electronics

Article

VAT-SNet: A Convolutional Music-Separation Network Based on Vocal and Accompaniment Time-Domain Features

Xiaoman Qiao [1], Min Luo [2], Fengjing Shao [1], Yi Sui [1], Xiaowei Yin [1] and Rencheng Sun [1,*]

[1] School of Computer Science and Technology, Qingdao University, Qingdao 266071, China
[2] Conservatory of Music, Qingdao University, Qingdao 266071, China
* Correspondence: src@qdu.edu.cn

Abstract: The study of separating the vocal from the accompaniment in single-channel music is foundational and critical in the field of music information retrieval (MIR). Mainstream music-separation methods are usually based on the frequency-domain characteristics of music signals, and the phase information of the music is lost during time–frequency decomposition. In recent years, deep learning models based on speech time-domain signals, such as Conv-TasNet, have shown great potential. However, for the vocal and accompaniment separation problem, there is no suitable time-domain music-separation model. Since the vocal and the accompaniment in music have a higher synergy and similarity than the voices of two speakers in speech, separating the vocal and accompaniment using a speech-separation model is not ideal. Based on this, we propose VAT-SNet; this optimizes the network structure of Conv-TasNet, which sets sample-level convolution in the encoder and decoder to preserve deep acoustic features, and takes vocal embedding and accompaniment embedding generated by the auxiliary network as references to improve the purity of the separation of the vocal and accompaniment. The results from public music datasets show that the quality of the vocal and accompaniment separated by VAT-SNet is improved in GSNR, GSIR, and GSAR compared with Conv-TasNet and mainstream separation methods, such as U-Net, SH-4stack, etc.

Keywords: single-channel music separation; deep learning; time-domain analysis; music information retrieval

Citation: Qiao, X.; Luo, M.; Shao, F.; Sui, Y.; Yin, X.; Sun, R. VAT-SNet: A Convolutional Music-Separation Network Based on Vocal and Accompaniment Time-Domain Features. *Electronics* **2022**, *11*, 4078. https://doi.org/10.3390/electronics11244078

Academic Editor: Daniel Morris

Received: 28 October 2022
Accepted: 6 December 2022
Published: 8 December 2022

1. Introduction

With the development of the internet, massive amounts of music data have emerged online, and music in audio and video software is loaded with a tremendous amount of information. The effective extraction of this information has undergone vital research in the field of digital music. However, the music on the internet and in real life is usually single-channel music, that is, a piece of music that comprises two kinds of music signals: the vocal and the instrumental accompaniment. Additionally, because the vocal and the accompaniment of the same piece of music can display great similarity and synergy in rhythm and melody, it is not always easy to extract them clearly and can be quite challenging to obtain the required information from the music. Separating the vocal from the accompaniment in single-channel music is the primary concern of music information retrieval (MIR) [1], which contributes to music genre classification [2], melody extraction [3], singing voice detection [4], singer recognition [5], and other research in the field of music.

The methods of music separation can be roughly summarized by the following four categories: The first is the music-separation method based on computational auditory scene analysis (CASA), which follows the principle of using a computer system to simulate the laws of human-ear hearing for modelling so that the model is able to extract the target signal from mixed music signals [6–9]. The second is the music-separation method based on non-negative matrix decomposition (NMF), which contributes significantly to research related to speech enhancement and speaker extraction, and is being gradually applied to the field

of music separation, decomposing the music spectrum matrix for separation [10–12]. The third is the separation method based on music melody and its periodic characteristics, such as modelling the repetition structure based on the beat spectrum in music (REPET) [13–15]. The fourth is the music-separation method based on neural networks [16–19]. With the rise of neural networks, deep learning methods, along with the previous classical music-separation methods, are able to better solve the problem.

Regarding the neural network music-separation method, Wang et al. defined the separation problem as a binary classification problem and obtained an IBM (Ideal Binary Mask) by performing binary classification through a support vector machine. The disadvantage of this method is that if the estimation is wrong, it will lead to too much loss of information, and the timing correlation information of the audio cannot be obtained by using the DNN structure [16]. In order to circumvent the shortcomings of IBM and DNN, Huang et al. applied recurrent neural networks to the separation task, used an IRM (Ideal Ratio Mask) to optimize the predicted accompaniment and singing voice, and achieved good separation results [17]. Jansson et al. applied the image segmentation model U-Net to the field of music separation, analyzed the spectrogram of mixed audio through the U-Net structure, and predicted the time–frequency mask corresponding to a single sound source in order to achieve the separation of sound sources [18]. Park et al. proposed the separation of the accompaniment and the vocal using a Stacked Hourglass Network (SH-4stack), which learns features from spectrogram images across multiple scales to generate each music source mask [19].

It can be seen that the current music-separation method is still in the frequency-domain research stage. The idea of the frequency-domain music-separation method is to convert music waveform data into a spectrogram, and then, learn its features, which can be understood as essentially turning the music problem into an image problem. Therefore, some models in the image field can be improved in order to solve the problems of music recognition, classification, and separation. However, music is composed of waveform data with temporal attributes, and its features cannot be fully expressed by images alone. As shown in Figure 1, the frequency-domain music-separation method adopts time–frequency decomposition to obtain the amplitude spectrum and phase spectrum of the mixed signal.

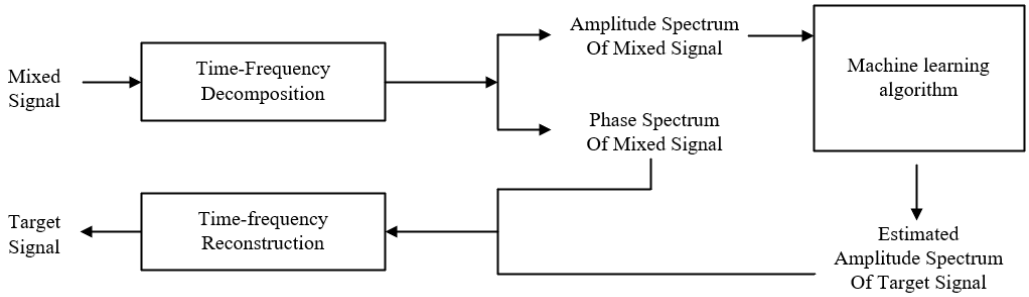

Figure 1. The limitation of the frequency-domain music-separation method.

Although the frequency-domain music-separation method can achieve the separation of the vocal and accompaniment, time–frequency reconstruction needs to combine the estimated amplitude spectrum of the target signal with the phase spectrum of the mixed signal due to the limitation of the frequency-domain music-separation method causing phase loss. Therefore, the traditional music-separation method will inevitably lose certain unique music features, resulting in music distortion and poor separation results.

Regarding the aforementioned issue, the main scope of this paper is as follows:

1. In order to avoid the defects of the traditional frequency-domain separation method, we build a time-domain music-separation model, VAT-SNet, which uses the music waveform data as input directly.
2. To mine the acoustic features of music, VAT-SNet transforms the waveform music into a nonlinear latent space with deep stacked convolutional layers in the encoder.

3. Aiming to improve the accuracy of the target music mask, a music feature extractor is designed in the auxiliary network based on the residual network, emphasizing the differences in data features between the vocal and accompaniment and assisting the training of a TCN music separator in order to derive the mask.
4. We compare VAT-SNet with mainstream music-separation models on a public dataset to verify the effectiveness of VAT-SNet. The experimental results show that the proposed model in this paper displays a significant improvement in terms of the objective evaluation indicators, demonstrating high performance music separation.

The remainder of this paper is structured as follows: Section 2 introduces the principle of single-channel music separation and summarizes the separation steps of previous deep-learning models in regard to this problem. Section 3 describes in detail the network structure of the VAT-SNet, the functions of each module, and the improvements compared to the baseline model. Section 4 introduces the dataset and environment related to the experiment and presents the experimental results. Finally, Section 5 presents the conclusions of this paper.

2. Related Principles

The purpose of single-channel music separation is to separate the vocal and accompaniment parts of the mixed music signals, that is, to separate two source signals from one mixed observation signal. This study conforms to the principle of independent component analysis (ICA) [20], and its mathematical model is expressed as follows:

$$x(t) = As(t) + n(t), \quad t = 1,2, \ldots, \mathrm{T} \tag{1}$$

where $x(t)$ is the observed signal vector, $\boldsymbol{A}$ denotes the mixing matrix, $s(t)$ represents the unknown signal vector from m sources, and $n(t)$ indicates the noise vector. Single-channel music separation is based on the blind source separation problem of the single-channel linear instantaneous mixed signal. Usually, only two source signals are considered, and the noise term $n(t)$ is ignored. Therefore, the mathematical model of single-channel music separation for linear instantaneous mixing can be defined as:

$$x(t) = As(t), \quad t = 1,2, \ldots, \mathrm{T} \tag{2}$$

where $s(t) = [s_1(t), s_2(t)]$ represents the signal vector from the two target music signals, and $s_1(t)$ and $s_2(t)$ are independent of each other and conform to the non-Gaussian distribution. By estimating the matrix A, and then, calculating the inverse matrix W of A, the independent components can be obtained using the following formula:

$$s(t) = Wx(t), \quad t = 1,2,\ldots,\mathrm{T} \tag{3}$$

It can be seen from the above that the key point of the music-separation model is to learn a mapping function from the mixed music signal to the target music signal [21], which has the same effect as W in Equation (3). In the study of separation problems, the mapping function is usually denoted as "Mask". Single-channel music separation belongs to the category of supervised learning, which extracts features from the true value of the target music signal and creates the corresponding mask. Based on the Ideal Binary Mask or Ideal Ratio Mask, the target music signal can be obtained by masking the interference signal in the mixed music signal.

Traditional music-separation research usually converts the time-domain signal of music into a two-dimensional frequency-domain signal and learns a mapping function corresponding to the frequency-domain characteristics of the target music. That is, previous research on music separation has focused on its frequency-domain characteristics, although music comprises time-domain signals. Existing deep learning models for music separation can generally be summarized as consisting of the following three steps:

1. Signal processing: The mixed music time-domain signal is decomposed into a two-dimensional frequency-domain signal representation using the time–frequency decomposition method. This process usually adopts the short-time Fourier transform (STFT) method to obtain the amplitude spectrum and phase spectrum of the mixed music signals.
2. Model training: By extracting and learning the transformed signal features and using the real values of the target music in the training set, a mapping function from the mixed music features to the target music is learned by the machine learning algorithm.
3. Waveform rebuilding: The mapping function acts on the amplitude spectrum of the mixed music signals in order to obtain the amplitude spectrum of the target music, which is then combined with the phase spectrum of the mixed music to obtain the waveform signal of the target music through the inverse Fourier transform.

3. Time-Domain Music-Separation Model

Due to the problems of traditional music-separation methods [22], as mentioned in Section 1, we no longer conduct research based on the frequency-domain signal of music but directly model the original waveform signal of music. The time-domain separation process is roughly the same as the frequency-domain separation process, which can also be divided into three parts: signal processing, model training, and waveform rebuilding. The difference is that the time-domain separation model directly performs convolutional coding operations on the music waveform data in the signal processing part without the need for time–frequency decomposition. Therefore, the problem of phase loss is not involved in the waveform reconstruction part, and only the corresponding deconvolution decoding operation is required.

This training method for feature extraction based on waveform signals has achieved breakthrough results in the speech-separation model Conv-TasNet [23] proposed by Luo et al. A schematic diagram of Conv-TasNet is shown in Figure 2. As a result, the time-domain convolution training method for waveform signals has attracted the attention of scholars in related fields. Based on Conv-TasNet, time-domain speech enhancement, speech recognition, multi-modal speech separation, and other models have been extended and applied in various fields [24–27]. Its higher performance further illustrates the feasibility and superiority of convolving the waveform signal directly.

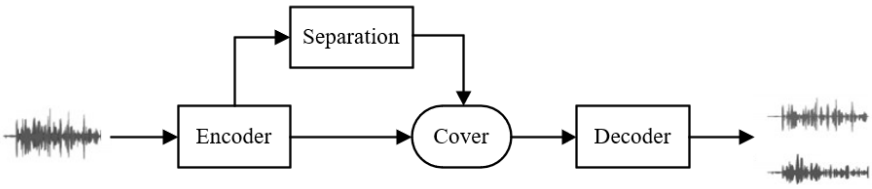

Figure 2. A schematic diagram of Conv-TasNet.

Music separation and speech separation have differences in the data characteristics of their separation results. The vocal and accompaniment in music are often strongly correlated, and both have the same periodic characteristics of pitch, melody, and rhythm. Due to the diversity of the acoustic features of music, the data distribution of music over time is more drastic and uneven than that of speech. In practical applications, the vocal and accompaniment are often parallel over time, while the voices of two speakers in speech separation are alternately vocalized, so separating the vocal and accompaniment is a greater challenge. The purity of the target music obtained is limited when simply separating the vocal and the accompaniment using the speech-separation model. Music separation and speech separation are consistent in the data types of the separation results, so the research idea based on the time-domain speech signal is instructive for the study of music separation. Therefore, based on the research idea in [23], we propose a time-domain vocal and accompaniment separation model, VAT-SNet, which optimizes the encoder–

decoder network structure to extract deep musical features and adds an auxiliary network to emphasize the difference between vocal and accompaniment data features.

3.1. Overall Framework of VAT-SNet

VAT-SNet mainly comprises two parts: the main network and the auxiliary network. Figure 3 shows the structure diagram of VAT-SNet. The middle column of the diagram is the main network, which takes the single-channel mixed music X as the input to generate the mask that corresponds to the target music. First, the time-domain music signal X is transformed into the processed music signal Y through the sample-level deep convolution of the encoder. The music signal Y and the corresponding music representation embedding generated by the auxiliary network are jointly trained in the separator to output the time-domain mask that corresponds to the target music. Then, the music signal Y is multiplied by the corresponding points of the mask to estimate the target music signal while suppressing other interference signals. Finally, the target music signal is deconvolved by the decoder to obtain pure waveform music in the time domain, and the output music file type of the model is consistent with the input music file type.

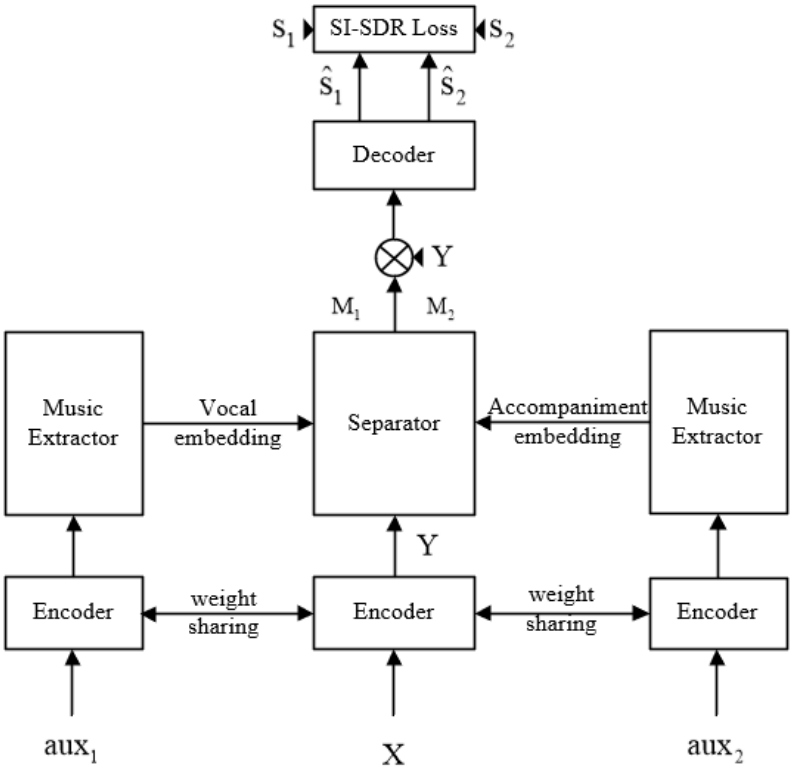

Figure 3. A schematic diagram of VAT-SNet.

The structure on both sides of the main network is an auxiliary network, which takes pure vocal music and pure accompaniment music as inputs, generates the embedding that corresponds to the target music by learning music features, and embeds them into the main network in order to constrain the weights of the corresponding masks, thus affecting the solution space of the model. Compared with Conv-TasNet, without adding the auxiliary network, the masks generated by the separation network in this paper are more accurate and have a significant inhibitory effect on interference signals.

In the model training stage, the gap between the predicted and true values is calculated using the loss function (SI-SDR Loss), and the masks are further modified in the subsequent training until the model achieves a smooth result. In the model inference stage, the vocal and accompaniment in music can be separated by inputting the single channel music into the model.

3.2. The Main Network of VAT-SNet

The main network focuses on generating the "Mask" and comprises three parts: the encoder, decoder, and separator. The separator is mainly composed of the time-domain convolutional network block (TCN-Block), and the other is the processing part that fuses the embedding generated by the auxiliary network with the main network data multiple times. The following is a detailed description of these modules.

3.2.1. Encoder

In MIR tasks, waveform music is usually input into the model after time–frequency decomposition, which is the signal processing part of traditional music separation. In [28], Lee concluded that a model set with a sufficient number of convolutional layers and convolutional kernel channels can tap deeper into acoustic features and effectively improve the model performance.

Inspired by [28], VAT-SNet feeds wave music into the encoder without any preprocessing. First, the original music signal is cut into K pieces of music with partial overlap and length L by the encoding module. $x_k \in R^{1\times L}$ represents each piece of music, and $k = 1, \ldots, K$ represents the index of the piece of music. Compared to the one-layer convolution of the baseline-model speech encoder, VAT-SNet deepens the number of encoder convolution layers and optimizes them. The encoder of VAT-SNet uses J-layer convolution. The first layer of convolution is a 1D convolution with a convolution kernel the size of L, a step the size of S = L/2, and an N number of convolution kernel channels to linearly transform x_k into the N-dimensional representation Y_1:

$$Y_1 = U_1 X \tag{4}$$

where $X \in R^{L\times K}$ is composed of all music clips x_k, $U_1 \in R^{N\times L}$ denotes the first layer of convolution with N channels (filters), and $Y_1 \in R^{N\times K}$ represents the N-dimensional representation of the music waveform data in the latent space after the first layer of convolution.

The remaining J-1 layers of convolution have the same structure. Each layer consists of a one-dimensional convolution with a kernel size of 3, a stride of 1, and an N number of convolution channels and takes PReLu as the activation function:

$$Y_j = \mathrm{PReLU}(U_j * Y_{j-1}) \tag{5}$$

where $*$ denotes the convolution operator, $j = 2, 3, \ldots, \mathrm{J}$ represents the layer index, $U_j \in R^{N\times N\times 3}$ represents the convolution operation of the layer, and $Y_j \in R^{N\times K}$ is the layer output. $Y_\mathrm{J} \in R^{N\times K}$ obtained by the convolution of the J-th layer is the final output value of the encoder in Figure 3.

Compared with speech data, the vocal and accompaniment in music have richer acoustic features, such as melody, rhythm, and beat. Using deep stacked convolutional layers in the encoder can transform waveform music into nonlinear latent space hierarchically. This sample-level convolution ensures a longer time-sensitive field while making the encoded processed music signal more conducive to the subsequent feature extraction of the target music, effectively improving the model's separation performance.

3.2.2. Separator

The separator takes the music signal processed by the encoder as input. It combines the embedding from the auxiliary network to generate the mask, which corresponds to the target music. Since the encoder no longer uses time–frequency decomposition for preprocessing, the mask in VAT-SNet is different from that in the frequency-domain music-separation method. The frequency-domain mask can be understood as the distribution of the target music signal in the mixed music signal. In VAT-SNet, the mask can be understood as the weight value of the target music signal in the mixed music signal.

Figure 4 shows the internal structure diagram of the separator. The encoded music signal Y and the embedding of target music generated by the auxiliary network are present

in this part for multiple fusions, and the characteristics of the target music are fully learned by this network to obtain a more accurate mask. Taking the generation of the vocals mask M_1 as an example, first, the encoded music signal Y is concatenated with the vocal embedding and point-by-point convolution is performed. After PReLU activation and normalization (GLN), it is processed by the temporal convolutional network block (TCN-Block). In this moment, the first fusion convolution ends. Then, the first fusion convolution result is spliced with the vocal embedding and the fusion convolution is performed again. After repeating the fusion convolution operation a total of Z times, Sigmoid activation is used to generate the vocal mask M_1.The generation of the mask can be summarized as Equation (6):

$$M_i = F\left(Y, E_i\right) \tag{6}$$

where $i = 1, 2$, M_1 and M_2 represent the vocal mask and the accompaniment mask, E_1 and E_2 represent the vocal embedding and the accompaniment embedding, and F represents the total fusion convolution operation of the mixed music signal Y and the target music embedding. The target music embedding is generated by the auxiliary network, and the specific content is described later in this paper.

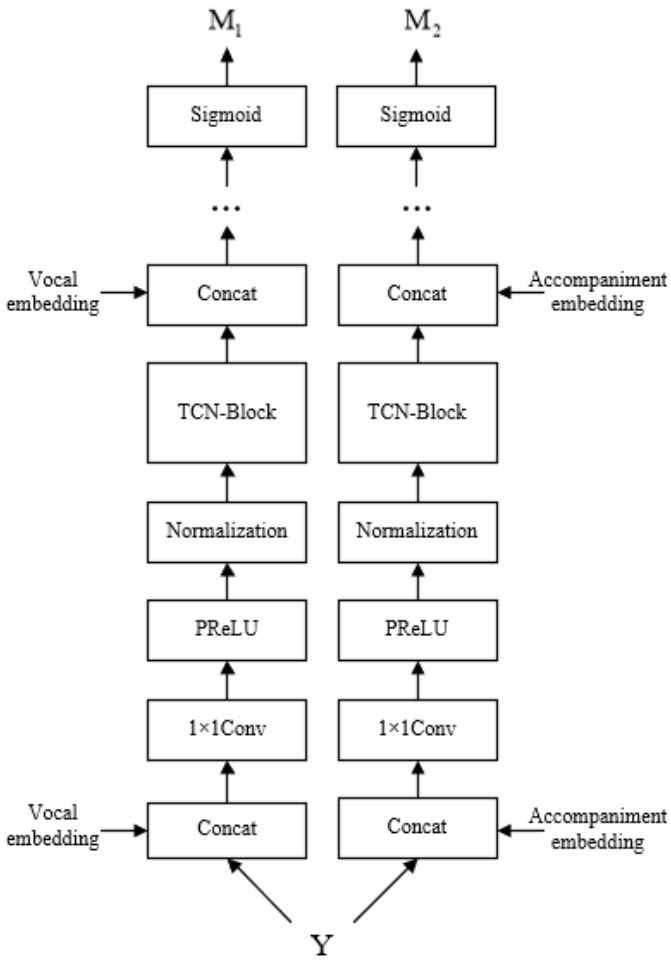

Figure 4. Internal structure diagram of the separator.

Figure 5 is a schematic diagram of the TCN-Block in the separator, the core of which is TCN [29]. Compared with RNN or CNN, VAT-SNet using TCN may be more suitable for processing the various phase data in the time-domain representation of music, and the parallel processing method of TCN can improve the operation efficiency effectively. In order to keep the size of the model small and reduce the number of parameters and computational cost, the structure of the baseline model is continued in TCN-Block, and a

depth-wise separable convolutional network [30] is used in TCN to replace the standard convolutional network. The expansion factor in the TCN-Block increases exponentially, ensuring that enough receptive fields are obtained in the feature extraction operation, which aligns with the long-term dependence of music signals.

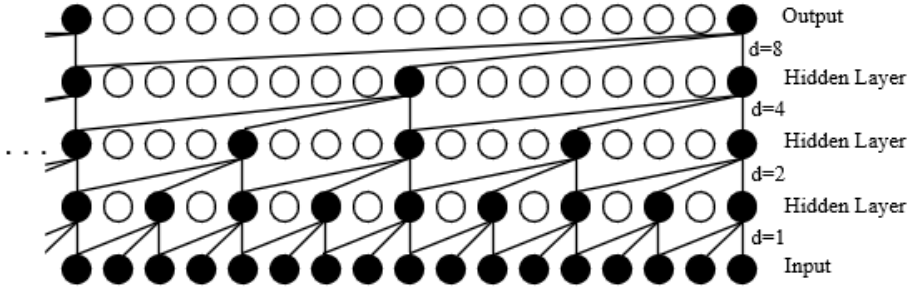

Figure 5. A schematic diagram of TCN-Block.

3.2.3. Decoder

The mask generated by the separator acts on the music signal obtained by the encoder. After being processed by the decoder, the vocal and accompaniment in the original mixed music can be obtained, as shown in Equation (7):

$$\hat{s}_i = V(Y \odot M_i) \tag{7}$$

where $\hat{s}_i \in R^{1\times L}$ and $i = 1, 2$, so $\hat{s}_1$ and $\hat{s}_2$ represent the rebuilt vocal waveform and accompaniment waveform; $\odot$ denotes the multiplication of the encoded original mixed music signal Y by the corresponding point of the corresponding target music mask M_i, and V represents the process whereby the decoder rebuilds the target music waveform.

In VAT-SNet, the decoder has a J-layer deconvolution operation that is symmetrical to the encoder. Its function is to rebuild the waveform music according to the results of the separator. The former J-1 layers of the decoder correspond to the latter J-1 layers of the encoder, and each layer consists of a 1D transposed convolution with a kernel size of 3, a stride of 1, and a PReLU activation function. The last layer of the decoder corresponds to the first layer of the encoder and performs deconvolution processing with a convolution kernel size of L and a step size of S = L/2. Since the music in the encoder is truncated into a plurality of partially overlapping music segments, the decoder needs to add the partially overlapping reconstructed segments to generate the final target music waveform.

VAT-SNet uses deep stack convolution in the encoder to replace the common STFT, which converts the signal into the latent space and avoids the loss of phase information. Therefore, the mask in the separator is generated according to the complete features of the music signal. As stated previously, the mask can be understood as the weight value. Therefore, the weight function is applied to the encoded original mixed music signal to generate the target signal when masking the latent space. The deconvolution layer of the decoder replaces the common ISTFT, so the whole separation process of VAT-SNet does not involve the decoupling of the magnitude and the phase of the music.

3.3. The Auxiliary Network of VAT-SNet

The focus of the single-channel music-separation model is to learn a mapping function from the mixed music signal to the target music signal, that is, the mask. In the speech-separation problem, the two speakers' speech rates, rhythms, and voiceprints are different. However, in the music-separation problem, the vocal and accompaniment are two parts of the same piece of music. The vocal and accompaniment have greater synergy and similarity in melody, rhythm, and beat, so the accuracy of the mask needs to be further improved. VAT-SNet adds an auxiliary network to extract the deep acoustic features of the target music; moreover, it generates vocal embedding, which corresponds to the vocal features, and BGM embedding, which corresponds to the accompaniment features, and fuses them

into the convolution process of generating the target music mask. The auxiliary network enables VAT-SNet to focus on the data features of the current target music during training, effectively optimizing the weights.

In order to represent the encoded mixed music signal and the encoded target music reference signal in the same latent feature space and to facilitate the fusion of the mixed music signal Y and the embedding of the target music in the separator, VAT-SNet sets up the same network structure and its weights are shared between the encoders of the main and auxiliary networks. After encoding the target music reference signal, the encoding result is put into the music extractor. Figure 6a shows a schematic diagram of the music extractor. The first step in utilizing the music extractor is channel-wise normalization, which performs a normalization operation on the reference music sequence. This operation does not change the data shape but changes the data distribution so that the drastically changing music is evened out. Then, the normalized music sequence is processed via 1D convolution and ResNet to extract its features. Finally, the 1D convolution projects the extracted musical features onto a fixed-dimensional embedding space.

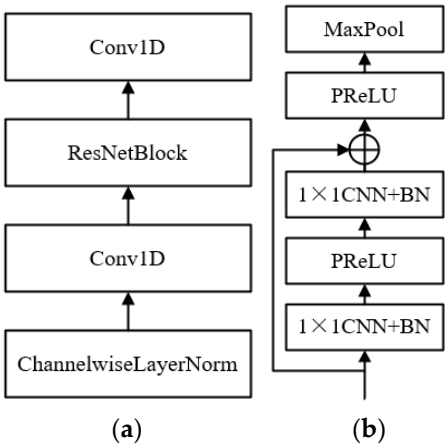

Figure 6. Internal structure diagram of the auxiliary network. (**a**) Music extractor (**b**) ResNetBlock.

The ResNet in the music extractor consists of ResNetBlocks with a number of R. As shown in Figure 6b, one ResNetBlock has two convolutional layers with a kernel size of 1×1. Each convolutional layer is followed by batch normalization (BN layer) and nonlinear activation (PReLu). ResNetBlock sets jump-join before the second PReLu and one layer of maximum pooling (MaxPool) after the second PReLu. Due to the characteristics of the ResNet, the jumping connections in the ResNet protect the integrity of the music-related information in the data and give the model stronger generalization capability compared to the original convolution.

3.4. Loss Function

The loss function is designed based on the SI-SNR (scale-invariant signal-to-noise ratio), which calculates the loss of the estimated value of the target source music signal obtained from the decoder with the true value. SI-SNR calculates the ratio of the original signal to the separation error, and the higher the value, the smaller the separation error and the better the obtained separation performance. The loss function needs to be minimized during model training, so VAT-SNet takes the negative SI-SNR as the loss function. The process of minimizing the loss function maximizes the SI-SNR and optimizes the separation performance. The SI-SNR formula is defined as:

$$\begin{cases} s_{\text{target}} = \frac{\langle \hat{s}, s \rangle s}{\|s\|^2} \\ s_{\text{noise}} = \hat{s} - s_{\text{target}} \\ \text{SI-SNR} = 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|s_{\text{noise}}\|^2} \end{cases} \tag{8}$$

where $\hat{s}$ represents the estimated target music signal, s represents the original pure target music signal, and $\|s\|^2 = \langle s, s \rangle$ represents the signal power and the inner product. $\hat{s}$ and s need to be normalized to zero mean before the formula is calculated to ensure scale invariance.

4. Experimental Results and Analysis

4.1. Experimental Environment and Dataset

The experiment in this paper was carried out using a CPU Intel Xeon W-2133 with six cores and twelve threads at 3.6 GH, a memory size of 32 GB, GPU RTX3090, CUDA11.1, Python 3.6, and pytorch 1.7.1. The experiment used the public dataset MIR-1K, released by Hsu (http://sites.google.com/site/unvoicedsoundseparation/mir-1k (accessed on 1 September 2021)). The dataset consists of 1000 16 kHz sampled music clips professionally cut from 110 Chinese karaoke songs, with durations of 4 s to 13 s. In our experiment, we cut all music clips to 4 s uniformly, and all waveforms were resampled to 8 kHz. A total of 800 music clips in MIR-1K were set as the training dataset, and the other 200 music clips were set as the evaluation dataset to evaluate the model performance.

4.2. Experimental Setup

The experiment learns 100 epochs on the training dataset, and the initial learning rate is set to 10^{-3}. If the accuracy does not improve within three consecutive epochs, the learning rate is halved. The normalization process uses global layer normalization (GLN). The contents of Table 1 are the settings of the remaining hyperparameters in VAT-SNet.

Table 1. Training model parameter setting.

Symbol	Quantity	Parameter Value
N	Number of channels in the encoder	512
J	Number of convolutional layers in the encoder	4
L	Size of convolution kernel in the encoder	16
B	Number of channels in 1 × 1 convolutional blocks	128
H	Number of channels in convolutional blocks	512
Q	Size of convolution kernel in 1D convolution	3
R	Number of ResNetBlocks in the music extractor	3
Z	Number of fusion times of the encoded music Signal Y and the target music embedding	4

4.3. Evaluation Indicators

Music-separation research usually uses objective evaluation metrics, and the Févotte separation toolbox (Blind Source Separation Evaluation, BSS _EVAL) provides a set of metrics aimed at quantifying the quality of the separation between the source signal and its estimated signal, the principle of which is to decompose the estimated signal into four parts, as follows:

$$\hat{s}(t) = s_{\text{target}}(t) + e_{\text{interf}}(t) + e_{\text{artif}}(t) + e_{\text{noise}}(t) \tag{9}$$

where $s_{\text{target}}(t)$ represents the part of the estimated signal of the target sound source that is related to the original pure signal, $e_{\text{interf}}(t)$ represents the interference components of other sound sources, $e_{\text{artif}}(t)$ represents the system error, and $e_{\text{noise}}(t)$ represents the disturbance noise. Disturbance noise is usually not considered in the separation research of vocals and accompaniments, that is, $e_{\text{noise}}(t)$ can be omitted. Therefore, the source-to-interference ratio (SIR) and the source-to-artifacts ratio (SAR) are defined as follows:

$$\text{SIR} = 10 \log_{10} \frac{\|s_{\text{target}}(t)\|^2}{\|e_{\text{interf}}(t)\|^2} \tag{10}$$

$$\text{SAR} = 10 \log_{10} \frac{\|s_{\text{target}}(t) + e_{\text{interf}}(t)\|^2}{\|e_{\text{artif}}(t)\|^2} \tag{11}$$

For the separation quality of one song, we use the signal-to-noise ratio (SNR) for evaluation, and the definition of SNR is shown in Equation (12). SNR is able to fully reflect the advantages and disadvantages of each algorithm in amplitude prediction and intuitively compare the separation performance of each algorithm. Comparing VAT-SNet with the traditional music-separation model based on this indicator, we can draw a more objective display of separation performance.

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left(\hat{Y}(i,j)\right)^2}{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left(Y(i,j)-\hat{Y}(i,j)\right)^2}\right) \tag{12}$$

where Y and $\hat{Y}$ represent the time–spectrum matrixes of pure signals and estimated signals, and m and n are the lengths of the frequency axis and time axis in the spectrum graph. $Y(i,j)$ and $\hat{Y}(i,j)$ are the amplitude values of the pure signal and the estimated signal in the (i,j) time–frequency unit.

To globally assess the separation quality of all music segments in the test set, the GSNR is defined and calculated as follows:

$$\mathrm{GSNR} = \frac{\sum_k w_k \cdot \mathrm{SNR}_k}{\sum_k w_k} \tag{13}$$

where SNR_k represents the signal-to-noise ratio of the k-th music; w_k represents the time weight, which is the duration of the k-th music; and GSAR and GSIR represent the values obtained by adding time weights to SAR and SIR. The calculation methods of GSAR and GSIR are the same as those of GSNR and will not be described in detail.

4.4. Evaluation Indicators

To verify the validity of VAT-SNet, we randomly selected one piece of music in the evaluation dataset and used VAT-SNet to separate the vocal from the accompaniment in that specific piece of music. In a real environment, human ears can hear that there is basically no accompaniment in the separated vocal and also no vocal in the separated accompaniment. Limited by the paper presentation, the pure music and the separated resulting music were visualized and printed as waveforms. As shown in Figure 7, the separated music waveform and the original pure music waveform are basically the same in shape.

Some music examples are available at https://qiaomusic.github.io/VAT-SNetDemo/ (accessed on 5 December 2022), which includes the mixed music and the separation results using VAT-SNet.

As mentioned, VAT-SNet makes several improvements on the basis of Conv-TasNet and continues to use the parameter configuration in [23] to achieve the optimal separation effect. Since VAT-SNet changes the structure of the original encoder and adds an auxiliary network to enhance feature extraction, several parameters were added compared to the original model. The specific parameter settings are shown in Table 1. We conducted experiments to verify the effect of parameters J and Z on model performance and size. We used the music dataset MIR-1K to train Conv-TasNet and VAT-SNet, respectively.

As shown in Table 2, the loss value obtained from Conv-TasNet is −9.73, and the loss value obtained from VAT-SNet is −10.87. It can be seen that VAT-SNet improves the SI-SNR value by 1.14 dB compared with the baseline model, and the ability of VAT-SNet for vocal and accompaniment separation is enhanced by about 11.7% compared with the original model. J represents the number of convolutional layers in the encoder, and the encoder of Conv-TasNet has only one layer of convolution, which means that J = 1 in Conv-TasNet.

We used stacked convolution layers in VAT-SNet to fully convolve the music sequences, and the loss values of VAT-SNet for J = {1,2,3,4,5,6} are listed in Table 2.

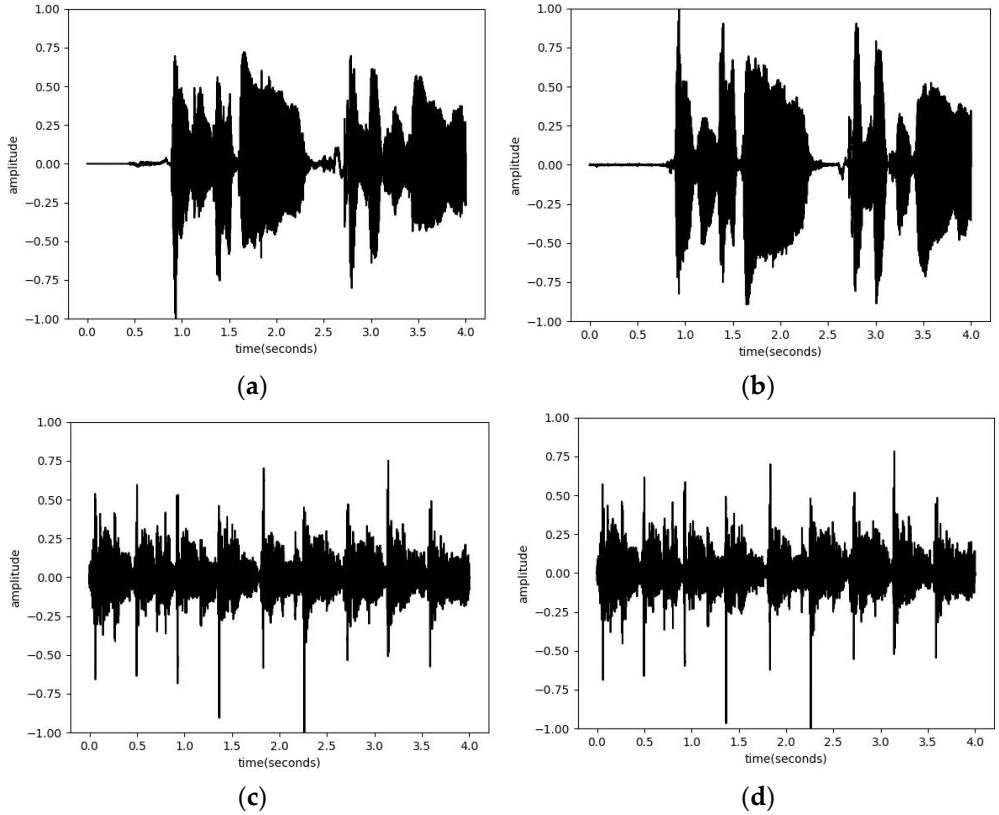

Figure 7. Comparison of waveforms before and after separation. (**a**) Primitive vocal waveform; (**b**) separated vocal waveform; (**c**) primitive accompaniment waveform; (**d**) separated accompaniment waveform.

Table 2. Training model parameter settings.

Method	J	Z	Size	SI-SNR
Conv-TasNet	1	/	5.6M	9.73
Conv-TasNet-4layers	4	/	5.6M	10.02
VAT-SNet-1layers	1	4	7.8M	10.55
VAT-SNet-2layers	2	4	7.8M	10.64
VAT-SNet-3layers	3	4	7.8M	10.75
VAT-SNet	4	4	7.8M	10.87
VAT-SNet-3Z	4	3	7.5M	10.77
VAT-SNet-5Z	4	5	8.7M	10.91
VAT-SNet-5layers	5	4	7.8M	10.70
VAT-SNet-6layers	6	4	7.8M	10.65

From the experiments, it can be seen that the separation performance of the model becomes better as the number of convolutional layers increases. The most positive separation performance of VAT-SNet is achieved at J = 4, which, in turn, verifies that deeper convolutional layers can extract richer acoustic features, and thus, improve the separation performance. However, as the number of layers deepens further, it has a negative impact on the model separation performance at J = 5 and J = 6. We conjecture that this is due to over-deep convolutional layers causing network degradation and overfitting, so we set J = 4 in VAT-SNet. We used the encoder and the decoder of the corresponding structure made of four convolutional layers stacked for Conv-TasNet, namely Conv-TasNet-4layer, and the resulting loss value was improved by 0.29 dB compared with the original Conv-TasNet.

VAT-SNet-4layer (VAT-SNet) improves about 0.85 dB compared with Conv-TasNet-4layer, which shows that the auxiliary network in VAT-SNet also has a positive effect.

Z represents the number of fusion times of the encoded music signal and the target music embedding. We set Z = 4 in VAT-SNet and perform experiments for Z = 3 and Z = 5. Although increasing the number of fusions can improve the model performance, it also increases the model size, so we ultimately set the fusion operation in VAT-SNet to be performed four times.

From the experimental results, it is clear that increasing the number of encoder and decoder network layers has almost no effect on the model size, and the reason for the larger size of VAT-SNet than Conv-TasNet is mainly the addition of the auxiliary network. The separation ability of Conv-TasNet on the music dataset MIR-1K is significantly inferior to that on the speech dataset WSJ0 in [23], illustrating the limitations of the baseline model in solving the vocal and accompaniment separation problem. As can be seen in the experimental results, the improved encoder structure and auxiliary network in VAT-SNet can effectively improve the SI-SNR.

In order to separate the vocal from the accompaniment in single-channel music, we proposed a time-domain separation model VAT-SNet. The other current time-domain-based separation techniques are aimed at the problem of speech separation, and they are also capable of separating the vocal and accompaniment in music to a certain extent. Therefore, we used the music dataset MIR-1K to train the other current time-domain separation methods, such as GALR [31], DPTNet [32], and DPRNN [33]. Table 3 shows a comparison of the SI-SNR values obtained by various models. It can be intuitively seen that VAT-SNet has a better ability to separate the vocal from the accompaniment in music compared with other time-domain methods.

Table 3. Comparison of time-domain methods.

Method	SI-SNR
GALR	8.01
DPTNet	8.56
DPRNN	9.46
Conv-TasNet	9.73
VAT-SNet	10.87

In other objective evaluation metrics, VAT-SNet has significantly improved compared to the baseline model and the classical music-separation method, as we will further demonstrate in the next section.

4.5. Separation Performance Comparison

To further illustrate the contribution of VAT-SNet to solving the single-channel music-separation problem, we compared the separation performance of VAT-SNet with classical music-separation methods (such as the algorithm proposed by Huang et al. [17], U-Net [18], and SH-4stack [19]) and the baseline model Conv-TasNet [23]. We randomly selected five music clips in the MIR-1K dataset for separation using different algorithms.

Figure 8 shows a comparison of the SNR between VAT-SNet and other methods after the separation of each music clip. It can be seen that VAT-SNet has at least a 5 dB improvement in SNR over other music-separation algorithms when separating the vocal and accompaniment and also has significant improvement compared to Conv-TasNet.

In addition, we used VAT-SNet and other separation algorithms to separate the vocal from the accompaniment in 200 music clips in the test set. Then, we evaluate the separation performance of each algorithm through three indicators: GSNR, GSIR, and GSAR. The results are shown in Table 4.

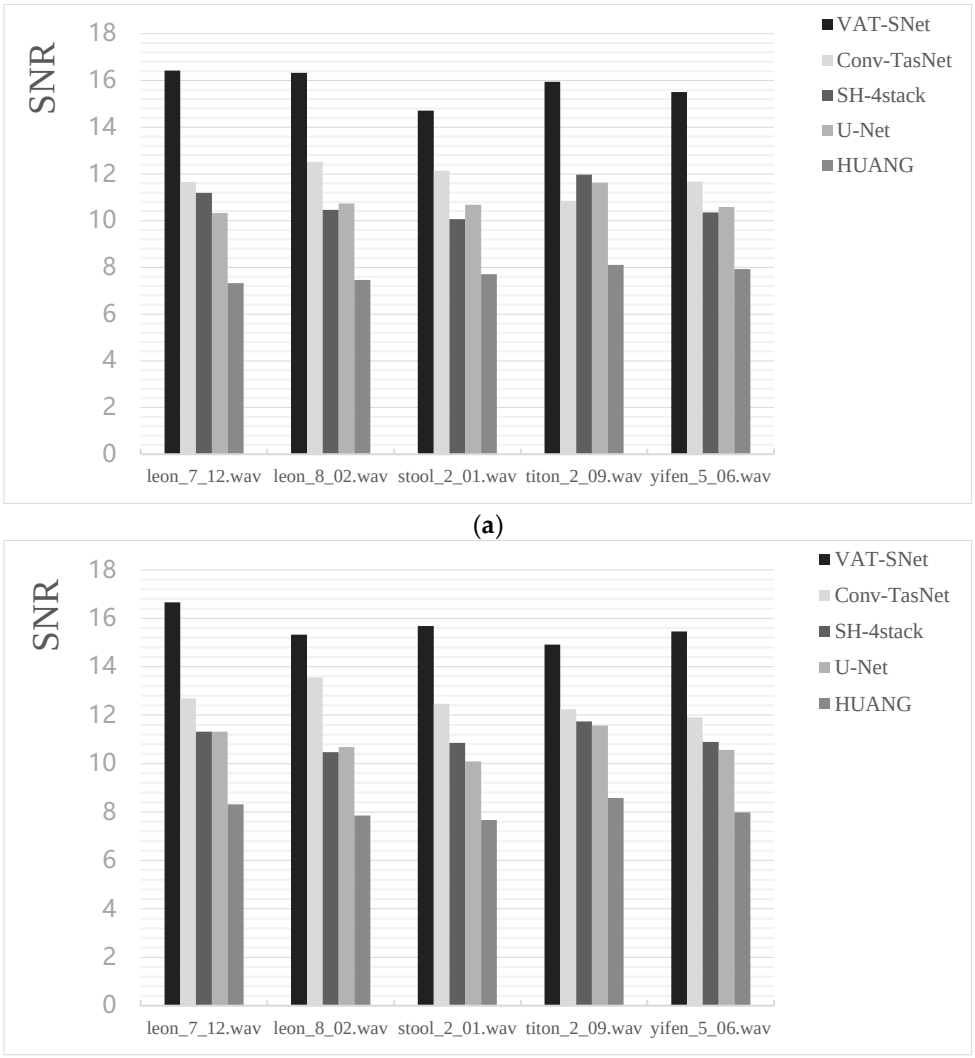

Figure 8. Comparison chart of the separation performance indexes of five random songs. (**a**) SNR(dB) of vocal separation contrast diagram; (**b**) SNR(dB) of accompaniment separation contrast diagram.

Part a of Table 4 shows the performance of the separated vocal music on various indicators. Compared with the baseline and other mainstream music-separation methods, VAT-SNet improves GSNR by about 3~8 dB, GSIR by about 7~13 dB, and GSAR by about 1~5 dB. Part b of Table 4 shows the performance of the separated accompaniment music on various indicators. Compared with the baseline and other music-separation methods, VAT-SNet improves GSNR by about 2~7 dB, GSIR by about 7~13 dB, and GSAR by about 1~5 dB. It can be seen that the quality of the vocal and accompaniment obtained through VAT-SNet separation is higher than other methods in GSNR, GSIR, and GSAR. In particular, GSIR is improved by at least 7.88 dB compared with the other models. The experimental results further validate the superiority of VAT-SNet over other separation methods.

The experimental data show that the speech-separation model Conv-TasNet has some applicability to music separation, but the separated vocal and accompaniment have no significant improvement in each separation evaluation index compared with the frequency-domain method. However, VAT-SNet is significantly better than the frequency-domain separation method and Conv-TasNet in the SNR obtained after separating five random music clips, and the GSNR, GSIR, and GSAR obtained after separating 200 music clips. Based on the above, we conclude that the deep stacked convolutional structure of VAT-SNet for the music signal and the residual network of the auxiliary network can extract deep

acoustic features, thereby effectively improving the model's ability to separate the vocal from the accompaniment.

Table 4. Comparison of separation performance of 200 music clips in MIR-1K (average result). (**a**) Separation results for the vocal. (**b**) Separation results for the accompaniment.

(a)			
Model	**Vocal**		
	GSNR	**GSIR**	**GSAR**
Huang	7.21	17.84	9.32
U-Net	10.09	11.96	11.30
SH-4stack	11.29	15.93	13.52
Conv-TasNet	12.16	17.38	14.29
VAT-SNet	14.57	25.53	15.02
(b)			
Model	**Accompaniment**		
	GSNR	**GSIR**	**GSAR**
Huang	7.49	16.89	9.97
U-Net	10.32	11.65	11.42
SH-4stack	12.07	15.21	14.95
Conv-TasNet	12.57	16.87	13.64
VAT-SNet	15.88	24.77	15.47

5. Conclusions

In this paper, a music-separation model based on a time-domain convolutional network, named VAT-SNet, is proposed. In the encoder of VAT-SNet, the common time–frequency decomposition is replaced by the convolution operation of the deep network stack. In the VAT-SNet separator, we designed a music extractor to generate the embedding of target music and incorporate the embedding into the time-domain convolutional network to generate a mask. Such joint learning improves the accuracy of the mask and compensates for the inadequate extraction of the acoustic features of music data by the baseline model. The mask of the target music is applied to the original mixed music and the waveform is reconstructed by the decoder to separate the vocal from the accompaniment in the music.

VAT-SNet is trained directly on the time-domain music waveform in a data-driven manner, which achieves better separation performance with lower latency and avoids the decoupling problems of the phase and amplitude common to traditional music-separation methods. It can be seen from the experimental results that the VAT-SNet in this paper can separate vocal and accompaniment signals while maintaining high quality and high purity. Related research based on time-domain music signals is still in its infancy. The separation performance of VAT-SNet verifies the feasibility and superiority of time-domain music signal modeling, but there is still room for improvement in the performance of VAT-SNet. The feature extraction method in [34–37] is instructive for our future research. In the future, we will continue this research based on music's time-domain signals and optimize the structure of VAT-SNet to achieve higher separation performance, which will then be applied to real-time music separation.

Author Contributions: Conceptualization, X.Q. and M.L.; methodology, X.Q. and R.S.; writing—original draft preparation, X.Q.; writing—review and editing, X.Q. and X.Y.; project administration, R.S. and F.S.; funding acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Youth Foundation of China (41706198).

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

References

1. Downie, J.S. Music information retrieval. *Annu. Rev. Inf. Sci. Technol.* **2003**, *37*, 295–340. [CrossRef]
2. Kum, S.; Nam, J. Joint detection and classification of singing voice melody using convolutional recurrent neural networks. *Appl. Sci.* **2019**, *9*, 1324. [CrossRef]
3. Salamon, J.; Gomez, E.; Ellis, D.P.; Richard, G. Melody Extraction from Polyphonic Music Signals: Approaches, applications, and challenges. *IEEE Signal Process. Mag.* **2014**, *31*, 118–134. [CrossRef]
4. You, S.D.; Liu, C.H.; Chen, W.K. Comparative study of singing voice detection based on deep neural networks and ensemble learning. *Hum. Cent. Comput. Inf. Sci.* **2018**, *8*, 34. [CrossRef]
5. Sharma, B.; Das, R.K.; Li, H. On the Importance of Audio-Source Separation for Singer Identification in Polyphonic Music. In Proceedings of the Conference of the International Speech Communication Association, Graz, Austria, 15–19 September 2019; pp. 2020–2024.
6. Smoliar, S.; Bregman, A.S. Auditory Scene Analysis: The Perceptual Organization of Sound. *Comput. Music J.* **1992**, *15*, 74. [CrossRef]
7. Hu, G.; Wang, D. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans. Neural Netw.* **2004**, *15*, 1135–1150. [CrossRef]
8. Hu, G.; Wang, D.L. A Tandem Algorithm for Pitch Estimation and Voiced Speech Segregation. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 2067–2079.
9. Yan, Z.; Wang, D.L.; Johnson, E.M.; Healy, E.W. A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions. *J. Acoust. Soc. Am.* **2018**, *144*, 1627–1637.
10. Benzi, K.; Kalofolias, V.; Bresson, X.; Vandergheynst, P. Song recommendation with non-negative matrix factorization and graph total variation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2439–2443.
11. Yoshii, K.; Itoyama, K.; Goto, M. Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 51–55.
12. Xiong, M.; Zhang, T.Q.; Zhang, T.; Yang, K. *Non-Negative Matrix Music Separation Method Combined with Hpss*; Computer Engineering and Design: Singapore, 2018.
13. Rafii, Z.; Pardo, B. A simple music/voice separation method based on the extraction of the repeating musical structure. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 221–224.
14. Rafii, Z.; Pardo, B. Repeating pattern extraction technique (REPET): A simple method for music/voice separation. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *21*, 73–84. [CrossRef]
15. Dogan, S.M.; Salor, O. Music/singing voice separation based on repeating pattern extraction technique and robust principal component analysis. In Proceedings of the 5th International Conference on Electrical and Electronic Engineering (ICEEE), Istanbul, Turkey, 5 May 2018.
16. Wang, Y.; Wang, D.L. Towards scaling up classification-based speech separation. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 1381–1390. [CrossRef]
17. Huang, P.S.; Kim, M.; Hasegawa-Johnson, M.; Smaragdis, P. Deep learning for monaural speech separation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 1562–1566.
18. Jasson, A.; Humphrey, E.; Montecchio, N.; Bittner, R.; Kumar, A.; Weyde, T. Singing voice separation with deep U-Net convolutional net-works. In Proceedings of the 18th International Society for Music Infor-mation Retrieval Conference (ISMIR), Suzhou, China, 23–27 October 2017; pp. 745–751.
19. Park, S.; Kim, T.; Lee, K.; Kwak, N. Music source separation using stacked hourglass networks. *arXiv* **2018**, arXiv:1805.08559.
20. Meinecke, F.; Ziehe, A.; Kawanabe, M.; Muller, K.R. Independent component analysis, a new concept? *Signal Process.* **1994**, *36*, 287–314.
21. Wang, Y.; Narayanan, A.; Wang, D.L. On training targets for supervised speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1849–1858. [CrossRef] [PubMed]
22. Wang, D.L.; Chen, J. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1702–1726. [CrossRef]
23. 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]
24. Xu, L.; Wang, J.; Yang, W.J.; Luo, Y.Y. Multi Feature Fusion Audio-visual Joint Speech Separation Algorithm Based on Conv-TasNet. *J. Signal Process.* **2021**, *37*, 1799–1805.
25. Hasumi, T.; Kobayashi, T.; Ogawa, T. Investigation of Network Architecture for Single-Channel End-to-End Denoising. In Proceedings of the European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021; pp. 441–445.
26. Zhang, Y.; Jia, M.; Gao, S.; Wang, S. Multiple Sound Sources Separation Using Two-stage Network Model. In Proceedings of the International Conference on Information Communication and Signal Processing (ICICSP), Shanghai, China, 24–26 September 2021; pp. 264–269.

27. Jin, R.; Ablimit, M.; Hamdulla, A. Speech Separation and Emotion Recognition for Multi-speaker Scenarios. In Proceedings of the International Conference on Pattern Recognition and Machine Learning (PRML), Chengdu, China, 22–24 July 2022; pp. 280–284.
28. Lee, J.; Park, J.; Kim, K.L.; Nam, J. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. *arXiv* **2017**, arXiv:1703.01789.
29. Bai, S.; Kolter, L.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:1803.01271.
30. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
31. Lam, M.W.Y.; Wang, J.; Su, D.; Yu, D. Effective Low-Cost Time-Domain Audio Separation Using Globally Attentive Locally Recurrent Networks. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021.
32. Chen, J.; Mao, Q.; Liu, D. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation. *arXiv* **2020**, arXiv:2007.13975.
33. 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.
34. Zhang, X.; Yu, Y.; Gao, Y.; Chen, X.; Li, W. Research on Singing Voice Detection Based on a Long-Term Recurrent Convolutional Network with Vocal Separation and Temporal Smoothing. *Electronics* **2020**, *9*, 1458. [CrossRef]
35. Chen, H.; Sui, Y.; Shang, W.L.; Sun, R.; Chen, Z.; Wang, C.; Han, C.; Zhang, Y.; Zhang, H. Towards renewable public transport: Mining the performance of electric buses using solar-radiation as an auxiliary power source. *Applied Energy* **2022**, *325*, 119863. [CrossRef]
36. Si, J.; Huang, B.; Yang, H.; Lin, W.; Pan, Z. A no-Reference Stereoscopic Image Quality Assessment Network Based on Binocular Interaction and Fusion Mechanisms. *IEEE Trans. Image Process.* **2022**, *31*, 3066–3080. [CrossRef] [PubMed]
37. Gao, Y.; Zhang, X.; Li, W. Vocal Melody Extraction via HRNet-Based Singing Voice Separation and Encoder-Decoder-Based F0 Estimation. *Electronics* **2021**, *10*, 298. [CrossRef]