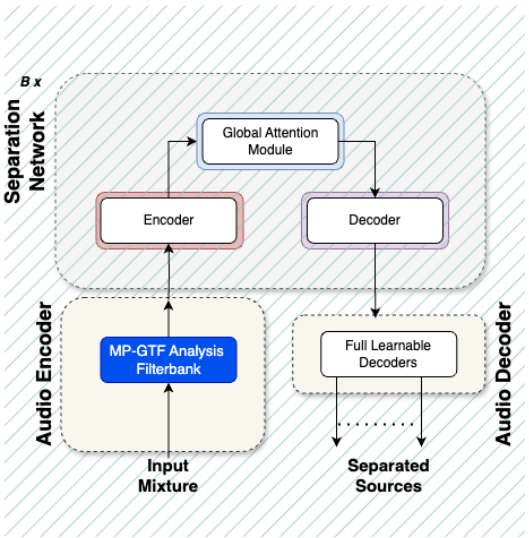


Figure 4. TDANet with MP-GTF as audio encoder.

The multi-phase gammatone filterbank (MP-GTF) improves upon the fully trainable convolutional encoder in TDANet by mimicking the signal encoding process in human auditory perception. As an auditory-inspired filterbank, MP-GTF has already demonstrated its effectiveness in Conv-TasNet [39]. In terms of computational efficiency, MP-GTF eliminates the need for the additional parameters required by the learnable convolutional encoder in the original TDANet, making it more resource-efficient, while maintaining strong performance in speech separation tasks.

Further discussions about the presented TDANet with MP-GTF are given, as follows:

- On the one hand, MP-GTF reduces the number of parameters compared to the convolutional encoder in the original TDANet and is expected to exhibit improved separation performance. On the other hand, the implementation of this revised method with MP-GTF introduces longer training epochs, due to the computational complexity during the filterbank construction phase, which currently relies on CPU processing, since CUDA support for GPU computation is not yet available [39]. Thus, while the use of MP-GTF reduces the model complexity, it may increase the training time and computational overhead while training the model.
- It is important to note that using a hand-crafted filterbank like the multi-phase gamma tone filterbank (MP-GTF) in speech separation tasks might not be essential, as recent state-of-the-art methods have used fully trainable convolutional filterbanks. However, choosing between a well-understood deterministic encoder based on signal processing principles and auditory perception, and a data-driven learned encoder raises an important question. This debate has gained attention in recent research, examining whether the benefits of deterministic designs, such as interpretability and reliance on perceptual features, outweigh the flexibility and potential performance advantages of fully trainable architectures. Ultimately, this decision depends on the specific requirements of the task and the trade-offs between these approaches.

4. Proposed Method B: Parameter Sharing Top-Down Selective Auditory Attention System

4.1. Introduction

Recurrent neural networks (RNNs) driven by multiplicative gate mechanisms, such as LSTM and GRU, were cornerstones of deep learning-based audio processing systems prior to the introduction of the transformer [41] architecture. They were widely used in many areas, such as speech recognition, synthesis, enhancement, and separation. The combination of a dual-path architecture and bidirectional long short-term memory (BLSTM) [42] has been demonstrated to be highly effective for speech separation in dual-path RNN (DPRNN) [43]. This architecture involves the splitting of a sequential input into overlapping chunks, and then alternate and iterative intra-chunk (local) and inter-chunk (global) processing. However, RNNs' sequential nature makes parallel processing less useful, especially when working with large datasets and long sequences. Comparatively, transformer-based models, such as DPTNet [35] and SepFormer [34], use a multi-head attention mechanism within a dual-path design to get around the problems with RNNs. These models allow the input series to be processed in parallel for both global and local dependencies. This makes speech separation much faster and better. Even though these models are lighter than models like [27,32], they still need a lot of learnable parameters and cannot be used for on-device real-time applications.

The chosen baseline system, TDANet, strives to reduce the differences between high-performance large transformer models and lightweight, efficient models with decent performance. To do this, it uses both top-down global attention and local attention on acoustic features at multiple scales. The crucial global attention module of TDANet, on the other hand, only has one transformer layer. This layer is made up of two parts: the multi-head self-attention (MHSA) and feedforward network (FFN) levels. Following [34], we believe that better speaker acoustic patterns can be derived from global contexts by letting the multi-scale features from the TDANet's encoder pass through a full transformer encoder block with multiple transformer encoder layers.

As such, here, we propose using a parameter-sharing scheme across the layers in the extended transformer block, so that the extra parameters and complexity that come with more transformer layers can be processed quickly. The evaluation results showed that this revision makes TDANet's separation performance even better, while keeping the model complexity at a level similar to that of lightweight separation systems.

4.2. Background: Parameter Sharing for Transformers

In transformers, parameter sharing is a method intended to improve model efficiency by means of parameter reuse among several levels or components of the model [44]. In traditional transformer designs, each layer has its own set of parameters. As the model depth increases, the total number of parameters increases by a large amount. Using parameter sharing methods lets some parameters be used in more than one layer or part of the model, which lowers the total number of parameters and the computational complexity.

Parameter sharing has several benefits. Mostly it results in a smaller model size, which makes it possible to use transformer models on devices with limited processing capacity such as edge computing environments and mobile devices. For real-time applications, a shorter inference time and lower latency, which result from this reduction, are also vital. Furthermore, by letting deeper models be trained with the same computational resources needed for shallower models, without parameter sharing, parameter sharing can result in enhanced training efficiency [45].

The first transformer parameter-sharing method was presented in the universal transformer [44], where the same set of parameters was utilized for all levels and only one layer of a transformer-based encoder–decoder. Although, in several cases of the application of natural language processing (NLP), the universal transformer outperformed conventional transformers, its great dimension size resulted in noticeably longer computing times during training than the conventional transformer. In an effort to enhance this, three methods for relaxed parameter assignments across many layers were proposed in [46]: SEQUENCE, CYCLE, and CYCLE (REV). Under the same resource limitations, these methods were shown to beat the universal transformer in terms of parameter size and computing time.

4.3. Proposed Method: TDANet with Parameter Sharing

Based on the findings in [46], we propose redesigning TDANet by replacing its current one-layer transformer inside the global attention module with a six-layer parameter-sharing transformer block, which is depicted in Figure 5. Of the three methods, the CYCLE (REV) strategy has demonstrated the highest improvements in related applications, so we decided to use it as our parameter sharing method. By reusing parameters from lower levels in higher layers, the CYCLE (REV) method constructs a N-layer transformer-based encoder with M parameter-sharing layers. Achieving this involves, first, stacking M layers one after the other until there are $M \times (\lceil N/M \rceil - 1)$ layers. The remaining layers are constructed by stacking the M layers in the reverse order. For example, when $M = 3$ and $N = 6$, we first stack three encoder layers sequentially, and then stack the last three layers in reverse order. Algorithm 1 shows how each encoder layer is constructed utilizing the CYCLE (REV) parameter sharing strategy.

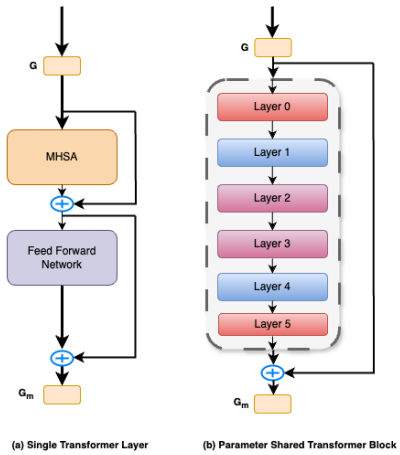


Figure 5. Comparison between (**a**) a single transformer block and (**b**) a six-layer parameter-sharing transformer block (transformer encoder layers with same color denote parameter sharing layers).

Algorithm 1: Encoder layer construction using CYCLE (REV) strategy

Input: The total number of layers N, Number of independent layers M
Output: $\text{enc}_1, \ldots, \text{enc}_N$

```
for i ∈ [1, ..., N] do
    if i == 1 then
        enc_i ← CreateNewLayer
    end
    else if i ≤ M then
        enc_i ← CreateNewLayer
    end
    else if i ≤ (M × (⌈N/M⌉ − 1)) then
        enc_i ← enc_((i−1) mod M)+1
    end
    else
        enc_i ← enc_(M−((i−1) mod M))
    end
    i ← i + 1
end
```

The "Reverse Cycle" parameter-sharing mechanism in the transformer encoder block significantly enhances the GPU memory utilization by allowing the reuse of parameters from lower layers in higher layers. This innovative approach effectively reduces the overall number of parameters needed, in stark contrast to traditional methods that allocate unique parameters to each layer, leading to higher memory consumption. By mitigating the redundancy associated with lower-layer parameters, the reverse cycle strategy strikes a crucial balance between efficiency and model expressiveness. This enables the construction of deeper models, without a corresponding increase in computational costs.

While the proposed parameter-sharing scheme offers substantial benefits, replacing the single multi-head self-attention (MHSA) layer in TDANet with a complete transformer encoder block could enhance the TDANet baseline. However, this modification also presents challenges, particularly regarding the increased computational complexity due to the additional attention and feedforward layers. This trade-off must be carefully considered to ensure that the benefits of improved performance do not come at the expense of efficiency.

5. Experimental Setup

5.1. Data

We used the Libri2Mix dataset and task [47] to evaluate the methods presented. This dataset is set up so that the test set and the training/validation set do not share any speaker identities. This design ensures that the test set includes data from speakers who are not present during the training phase, allowing for a fair and robust assessment of a model's generalization abilities. The target speech in the mix audio in this dataset comes from a random subset of LibriSpeech's train-100 set [48]. The chosen audio samples are mixed according to uniformly sampled loudness units relative to full scale (LUFS) [49], ranging from -25 to -33 dB. Compared to classical signal-to-noise ratios (SNRs), LUFS are better correlated with human perception, silence-invariant, and less sensitive to downsampling. Each mix audio clip has two separate speakers, lasts three seconds, and is recorded at an 8 kHz sampling rate.

5.2. Network Configurations

In order to ensure a fair comparison, the majority of the training hyperparameters for the revised TDANet network were set to match those of the baseline TDANet [31]. The first revision involved MF-GTF features, using 128 multiphase gamma-tone filters with a kernel size of 32 and a stride size of 8 for the audio encoder for the 8 kHz sampled data. The

downsampling factor was set to 4, and each convolutional layer contained 512 channels. For the multi-head self-attention (MHSA) layer, we used 8 attention heads with a dimension of 512, along with sine and cosine functions at different frequencies for positional encoding.

Next, in the second revision, regarding parameter-sharing transformers, we utilized a fully expanded transformer encoder block with six MHSA layers, three of which shared parameters using the CYCLE (REV) strategy. In order for the model to compute attention over the entire series, using both past and future context, the mask was turned off in each MHSA layer. The FFN layer consisted of three convolutional layers, each with 512, 1024, and 512 channels; kernel sizes of 1, 5, and 1; strides of 1; and biases set to false, true, and false, respectively. To prevent overfitting, dropout was employed with a probability of 0.1. The default separation network was set to repeat 16 times, except in the revised network with parameter-sharing, which reduced the repetition to 8 cycles before reaching the audio decoder.

The models were scheduled to be trained for 500 epochs, with a batch size of 4 at the utterance level. The Adam optimizer was employed with an initial learning rate of 0.001. The training objective was to maximize the scale-invariant signal-to-distortion ratio (SI-SDR) [50], and permutation-invariant training (PIT) [22] was utilized to address source permutation issues (which will be introduced in the next subsection). If the best model was not found in 15 consecutive epochs, the learning rate was reduced by half. If an improvement in relative loss failed to appear within 30 consecutive epochs, the training procedure was halted early according to the early stopping criterion. Gradient clipping was used with a maximum L2 norm of 5 to avoid gradient explosion. All experiments were carried out with a single GeForce RTX 4060 GPU for both training and testing.

5.3. Evaluation Metric

We used scale-invariant signal-to-distortion ratio improvement (SI-SDRi) as an objective evaluation metric of the separation precision. For the clean source signal s and the predicted speech signal $\hat{s}$, SI-SDR was defined as follows:

$$\rho(\hat{s}, s) = 10 \log_{10} \left(\frac{\left\| \frac{\langle \hat{s}, s \rangle}{\langle s, s \rangle} \right\|^2}{\left\| \frac{\langle \hat{s}, s \rangle}{\langle s, s \rangle} s - \hat{s} \right\|^2} \right), \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product operation. To ensure scale invariance, both $\hat{s}$ and s signals were normalized to zero-mean before calculating SI-SDR. In speech separation, the issue of permutation arises when multiple clean source signals are misaligned. To address this, permutation invariant training (PIT) [22] was employed to identify the optimal SI-SDR value.

PIT tackles the permutation problem by dynamically computing a training loss function. It first evaluates the loss between the reference features $\mathbf{S} = \{S_c\}_{c=1}^{C}$ and the estimated features $\hat{\mathbf{S}}$ across all possible permutations $\mathbf{P}$. The permutation that yields the minimum loss is selected, and this loss is then used for back-propagation to update the parameters of the separation network. The PIT loss $\mathcal{L}_{\text{PIT}}$ can be expressed as follows:

$$\mathcal{L}_{\text{PIT}}(\mathbf{S}, \hat{\mathbf{S}}) = \min_{\mathbf{P}} \sum_{c=1}^{C} \mathcal{L}\left(S_c, [\mathbf{P}\hat{\mathbf{S}}]_c\right), \tag{5}$$

where $\mathbf{P}$ denotes a permutation matrix and $\mathcal{L}$ represents the signal-level loss function to be minimized. The SI-SDRi is calculated as the difference between the SI-SDR of the separated signal and that of the original mixture. A higher SI-SDRi value reflects improved performance in the separation task.

5.4. Experimental Results and Discussion

The evaluation was performed on the original TDANet, as well as the two newly revised versions of TDANet with MF-GTF and parameter sharing transformers, respectively. In addition, several advanced speech separation methods, such as BLSTM-TasNet [37], Conv-TasNet [32], SuDoRM-RF1.0x and 2.5x [27] (which consist of 16 and 40 U-Convolution blocks, respectively), DPRNN [43], DPTNet [35] , A-FRRCNN-16 [29], and Sepformer [34], were also included in the evaluation.

Table 1 presents the SI-SDRi values for the test set across the various approaches. Additionally, it provides details on the model parameter numbers and the values of multiply–accumulate operations per second (MACs), which are indicators of model complexity. From this table, we can make the following observations:

1. The original TDANet model, with its compact architecture of just 2.3 million parameters, demonstrated remarkable efficiency in achieving a high performance in speech separation tasks. Despite its relatively small size, it attained an SI-SDRi that was either superior to, or on par with, several advanced resource-efficient separation techniques, including Conv-TasNet, SuDoRM-RF, and DPRNN. This is particularly noteworthy given that contemporary models often require significantly larger parameter counts to achieve comparable performance. The success of TDANet can be attributed to its innovative use of top-down attention mechanisms, which allow it to effectively focus on relevant features in the audio signal, while minimizing the computational overhead. This efficiency makes TDANet an attractive option for real-time applications where both performance and resource constraints are critical.
2. The first revision, TDANet with MF-GTF, showed an absolute improvement of 0.37 in SI-SDRi compared to the original model. This enhancement was achieved with an addition of only 0.3 million parameters, underscoring the model's ability to effectively leverage the handcrafted deterministic auditory filterbank. The MF-GTF approach allows TDANet to capture more nuanced auditory features by mimicking the human auditory system's response to sound, which is particularly beneficial in complex acoustic environments. The use of a deterministic filterbank, as opposed to a fully trainable convolutional filterbank, not only enhances the model's performance but also maintains a low computational cost. This balance between improved performance and a minimal parameter increase highlights the efficacy of integrating bio-inspired techniques into deep learning architectures for speech separation.
   We also found that the presented TDANet with MF-GTF outperformed state-of-the-art techniques like SuDoRM-RF and A-FRCNN in both separation performance and computational complexity. As shown in the table, our method achieved superior SI-SDRi, while requiring significantly fewer MACs and demonstrating a reduced number of parameters.
3. The second revision, TDANet with parameter-sharing transformers, demonstrated a remarkable enhancement in performance metrics, particularly in SI-SDRi measurements. By utilizing an additional 10.6 million parameters, this revised model not only surpassed the original TDANet but also outperformed several high-performing models, including DPTNet, A-FRCNN, and Sepformer. This achievement was significant as it highlighted the effectiveness of parameter sharing in optimizing model performance, while maintaining a manageable increase in complexity.
   The parameter-sharing approach allowed the model to leverage shared weights across different layers, which enhanced its ability to learn from the data, without exponentially increasing the number of parameters. Specifically, the revised TDANet utilized only 25.56% and 20.9% of the MACs required by DPTNet and A-FRCNN, respectively. This substantial reduction in computational complexity not only makes the model more accessible for deployment in real-time applications but also suggests that it can be effectively utilized in environments with limited computational resources.
   Furthermore, when compared to Sepformer, the revised framework showed a 3.52% improvement in SI-SDRi score, while achieving a remarkable 69.83% decrease in

MACs and a 50.38% reduction in parameters. This combination of enhanced performance and reduced computational demands makes TDANet with parameter-sharing transformers a promising solution in speech separation, demonstrating that innovative architectural strategies can yield significant benefits in both efficiency and effectiveness.

4. The integration of both revised methods with the original TDANet resulted in a significant increase in SI-SDRi score, achieving the highest score among all compared systems. This outcome underscores the synergistic effects of combining advanced techniques, such as parameter sharing and the MF-GTF approach, which together enhanced the model's ability to separate overlapping speech signals effectively. Notably, this combined framework operated with only 13.1 million parameters and 33.0 MACs, showcasing an impressive balance between performance and computational efficiency. The ability to achieve superior separation quality, while maintaining a low parameter count and minimal computational overhead is particularly advantageous for practical applications, where resource constraints are often a critical consideration.

Table 1. Performanceand complexity comparison between proposed methods and other existing models on Libri2Mix dataset (test set).

Model	SI-SDRi (↑)	Params (M)	MACs (G/s)
BLSTM-TasNet	7.9	23.6	43.0
Conv-TasNet	12.2	5.6	10.2
SudoRM-RF1.0x	13.5	2.7	4.6
SudoRM-RF2.5x	14.0	6.4	10.1
DPRNN	16.1	2.7	85.3
DPTNet	16.7	2.6	102.5
A-FRCNN-16	16.7	6.1	125.3
Sepformer	16.5	26.0	86.9
TDANet baseline	15.97	2.3	4.7
Proposed Method A (TDANet with MP-GTF)	16.34	2.6	9.1
Proposed Method B (Parameter-sharing TDANet)	17.08	12.9	26.2
Proposed Methods A and B (Combined)	17.38	13.1	33.0

Figure 6 provides a comprehensive overview of spectrograms related to mixed speech, the original sources, and the estimated sources produced by the baseline TDANet model, as well as our proposed methods A and B. Specifically, the waveform and spectrogram of an audio mixture randomly selected from the test set of Libri2Mix are depicted in Figure 6a.

Upon analyzing the spectrograms, it is evident that our proposed methods illustrated in Figure 6d,e demonstrated a significantly improved ability to separate the noise, which represented the signals of another speaker within the speech separation scenario. This enhanced separation capability is clearly demonstrated by the closer resemblance of the spectrograms produced by our proposed methods to the ground truth spectrograms shown in Figure 6b.

Furthermore, the spectrogram generated by the baseline TDANet model, as shown in Figure 6c, reveals limitation in its ability to detect intensities within the higher frequency bands, which are highlighted by the red outline. This limitation is in contrast to the performance of our proposed methods, highlighting their effectiveness in achieving better separation and clarity in the audio signals.

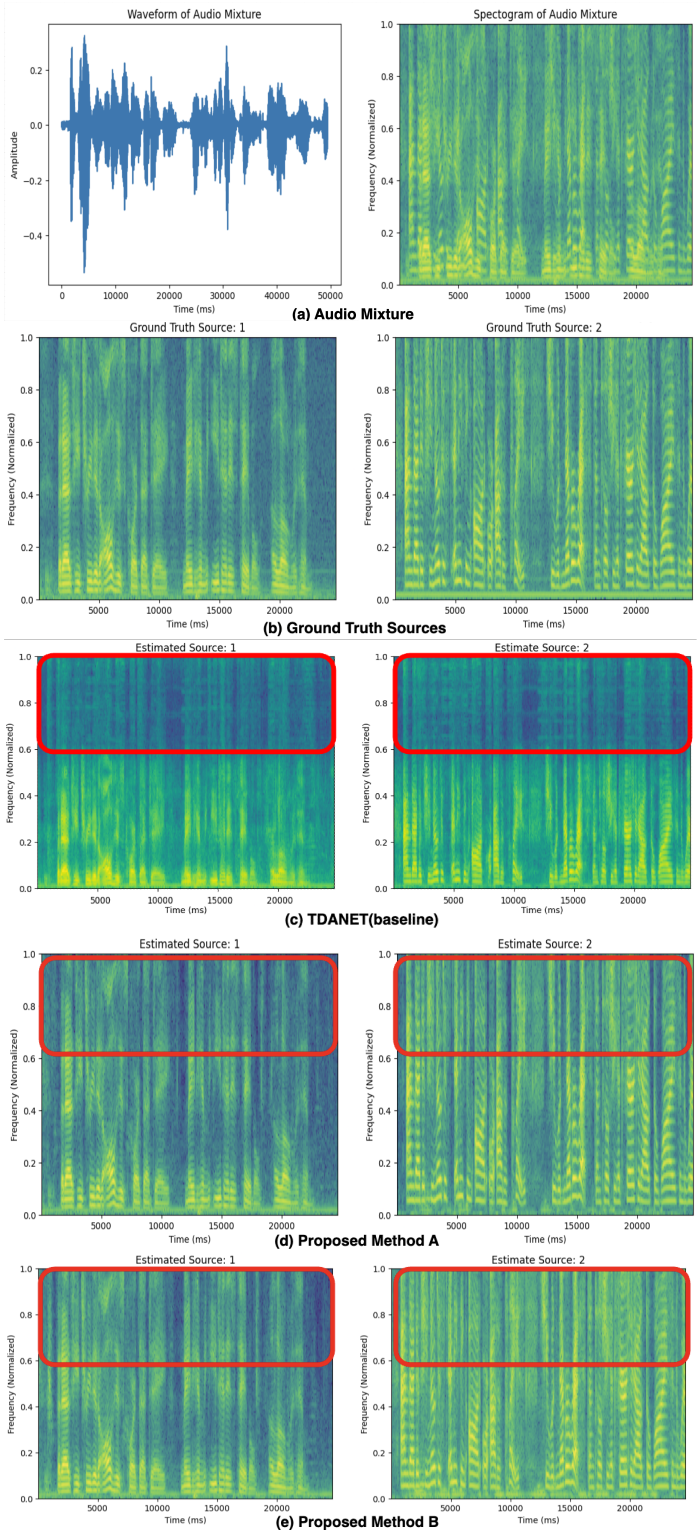


Figure 6. Comparison of spectrograms for ground truth sources and estimated sources produced by the baseline TDANet model and the proposed methods A and B. (**a**) Waveform and spectrogram of the audio mixture. (**b**) Ground truth spectrograms of speaker source 1 and speaker source 2. (**c**) Estimated spectrograms of speaker source 1 and speaker source 2 using the baseline TDANet model. (**d**) Estimated spectrograms of speaker source 1 and speaker source 2 using proposed method A. (**e**) Estimated spectrograms of speaker source 1 and speaker source 2 using proposed method B. The areas outlined by the red box are higher frequency bands.

6. Conclusions

In this study, we have made several key contributions to advancing deep-network-based speech separation methods. First, we proposed the implementation of a multi-phase gammatone filterbank (MP-GTF) as a deterministic, hand-crafted alternative to the trainable convolutional encoder in TDANet. This novel integration enhances feature extraction in overlapping speech scenarios, resulting in improved separation performance. Experiments conducted with the Libri2Mix dataset demonstrated that the MP-GTF feature achieved moderately better performance than the original TDANet, which was equipped with a learnable encoder. Second, we extended the global attention (GA) module of the original TDANet from a single multi-head self-attention (MHSA) layer to a comprehensive transformer encoder block comprising multiple MHSA layers. To optimize GPU memory utilization, we employed a reverse cycle parameter-sharing strategy across layers in the transformer-based encoder. Evaluation on the theLibri2Mix dataset demonstrated that this revised TDANet, with its extended GAmodule, achieved superior separation performance compared to the baseline TDANet and delivered competitive results relative to state-of-the-art (SOTA) methods, while maintaining a higher efficiency.

However, despite these advancements, certain limitations remain. The fixed parameters in the MP-GTF may hinder adaptability to varying acoustic conditions, and the current parameter-sharing strategy is limited to the transformer encoder layers, leaving other parts of the model underutilized. Therefore, in our future research, we will explore several avenues to further enhance the revised TDANet, without compromising its effectiveness. Specifically, for MP-GTF, we intend to experiment with training fixed parameters within the equivalent rectangular bandwidth (ERB), as proposed in [40]. This approach is expected to improve the accuracy and optimize performance in auditory feature extraction, thereby increasing the overall efficiency and effectiveness of MP-GTF. For parameter sharing in the transformer architecture, our future work will extend to exploring parameter sharing across branches and matrices. This technique has already shown promise in achieving a better convergence with reduced model complexity, as demonstrated in [51]. By integrating these advanced parameter-sharing strategies, we aim to further streamline the transformer-based encoder in TDANet and enhance its performance metrics.

Author Contributions: Conceptualization, A.N.A. and J.-w.H.; Methodology, A.N.A.; Software, A.N.A.; Validation, A.N.A. and J.-w.H.; Formal analysis, A.N.A. and J.-w.H.; Investigation, A.N.A. and J.-w.H.; Resources, J.-w.H.; Writing—original draft, A.N.A. and J.-w.H.; Writing—review and editing, J.-w.H.; Visualization, J.-w.H.; Supervision, J.-w.H.; Project administration, J.-w.H.; Funding acquisition, J.-w.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study require the Libri2Mix dataset, which is openly available to download at https://github.com/JorisCos/LibriMix, accessed on 22 October 2024.

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

References

1. Conway, A.R.; Cowan, N.; Bunting, M.F. The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychon. Bull. Rev.* **2001**, *8*, 331–335. [CrossRef] [PubMed]
2. Coch, D.; Sanders, L.D.; Neville, H.J. An Event-related Potential Study of Selective Auditory Attention in Children and Adults. *J. Cogn. Neurosci.* **2005**, *17*, 605–622. [CrossRef]
3. Hill, K.; Miller, L. Auditory Attentional Control and Selection during Cocktail Party Listening. *Cereb. Cortex* **2009**, *20*, 583–590. [CrossRef]
4. Shamma, S.A.; Elhilali, M.; Micheyl, C. Temporal coherence and attention in auditory scene analysis. *Trends Neurosci.* **2011**, *34*, 114–123. [CrossRef]
5. McDermott, J.H. The cocktail party problem. *Curr. Biol.* **2009**, *19*, R1024–R1027. [CrossRef] [PubMed]
6. Morís Fernández, L.; Visser, M.; Ventura-Campos, N.; Ávila, C.; Soto-Faraco, S. Top-down attention regulates the neural expression of audiovisual integration. *NeuroImage* **2015**, *119*, 272–285. [CrossRef] [PubMed]

7. Gazzaley, A.; Cooney, J.W.; McEvoy, K.; Knight, R.T.; D'Esposito, M. Top-down enhancement and suppression of the magnitude and speed of neural activity. *J. Cogn. Neurosci.* **2005**, *17*, 507–517. [CrossRef]
8. Chen, J.; Teo, T.H.; Kok, C.L.; Koh, Y.Y. A Novel Single-Word Speech Recognition on Embedded Systems Using a Convolution Neuron Network with Improved Out-of-Distribution Detection. *Electronics* **2024**, *13*, 530. [CrossRef]
9. Xu, C. Single Channel Multi-Talker Speech Separation with Deep Learning. Ph.D. Thesis, Nanyang Technological University, Singapore, 2020.
10. de Cheveigné, A. *Computational Auditory Scene Analysis*; Wiley: Hoboken, NJ, USA; ISTE: Washington, DC, USA, 2008.
11. Li, P.; Guan, Y.; Wang, S.; Xu, B.; Liu, W. Monaural speech separation based on MAXVQ and CASA for robust speech recognition. *Comput. Speech Lang.* **2010**, *24*, 30–44. [CrossRef]
12. Han, R.; Zhao, P.; Gao, Q.; Zhang, Z.; Wu, H.; Wu, X. CASA based speech separation for robust speech recognition. In Proceedings of the Interspeech, Pittsburgh, PA, USA, 17–21 September 2006.
13. Zeremdini, J.; Ben Messaoud, M.A.; Bouzid, A. A comparison of several computational auditory scene analysis (CASA) techniques for monaural speech segregation. *Brain Inform.* **2015**, *2*, 123–134. [CrossRef]
14. Wang, D.; Brown, G.J. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*; Wiley-IEEE Press: Hoboken, NJ, USA, 2006.
15. Pham, T.; Lee, Y.S.; Chen, Y.A. A review on speech separation using NMF and its extensions. In Proceedings of the 2015 International Conference on Audio, Language and Image Processing, Québec City, QC, Canada, 27–30 September 2015.
16. Nie, S.; Liang, S.; Liu, W.; Zhang, X.; Tao, J. Deep Learning Based Speech Separation via NMF-style Reconstructions. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 2043–2055. [CrossRef]
17. Abdali, S.; NaserSharif, B. Non-negative matrix factorization for speech/music separation using source dependent decomposition rank, temporal continuity term and filtering. *Biomed. Signal Process. Control.* **2017**, *36*, 168–175. [CrossRef]
18. Chollet, F. *Deep Learning with Python*, 1st ed.; Manning Publications: Shelter Island, NY, USA, 2017.
19. Chollet, F. *Deep Learning with Python*, 2nd ed.; Manning Publications: Shelter Island, NY, USA, 2021.
20. Aggarwal, C.C. *Neural Networks and Deep Learning: A Textbook*, 2nd ed.; Springer International Publishing: Cham, Switzerland, 2023.
21. Hershey, J.R.; Chen, Z.; Roux, J.L.; Watanabe, S. Deep clustering: Discriminative embeddings for segmentation and separation. *arXiv* **2015**, arXiv:1508.04306.
22. Yu, D.; Kolbæk, M.; Tan, Z.H.; Jensen, J. Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation. *arXiv* **2017**, arXiv:1607.00325.
23. Huang, P.S.; Kim, M.; Hasegawa-Johnson, M.; Smaragdis, P. Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 2136–2147. [CrossRef]
24. Huang, P.S.; Kim, M.; Hasegawa-Johnson, M.; Smaragdis, P. Deep learning for monaural speech separation. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 1562–1566. [CrossRef]
25. Wang, D.; Chen, J. Supervised Speech Separation Based on Deep Learning: An Overview. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1702–1726. [CrossRef]
26. Zhang, Y.; Liu, X. Deep Encoder/Decoder Dual-Path Neural Network for Speech Separation in Noisy Reverberation Environments. *EURASIP J. Audio Speech Music. Process.* **2023**, *2023*, 1–15. [CrossRef]
27. Tzinis, E.; Wang, Z.; Smaragdis, P. Sudo rm-rf: Efficient networks for universal audio source separation. In Proceedings of the 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP), Espoo, Finland, 21–24 September 2020; pp. 1–6.
28. Tzinis, E.; Wang, Z.; Jiang, X.; Smaragdis, P. Compute and Memory Efficient Universal Sound Source Separation. *J. Signal Process. Syst.* **2022**, *94*, 245–259. [CrossRef]
29. Hu, X.; Li, K.; Zhang, W.; Luo, Y.; Lemercier, J.M.; Gerkmann, T. Speech Separation Using an Asynchronous Fully Recurrent Convolutional Neural Network. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22509–22522.
30. Park, H.J.; Friston, K. Structural and functional brain networks: From connections to cognition. *Science* **2013**, *342*, 1238411. [CrossRef]
31. Li, K.; Yang, R.; Hu, X. An efficient encoder-decoder architecture with top-down attention for speech separation. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
32. Luo, Y.; Mesgarani, N. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Trans. Audio, Speech Lang. Process.* **2019**, *27*, 1256–1266. [CrossRef] [PubMed]
33. Li, K.; Hu, X.; Luo, Y. On the Use of Deep Mask Estimation Module for Neural Source Separation Systems. *arXiv* **2022**, arXiv:2206.07347.
34. Subakan, C.; Ravanelli, M.; Cornell, S.; Bronzi, M.; Zhong, J. Attention is All You Need in Speech Separation. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021.
35. Chen, J.; Mao, Q.; Liu, D. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 2642–2646. [CrossRef]

36. Wang, Z.Q.; Roux, J.L.; Hershey, J.R. Alternative Objective Functions for Deep Clustering. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 686–690. [CrossRef]
37. Luo, Y.; Mesgarani, N. TasNet: Time-domain audio separation network for real-time, single-channel speech separation. *arXiv* **2018**, arXiv:1711.00541.
38. Patterson, R.; Nimmo-Smith, I.; Holdsworth, J.; Rice, P. *An Efficient Auditory Filterbank Based on the Gammatone Function*; APU Report 2341, MRC; Applied Psychology Unit: Cambridge, UK, 1988. Available online: https://www.pdn.cam.ac.uk/system/files/documents/SVOSAnnexB1988.pdf (accessed on 22 October 2024).
39. Ditter, D.; Gerkmann, T. A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 36–40. [CrossRef]
40. Zhu, W.; Wang, M.; Zhang, X.L.; Rahardja, S. A comparison of handcrafted, parameterized, and learnable features for speech separation. In Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 14–17 December 2021; pp. 635–639.
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.U.; Polosukhin, I. Attention is All you Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
42. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
43. Luo, Y.; Chen, Z.; Yoshioka, T. Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 46–50. [CrossRef]
44. Dehghani, M.; Gouws, S.; Vinyals, O.; Uszkoreit, J.; Kaiser, L. Universal Transformers. *arXiv* **2018**, arXiv:1807.03819.
45. Gholami, S.; Omar, M. Do Generative Large Language Models need billions of parameters? *arXiv* **2023**, arXiv:2309.06589.
46. Takase, S.; Kiyono, S. Lessons on Parameter Sharing across Layers in Transformers. In Proceedings of the Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), Toronto, ON, Canada, 13 July 2023; Sadat Moosavi, N., Gurevych, I., Hou, Y., Kim, G., Kim, Y.J., Schuster, T., Agrawal, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 78–90. [CrossRef]
47. Cosentino, J.; Pariente, M.; Cornell, S.; Deleforge, A.; Vincent, E. LibriMix: An Open-Source Dataset for Generalizable Speech Separation. *arXiv* **2020**, arXiv:2005.11262.
48. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 19–24 April 2015; pp. 5206–5210. [CrossRef]
49. ITU-R. *Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level*; Recommendation ITU-R BS.1770; ITU-R: Geneva, Switzerland, 2012
50. Roux, J.L.; Wisdom, S.; Erdogan, H.; Hershey, J.R. SDR—Half-baked or well done? *arXiv* **2018**, arXiv:1811.02508.
51. Lin, Y.; Wang, M.; Zhang, Z.; Wang, X.; Xiao, T.; Zhu, J. Understanding Parameter Sharing in Transformers. *arXiv* **2023**, arXiv:2306.09380.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.