Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 4, pp. 2960–2972, DOI 10.12694/scpe.v25i4.2899

APPLICATION OF UNET-SE-BISRU ALGORITHM FOR MUSIC SIGNAL PROCESSING IN MUSIC SOURCE SEPARATION

TAO ZHANG[*]

Abstract. At present, the use of time-domain deep learning for end-to-end neural network models has the problem of long training time and poor performance in music source separation. To address this issue, a U-network squeezing excitation bidirectional simple recursive unit model was proposed based on the deep extractor model. Replace Unet SE Bisru with Unet SE Bisru in the following text. This model improves the bidirectional long short-term memory network into a bidirectional simple recurrent unit, and then introduces attention mechanisms in the generalized encoding and decoding layers. The squeezing excitation block is used to selectively extract features based on the type of audio to be separated. Finally, group normalization is added after one-dimensional convolution, And its effectiveness was verified. The experimental results show that the signal noise distortion ratio in the improved model is 5.68 decibels compared to the bidirectional simple recursive unit value, which is higher than the 5.55 decibels of bidirectional long short-term memory. After adding the squeezing excitation module, the overall increase is about 0.1-0.5 decibels. In addition, in the model comparison, the three indicators of the improved model with the same number of channels were 5.68 decibels, 5.91 decibels, and 11.28 decibels, respectively, higher than the benchmark model. Compared with other music source classification models, the improved model has better comprehensive separation performance. Although some indicators are lower than the comparison model, the signal noise distortion ratio of drum and bass is 6.11 decibels and 6.36 decibels, which is better than the comparison model. Overall, the improved model has high performance in music source separation for music signal processing and can be effectively applied in practical music source separation

Key words: Deep learning; Music source; Unet-SE-BiSRU; SDR; Number of channels

1. Introduction.

Mixed audio is the most commonly encountered form of audio in production and daily life. When the current listener is interested in mixed audio, they need to separate the target audio source from it to understand its semantics [1]. The music source separation (MSS) method is a special problem of sound source separation. The actual speech content generally has harmony, specifically manifested as the frequency spectrum of different sound sources often overlapping, which makes solving the problem of MSS more challenging [2]. Effective separation of sound sources in mixed audio is the first issue to be addressed in the subsequent deep utilization of massive music resources, which has important research value and broad application prospects [3]. MSS is an important research direction in current audio signal processing (ASP). It aims to extract one or more sound sources from the sound source and effectively suppress other sound sources and noise [4]. However, traditional music ASP methods have limited complexity and insufficient expressive power in their models. The time domain deep learning of end-to-end neural network model (TDDL-E2ENN) has problems such as long training time and poor separation performance. Based on this, this study proposes the TDDL-E2ENN model, which is the U network Squeeze Exit Indirect Simple Recurrent Units (Unet-SE-BiSRU), on the basis of the current Deep Extractor for Music Source Separation model (Demucs) with the highest time domain separation performance. Its purpose is to solve the relevant problems of the current MSS method and provide theoretical support for its application in ASP.

The study consists of four parts in total. Part 1 is a summary and discussion of the current hybrid MSS methods. Part 2 is an analysis of the MSS method using the Unet-SE-BiSRU algorithm. The third part is to verify the effectiveness of Unet-SE-BiSRU. The fourth part is a summary of the entire article.

2. Related works.

The rapid development of modern music signal processing and computer technology has made the study of hybrid MSS one of the increasingly popular research topics in the ASP field. Unlike traditional MSS methods, which mainly focus on the acoustic characteristics of sound sources for studying songs

[*] School of Music, Henan Vocational Institute of Arts, Zhengzhou, 451464, China (unwenzhy1234@163.com)

and instruments in mixed audio, in recent years, the growth of big data and computing technology has led to the introduction of more modern computing technologies into the field of audio processing, and has also brought new ideas to MSS [5-6]. Chen et al. addressed the problem of poor separation performance of multi-source mixed audio sources by utilizing three sets of channels to train the audio set based on deep learning, thereby improving separation performance and expanding its generalization [7]. Huang et al. proposed a compatible prediction model based on self supervised and semi supervised learning to address the issue of low prediction performance of audio element compatibility in current mixed audio MSS. It not only enhanced the classification system, but also enhanced the compatibility of mixed audio systems [8]. Ma B et al. proposed a method for solving blind source separation in frequency domain convolution based on typical correlation analysis to address the related issues in mixed hidden audio MSS. It effectively improved the effectiveness of convolutional mixed audio [9]. Zhou et al. proposed the optimal fusion structure based on multimodal analysis to effectively improve the performance of visual and audio separation in mixed audio MSS [10].

In addition, Colonel et al. proposed a method for constructing multi track reverberation parameters based on the reverberation module architecture to address the issue of poor actual performance of mixed audio MSS. This method not only improved separation performance but also improved the effectiveness of mixing perception [11]. Gupta et al. comprehensively discussed the perception, control, and rendering capabilities of enhanced and mixed reality technologies in audio MSS. It provided a theoretical basis for its application in mixed audio MSS [12]. Srinivasamurthy et al. explored the MSS problem of two music datasets in India through data collection, annotation, and organization processes, providing theoretical support for MSS and melody extraction [13]. Sheeja et al. proposed a lightweight convolutional neural network quantum teaching optimization algorithm based on discrete Fourier transform for MSS problems under noise interference. This algorithm effectively improved separation performance while eliminating blind separation noise [14].

From the research of domestic and foreign scholars, it can be seen that the main method to solve the problem of music source separation is deep learning, but the optimal method still chooses mask operation. In the presence of multiple instruments, there is a risk of information loss. Due to the fact that the signal cannot be simply shielded, the separation effect of the drum and bass sound sources is not good. In addition, the time-domain end-to-end model used to solve the problem of music source separation is still in the exploratory stage. Although its separation performance should theoretically reach the highest level, there is still a certain gap in performance compared to the frequency domain mask method. At the same time, the current optimal end-to-end model has problems such as long training time and poor separation performance [15]. Therefore, based on Demucs, the Unet-SE-BiSRU proposed in this study is innovative, which effectively solves three problems and improves the performance of MSS.

3. Analysis of MSS Method Based on Unet-SE-BiSRU Algorithm.

The current music source signal mainly consists of mixed music audio, which is crucial for improving MSS methods. Therefore, this section mainly analyzes the mixed music audio and deep neural network models in MSS, and based on this, introduces SE and BiSRU modules to construct an improved algorithm model based on the U-shaped network.

3.1. Hybrid Music Audio and Deep Neural Network Model in MSS.

The traditional music ASP method has limited model complexity and insufficient expression ability, while using TDDL-E2ENN model also has problems such as long training time and poor separation effect. Based on Demucs, Unet-SE-BiSRU is proposed. Currently, most mixed music audio is composed of a combination of human voice and instrument sound, and the time-frequency characteristics of the sound source signal are of great significance for the correct selection, understanding, and improvement of MSS technology. Therefore, this study first analyzes the characteristics of mixed music audio sources before proposing Unet-SE-BiSRU.

The human voice in mixed music audio is usually similar to a voice signal, which also includes pitch, intensity, length, and timbre. The pitch depends on the frequency at which the vocal organ vibrates within one second. Sound intensity is determined by amplitude and increases as the amplitude increases. The sound length is determined by the time it takes for the continuous vibration of the vocal organ. Timbre is a major feature that distinguishes different pronunciation units. In addition, there are two important acoustic characteristics in speech signals: pitch frequency and resonance peak. In terms of human auditory perception, differences in human voice are mainly reflected in three aspects: amplitude, duration, and fundamental frequency [16-17]. The instruments in audio can be divided into string instruments, wind instruments, and percussion instruments

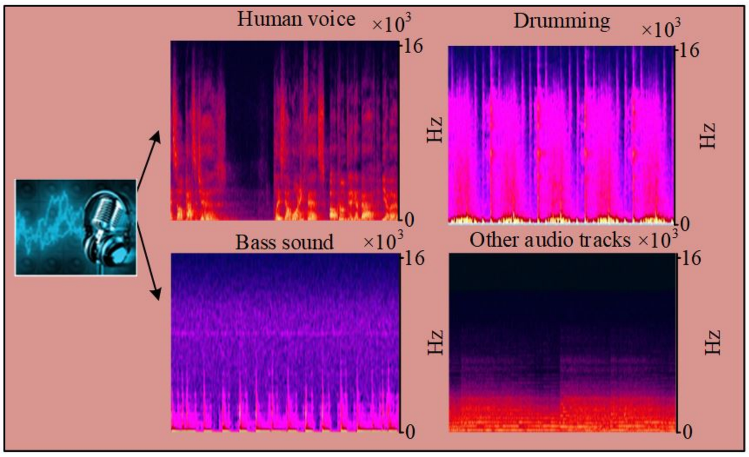


Fig. 3.1: Spectrum diagram of human sound source and some instrument sound sources in music signals

due to their different sound production mechanisms. Due to certain differences in the spectrum diagrams of different sound sources, the spectrum diagrams of human sound sources and some instrument sound sources in music signals are shown in Figure 3.1.

The audio in Figure 3.1 is the 10 second audio of the first track in the training set from the music track splitting dataset (MUSDB18). Human voice audio has a high frequency spectrum and a wide range of energy, while the frequency spectrum of bass sound is mostly low frequency, while the high energy range of other tracks is mostly below 2 kHz. In addition, according to the differences in the collection environment, the audio source mixing model can be divided into linear and convolutional models. In MSS, the mixing signal of the linear mixing model can be represented as a linear function of each channel source, as shown in equation (3.1).

$$p_i\left(\tau\right) = \sum_{k=1}^{W} b_{ik} z_k\left(\tau\right) \tag{3.1}$$

In equation (3.1), $p_i\left(\tau\right)$ represents a mixed signal. τ is the time. b_{ik} represents the actual transmission channel from the k-th source signal to the i-th sensor in an ideal environment. W represents the actual quantity of the original model. $z_k\left(\tau\right)$ represents the source signal. However, in practical environments, sources may experience delays, reflections, and other phenomena during transmission, making the mixing form of multiple sources more suitable for convolutional mixing models as it is not simply instantaneous mixing. The signal received by the j-th microphone in this model is expressed as shown in equation (3.2).

$$p_j\left(\tau'\right) = \sum_{k=1}^{W} g_{jk}\left(\tau'\right) * z_k\left(\tau'\right) = \sum_{k=1}^{W} \sum_{\tau'=0}^{O} g_{jk}\left(\gamma\right) z_k\left(\tau' - \gamma\right) \tag{3.2}$$

In equation (3.2), g_{jk} represents the actual transmission channel from the k-th source signal to the j-th microphone. τ' represents discrete time. γ represents the time delay of the source's transmission in the actual environment. O represents the order of the transfer function. $*$ represents the convolution operator. Of course, whether it is a linear or convolutional hybrid model, it can be simplified using a matrix, as expressed in equation (3.3).

$$q\left(\tau'\right) = Mp\left(\tau'\right) \tag{3.3}$$

In equation (3.3), $q\left(\tau'\right)$ represents the estimated signal. M represents the separation matrix to be solved. It is very difficult to restore the original signal from the mixed signal when both the source and channel parameters

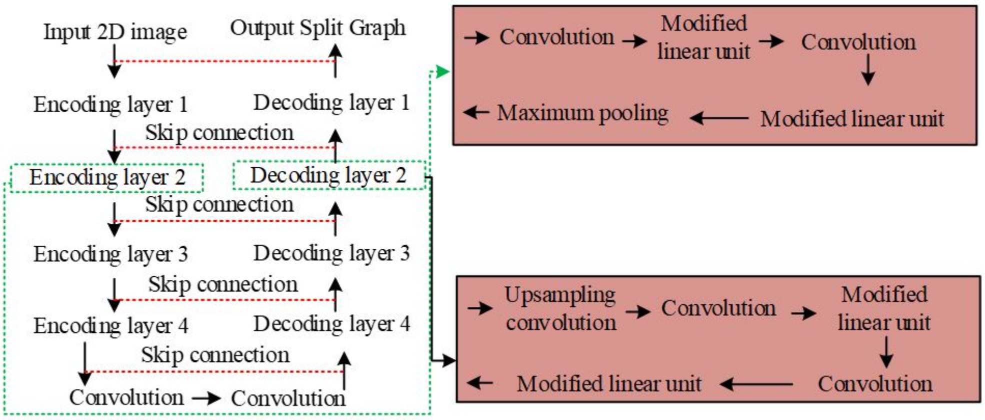


Fig. 3.2: Schematic diagram of U-net network structure

are uncertain. Usually, the original signal is restored by finding the inverse matrix of the mixed matrix. Due to the fact that the proposed MSS method belongs to the deep neural network class model, it is necessary to introduce the basic network algorithms it constitutes. Due to space limitations, only U-net will be analyzed here. The structure diagram of U-net is Figure 3.2.

In Figure 3.2, the U-net network consists of four levels of encoding layers and a decoder. Each layer sequentially includes convolutional layer, maximum buffer layer, exchange buffer layer, and skip connection. The U-net encoding layer requires 4 downsamples. At the same time, in response to the symmetry of U-net, the decoding layer also adopts 4 upsampling to restore the low-level high-level semantic image to the original image. Compared with traditional methods such as fully convolutional networks, U-net only performs 4 data upsampling and adopts a "skip" connection method, avoiding supervision and loss of high-level semantic features. Therefore, this method can fuse multiple low rank features to achieve multi-scale and deep supervision. To address the issue of sound source separation, the use of U-net's U-shaped structure and skip connection technology can achieve multi-scale characterization of four types of signal sources and obtain more accurate waveforms.

3.2. MSS method combined with Unet-SE-BiSRU. After analyzing the characteristics of mixed music audio sources, this study began to propose research methods for MSS. Unet-SE-BiSRU is based on the U-net network structure, combining SE module and SRU. The SE module is a component used in convolutional neural networks to enhance feature representation by explicitly modeling inter channel dependencies. The main advantage of the SE module is its ability to adaptively recalibrate feature maps, thereby improving model performance. The SE module focuses on learning inter channel dependencies in feature maps. By clearly modeling these dependencies, the network can dynamically adjust the importance of different channels at each spatial location, and its architecture is Figure 3.3.

In Figure 3.3, Unet-SE-BiSRU is mainly composed of three parts: encoder, loop network layer, and decoder. Unet-SE-BiSRU's codec also uses the skip method. Although this process is determined by experience, skip connections allow for direct access to the source signal, allowing the phase of the input signal to be transmitted directly to the output. On this basis, this study considers music audio as a time series, and therefore chooses SRU in a recurrent neural network [18], as shown in Figure 3.4.

In Figure 3.4, compared to traditional Long Short-Term Memory (LSTM), the SRU unit mainly utilizes the cell state of the previous moment to make temporal connections. Therefore, it is possible to perform gate state calculations based solely on the input of the current time, which has higher computational efficiency. The forgetting gate, reset gate, and intermediate output state expressions in Figure 3.4 are shown in equation (3.4).

$$e_t = \rho\left(C_e s_t + a_e\right) \tag{3.4}$$

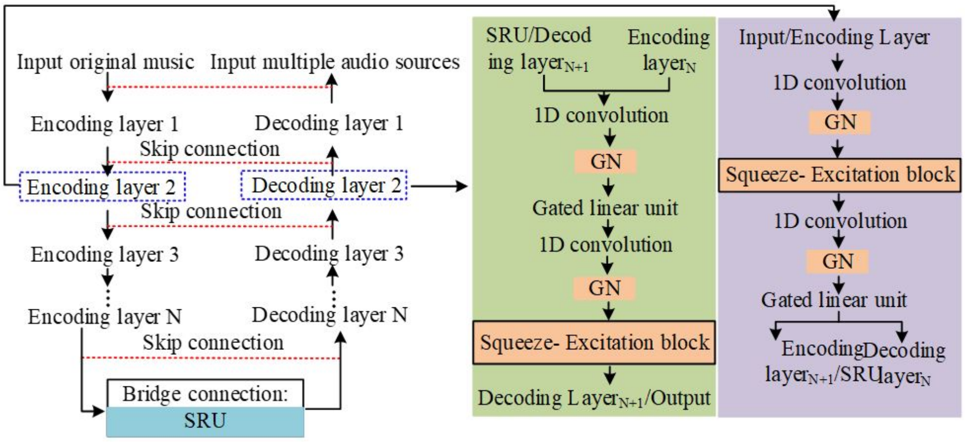


Fig. 3.3: Schematic diagram of Unet-SE-BiSRU network structure

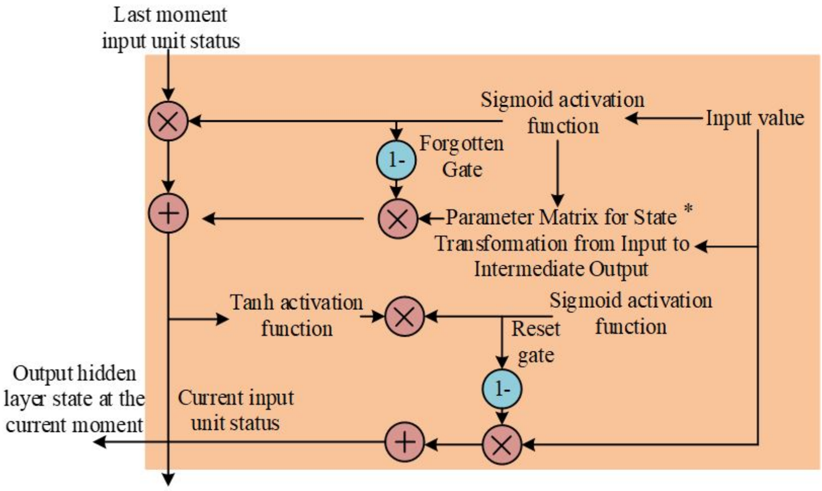

Last moment input unit status
Sigmoid activation function
Input value
1-
Forgotten Gate
Parameter Matrix for State Transfomation from Input to Intermediate Output
*
Tanh activation function
Sigmoid activation function
Reset gate
1-
Output hidden layer state at the current moment
Current input unit status

Fig. 3.4: Schematic diagram of SRU unit structure

In equation (3.4), e_t represents the forgetting gate. ρ represents the sigmoid activation function. C_e represents the parameter matrix of the transition from the forgetting gate to the intermediate output state. s_t represents the input value. a_e represents the bias value of the forgetting gate.

$$h_t = \rho\left(C_h s_t + a_h\right) \tag{3.5}$$

In equation (3.5), h_t represents the reset gate. C_h represents the parameter matrix for the transition from reset gate to intermediate output state. a_h represents the offset value of the reset gate.

$$\tilde{b}_t = C s_t \tag{3.6}$$

In equation (3.6), $\tilde{b}_t$ represents the intermediate output state. C represents the parameter matrix of the input to intermediate output state transformation. For a sequence composed of all input values s_t, batch

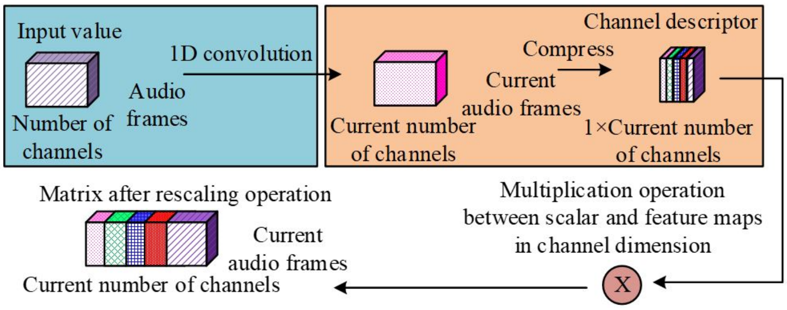

Input value
1D convolution
Audio frames
Number of channels
Channel descriptor
Compress
Current audio frames
Current number of channels
1×Current number of channels
Matrix after rescaling operation
Multiplication operation between scalar and feature maps in channel dimension
Current audio frames
Current number of channels
X

Fig. 3.5: Schematic diagram of 1D-SE basic architecture

allocation of matrix multipliers to various time steps can significantly improve computational efficiency, and the expression of batch multiplication is shown in equation (3.7).

$$U^T = \begin{bmatrix} C_e^T & C_h^T & C^T \end{bmatrix} \begin{bmatrix} s_1 & s_2 & \dots & s_t \end{bmatrix} \tag{3.7}$$

In equation (3.7), U^T represents batch processing of multiplication values. When the input vectors are equal, SRU and LSTM networks of the same size have smaller SRU weights and can reduce the computational burden during training. In addition, Unet-SE-BiSRU adds an SE module after convolutional operation, which adaptively weights each channel through attention mechanism to improve its expression ability. Because the SE module is used to process 2D images, it needs to be improved on Unet-SE-BiSRU to adapt to one-dimensional audio signals. The two main operations of 1D-SE after one-dimensional optimization are squeezing and excitation, and the basic architecture of 1D-SE is Figure 3.5.

In Figure 3.5, the first step is to obtain the corresponding feature through one-dimensional convolution and compress it into a channel descriptor to transmit the feature. The expression of the d-th element of the channel descriptor is equation (3.8).

$$Q_d = D_{sq}\left(v_d\right) = \frac{1}{N}\sum_{c=1}^{N} v_d\left(c\right) \tag{3.8}$$

In equation (3.8), Q_d represents the value of the d-th element. D_{sq} represents a compression operation. v_d represents the characteristic of the d-th element. c represents the number of audio frames, with a maximum value of N. Secondly, adaptive recalibration is carried out, and finally, the network expression ability is improved by explicitly modeling the correlation between convolutional feature channels, as shown in equation (3.9).

$$\Im = D_{ex}\left(Q, A\right) = \psi\left(r\left(Q, A\right)\right) = \psi\left(A_2\left(A_1 Q\right)\right) \tag{3.9}$$

In equation (3.9), $\Im$ represents the expression of interdependence. D_{ex} represents the compression operation from the channel descriptor module to the interdependence expression module. ψ represents a modified linear unit function. A, A_1 and A_2 represent parameters that consider model complexity and generalization. The final output expression of this module is shown in equation (3.10).

$$\tilde{X}_d = D_{scale}\left(v_d, \phi_d\right) = \phi_d v_d \tag{3.10}$$

In equation (3.10), $\tilde{X}_d$ represents the matrix after the rescaling operation. $D_{scale}\left(v_d, \Im_d\right)$ represents the multiplication operation between scalar ϕ_d and feature maps in the channel dimension. The 1D-SE module performs channel recalibration on one-dimensional convolution, which has the characteristic of lightweight and requires less model complexity and computational complexity [19]. Finally, in the group normalization module of Unet-SE-BiSRU, this study chose Group Normalization (GN) as the improvement scheme for Batch

Normalization (BN). It groups channels without using batch dimension data, so the calculation does not depend on the number of batches. The calculation expression of GN is equation (3.11).

$$\tilde{b}_\alpha = \frac{1}{\varsigma_\alpha}\left(b_\alpha - \eta_\alpha\right) \tag{3.11}$$

In equation (3.11), ϕ_d and $\tilde{b}_\alpha$ represent channel data before and after GN. ς represents the mean of the input data. ϕ_d represents the standard deviation. α represents the 3D vector after the corresponding feature index. The expression of standard deviation and input data mean is shown in equations (3.12) and (3.13).

$$\eta_\alpha = \frac{1}{n}\sum_{l \in J_\alpha} b_l \tag{3.12}$$

In equation (3.12), J_α represents the set of average audio frame values and standard deviations. n represents the actual size of the set.

$$\varsigma_\alpha = \sqrt{\frac{1}{n}\sum_{l \in J_\alpha}\left(b_l - \eta_\alpha\right)^2 + \theta} \tag{3.13}$$

In equation (3.13), θ represents a small constant. The expression of J_α calculation is equation (3.14).

$$J_\alpha = \left\{ l \middle| \, l_v = \alpha_v, \left\lfloor \frac{l_\aleph}{\aleph/Y} \right\rfloor = \left\lfloor \frac{\alpha_\aleph}{\aleph/Y} \right\rfloor \right\} \tag{3.14}$$

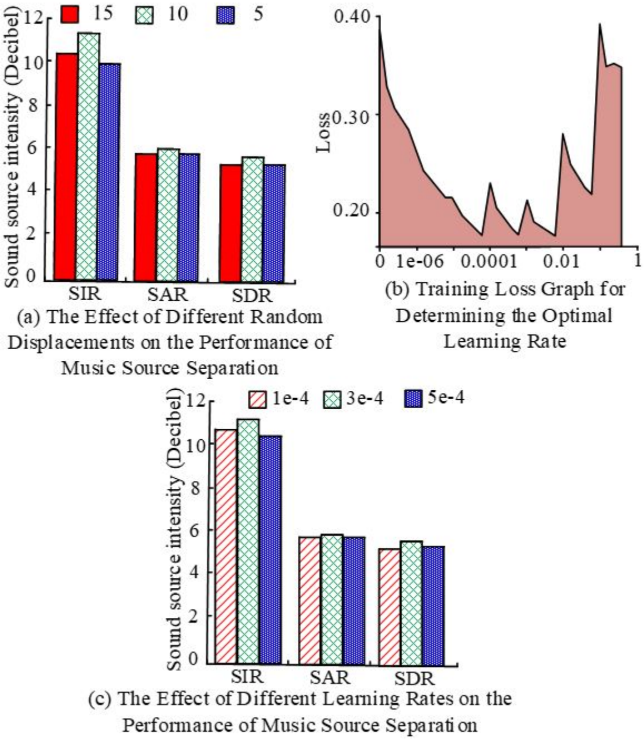
In equation (3.14), l_v and α_v represent sub-indices along the channel axis. v represents the index number. $\aleph$ represents the channel axis. $\aleph$ represents the number of groups, set to 32 based on actual needs. The final output formula is expressed as shown in equation (3.15).

$$f_\alpha = \mu\tilde{b}_\beta + \varepsilon \tag{3.15}$$

In equation (3.15), f_α represents the final output value. μ represents the scaling factor that can be learned. ε represents the offset.

4. Experimental Analysis of MSS Method Based on Unet-SE-BiSRU. Conducting simulation experiments on algorithms is a way to verify their performance. Therefore, this section mainly tests and evaluates Unet-SE-BiSRU, including parameter settings, analysis of different modules, and comparative experiments with other MSS methods.

4.1. Experimental parameter setting and analysis of different modules. To verify the MSS effectiveness of Unet-SE-BiSRU, this study conducted experimental and comparative validation. The open source dataset MUSDB18 used in Figure 1 was selected for the experiment, which is a commonly used dataset for music source separation research. This dataset contains a mixed version of music tracks, each containing multiple audio tracks such as vocals, guitars, drums, etc. It contains 150 pieces of music and is divided into training and testing sets in a 2:1 manner. All music audio data is dual channel, with a sampling rate of 44.1 Khz. All music audio data is dual channel, with a sampling rate of 44.1 Khz. Prior to this, to effectively enhance the generalization ability of the actual separation network, the original data was expanded in the experiment. The first step is to randomly switch between each sound source. The second is to stack and scale the amplitude of the sound source. Then, each sound source is randomly grouped and formed into a series, and then randomly mixed from different tracks. Finally, multiply each source waveform by ± 1. The experimental environment includes the Ubuntu 16.04 operating system, the Intel Core i9-10900 central processor, three 12G NVIDIA produced RTX2080Ti graphics cards, and a Python deep learning framework. The batch processing size is set to 12. The evaluation criteria selected in the experiment are Signal to Distortion Ratios (SDR), Source to Interference Ratio (SIR), and Sources to Noise Ratio (SAR). In audio enhancement tasks, SDR is more important because it directly reflects the quality of the reconstructed signal; In the task of environmental noise



(a) The Effect of Different Random Displacements on the Performance of Music Source Separation

(b) Training Loss Graph for Determining the Optimal Learning Rate

(c) The Effect of Different Learning Rates on the Performance of Music Source Separati on

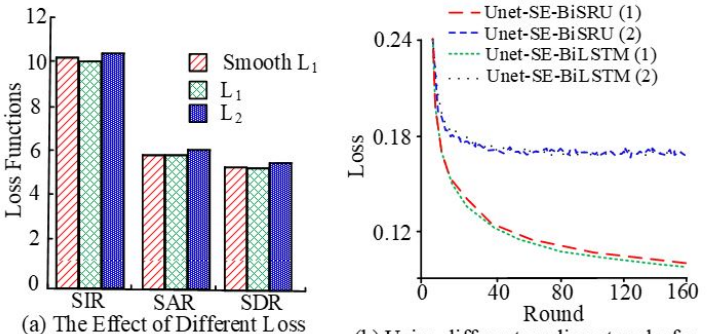
Fig. 4.1: Predicting Random Offset Values and Optimal Learning Rate Results

suppression, SIR and SAR are more crucial. In the optimal parameter setting, analysis was conducted on the predicted random offset value and the optimal learning rate, and the results are shown in Figure 4.1.

In Figure 4.1, when the predicted random offset value was 5, the values of indicators A, B, and C were 5.20 decibels, 5.79 decibels, and 9.90 decibels, and when it was 15, the values of the three indicators were 5.29 decibels, 5.79 decibels, and 10.38 decibels, respectively. At 10:00, the values of the three indicators were 5.68 decibels, 5.91 decibels, and 11.28 decibels, which were higher than the results of the other two predicted random offset values. In addition, the learning rate analysis showed a corresponding increase in learning rate every 10 rounds, resulting in a total of 70 training rounds. The results showed that at the beginning of training, the loss decreases with increasing rounds. As the learning speed increased, the loss began to increase, which meat that the accuracy learning rate of 1e-4 was sufficient to make it jump out of local optima. Therefore, when using 1e-4 as the base learning rate to validate the analysis, the SDR value at 3e-4 was 5.68, which was higher than the other learning rates. Based on this, in the experiment, the predicted random offset value was set to 10, and the learning rate was 3e-4.

In the experimental verification of Unet-SE-BiSRU, the effectiveness of its loss function and BiSRU was first verified. The L_1 norm loss function (L_1), L_2, and smooth L_1 norm loss function (Smooth L_1) was selected for comparison, and the results are shown in Figure 4.2.

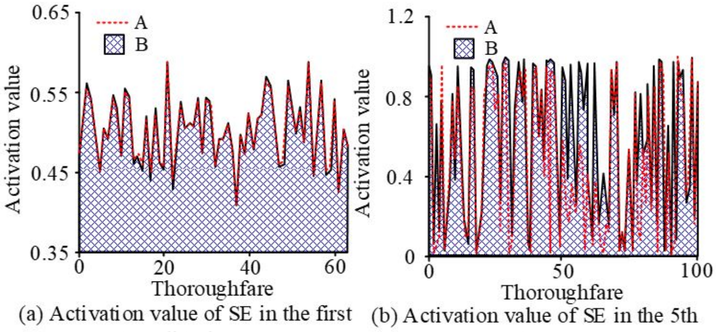
In Figure 4.2 (b), 1 and 2 represent training loss and validation loss, respectively. The SDR value of L_1 is 5.68 decibels, which is higher than the comparison loss function. Therefore, L_1 was selected as the optimization target in the Unet-SE-BiSRU network. Overall, the actual training speed of BiSRU was significantly higher than that of BiLSTM. The former required a total of 18 minutes per training round, while the latter only required 12 minutes. And the actual parameter quantity of BiSRU was 3.33×10^4, only half of BiLSTM. In the comparison of SDR indicators, the BiSRU value was 5.68 decibels and the BiLSTM value was 5.55 decibels. Therefore, SRU units had higher performance and application effectiveness. This study analyzed the effectiveness of attention



(a) The Effect of Different Loss Functions on the Performance of Music Source Separation

(b) Using different cyclic networks for each round loss value

Fig. 4.2: The Influence of Different Loss Functions on the Performance of Music Source Separation and the Effectiveness Analysis of BiSRU



(a) Activation value of SE in the first encoding layer

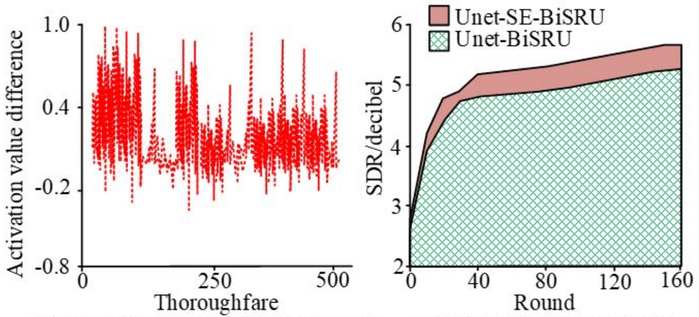
(b) Activation value of SE in the 5th encoding layer

Fig. 4.3: Activation value results of SE on the first and fifth encoding layers of any two music in the test set

mechanism SE blocks and group normalization. Among them, by assigning SE activation values at each coding level, the effect of SE feature weighting could be tested. Therefore, taking any two pieces of music in the test set (represented by A and B) as examples, the activation values of SE on the first and fifth encoding layers are shown in Figure 4.3.

In Figure 4.3, in the first encoding layer of these two pieces of music, the actual activation values of SE were mainly distributed between 0.4 and 0.6. The activation values of the two songs had high similarity on different channels, indicating that during MSS, the shallow layers in the network were likely to extract common characteristics of sound. As the depth of the network increased, the characteristics of network retrieval gradually focused on differences in performance sources. In addition, the gradual deepening of network depth had led to a gradual differentiation of the activation value distribution of SE, and its practical effect had become increasingly apparent. To effectively compare the differences in activation values between two pieces of music in the fifth encoding layer and analyze the effect of adding an SE module, the results of subtracting the activation values of each channel and comparing them with the Unet-BiSRU network without an SE module were analyzed. The results are shown in Figure 4.4.

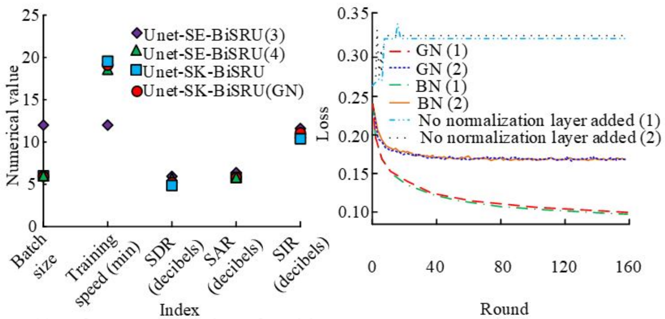
In Figure 4.4, the SE module selectively enhanced the effective feature weights and suppresses useless features through adaptive recalibration of multiple audio tracks, thereby enhancing the network's representation ability and improving the actual separation effect. After adding the SE module, the SDR values during the experimental rounds were significantly higher than those of the Unet-BiSRU network without adding the SE



(a) The difference in activation values between two pieces of music in SE on the 5th encoding layer

(b) SDR indicator results for Unet-BiSRU and Unet-SE-BiSRU

Fig. 4.4: The result of subtracting the activation values of each channel and comparing it with the Unet-BiSRU network without the SE module



(a) Performance Comparison of Models Using Different Attention Blocks

(b) Loss values for each round using different normalization strategies

Fig. 4.5: Comparison of different typical attention mechanism modules and GN effectiveness results

module, with an overall increase of about 0.1-0.5 decibels. This indicated the effectiveness of the attention mechanism SE module. On this basis, the effectiveness of different typical attention mechanism modules and GN modules was analyzed: the introduction of Selective Kernel (SK) attention mechanism was compared. Figure 10 shows the results.

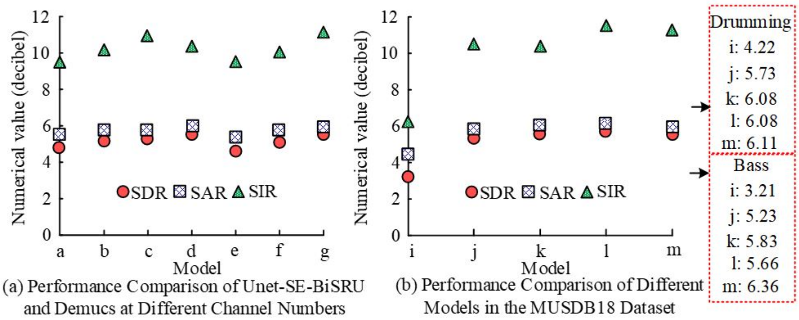
In Figure 4.5 (a), 3 and 4 represent Unet SE BiSRUs with batch sizes 12 and 6. The maximum batch processing sizes that Unet-SE-BiSRU and Unet-SK-BiSRU could support were 12 and 6, respectively, with SDR values of 5.68 decibels and 5.21 decibels for both. Overall, the SE module was more suitable for studying models under the same limitations of memory size. Meanwhile, the final SDR value obtained by GN was 5.68 decibels, significantly higher than BN's 5.34 decibels. Overall, the SE module in the proposed Unet-SE-BiSRU network had higher performance compared to GN, thereby proving the robustness of the research network.

4.2. Contribution of different modules and comparative experimental analysis. As shown in Table 4.1, based on the verification of the performance of each part in the Unet-SE-BiSRU network, this study began to analyze the contribution of innovation in each part to separation performance.

In Table 4.1, 1∼6 represent the Unet-SE-BiSRU, Unet-BiSRU, Unet-SE-BiLSTM, Unet-SE-BiGRU, Unet-SE-BiSRU (BN) and Unet-SE-BiSRU (Unnormalized), respectively. Among the three optimizations, SE had the greatest impact on the separation effect SDR, followed by GN, and finally SRU. With the removal of SE, Unet-SE-BiSRU saved time, but the SDR value decreased by 0.43 decibels. After replacing LSTM or GRU with

Table 4.1: Contribution of innovation in various parts to actual separation performance

-	1	2	3	4	5	6
Differences from Unet-SE-BiSRU	-	Remove SE	Replace SRU with LSTM	Replace SRU with Gate Recurrent Unit	Replace GN with BN	Remove GN
Training speed	12.0 min	11.5 min	18.0 min	16.7 min	12.0 min	12.4 min
SDR (decibel)	5.68	5.24	5.54	5.36	5.33	0.91
SAR (decibel)	5.91	5.86	5.88	5.76	6.06	5.18
SIR (decibel)	11.28	10.46	11.17	11.03	10.17	-2.23



(a) Performance Comparison of Unet-SE-BiSRU and Demucs at Different Channel Numbers

(b) Performance Comparison of Different Models in the MUSDB18 Dataset

Fig. 4.6: Comparison results between Unet-SE-BiSRU and other models

bridge connected SRU, its performance decreased by 0.13 decibels. Overall, the three improvements effectively improved the actual separation performance of the network while not increasing the time complexity. Finally, the study compared it with Wave-U-Net, the Audio Separation Model using Convolutional Neural Networks (Open-Unmix), the Reference Model Demucs, and the Fully Convolutional Time Domain Audio Separation Network (Conv Tasnet). Each model is represented by i, j, k, l in sequence, and the research network model is m. The results are shown in Figure 4.6.

In Figure 4.6 (a), a∼d represent Demucs models with channel numbers of 16, 32, 64, and 100. The training speeds of the first three are 8min, 11min, and 20min respectively, with 240 iteration rounds. e∼g represents Unet-SE-BiSRU with channel numbers of 16, 32, and 64, with 160 iteration rounds. Figure 4.6 shows that the Unet-SE-BiSRU model with the same number of channels has significantly better SDR values than the Demucs model, indicating its effectiveness in optimization. Compared with other models, the SDR value of Unet-SE-BiSRU has currently shown better comprehensive separation performance. The comparison between the Wave U Net, Deep Clustering, and Deep attractor network algorithm models introduced in the study and the algorithm models proposed in this study is shown in Table 4.2 [20-22].

According to Table 4.2, the ACC, RMSE, AOC, F1, mAP, and Loss of the proposed Unet SE BiSRU algorithm model are 0.92, 0.03, 0.98, 0.97, 0.92, and 0.064, respectively. The experimental results show that the proposed Unet SE BiSRU algorithm model has excellent performance among the four algorithm models.

5. Conclusion. In response to the limited complexity and insufficient expression ability of traditional music audio signal processing methods, as well as the long training time and poor separation effect of end-to-end neural network models using time-domain deep learning, this study proposes the Unet-SE BiSRU model based on Demucs and analyzes its comprehensive separation performance. The experimental results show that when predicting a random offset value of 10, the values of the three indicators are 5.68, 5.91, and 11.28, respectively. When learning rate is 3e-4, the values of the three indicators are 5.68, 5.91, and 11.28, respectively. Therefore, the two parameters are set to 10 and 3e-4. In addition, the training speed of BiSRU is significantly higher than that of BiLSTM. The former requires a total of 18 minutes for each round of training, while the latter only

Table 4.2: Performance Comparison of Various Algorithms

Model	ACC	RMSE	AOC	F1	mAP	Loss
Wave-U-Net	0.83	0.21	0.86	0.82	0.83	0.097
Deep Clustering	0.87	0.11	0.92	0.87	0.84	0.082
Deep attractor network	0.90	0.06	0.96	0.91	0.88	0.076
Unet-SE-BiSRU	0.92	0.03	0.98	0.97	0.92	0.064

requires 12 minutes. The gradual deepening of network depth causes the distribution of SE activation values to gradually differentiate. At the same time, in model comparison, the Unet SE BiSRU model with the same number of channels has a significantly better SDR value than the Demucs model, with an overall increase of about 0.2 decibels. Compared with the current better audio separation model, it also shows good performance. Overall, the Unet SE BiSRU model is effective in music source separation for music signal processing. However, the introduction of the extrusion excitation module can improve the end-to-end network expression ability in the time domain, but there is still significant room for improvement, and the richness of the dataset used is also insufficient. In the future, it is necessary to enhance the performance of speech feature extraction and use a larger dataset to achieve better sound source separation.

REFERENCES

[1] Gong, Y., Dai, L. & Tang, J. A selection function for pitched instrument source separation. *Multimedia Systems.* **28**, 311-319 (2022)

[2] Gauer, J., Nagathil, A., Eckel, K., Belomestny, D. & Martin, R. A versatile deep-neural-network-based music preprocessing and remixing scheme for cochlear implant listeners. *The Journal Of The Acoustical Society Of America.* **151**, 2975-2986 (2022)

[3] Rixen, J. & Renz, M. Sfsrnet: Super-resolution for single-channel audio source separation. *Proceedings Of The AAAI Conference On Artificial Intelligence.* **36**, 11220-11228 (2022)

[4] Chen, Y., Hu, Y., He, L. & Huang, H. Multi-stage music separation network with dual-branch attention and hybrid convolution. *Journal Of Intelligent Information Systems.* **59**, 635-656 (2022)

[5] Liu, S., Keren, G., Parada-Cabaleiro, E. & Schuller, B. N-HANS: A neural network-based toolkit for in-the-wild audio enhancement. *Multimedia Tools And Applications.* **80**, 28365-28389 (2021)

[6] Tzinis, E., Wang, Z., Jiang, X. & Smaragdis, P. Compute and memory efficient universal sound source separation. *Journal Of Signal Processing Systems.* **94**, 245-259 (2022)

[7] Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T. & Dubnov, S. Zero-shot audio source separation through query-based learning from weakly-labeled data. *Proceedings Of The AAAI Conference On Artificial Intelligence.* **36**, 4441-4449 (2022)

[8] Huang, J., Wang, J., Smith, J., Song, X. & Wang, Y. Modeling the compatibility of stem tracks to generate music mashups. *Proceedings Of The AAAI Conference On Artificial Intelligence.* **35**, 187-195 (2021)

[9] Ma, B., Zhang, T., An, Z. & Yi, C. Measuring dependence for permutation alignment in convolutive blind source separation. *IEEE Transactions On Circuits And Systems II: Express Briefs.* **69**, 1982-1986 (2021)

[10] Zhou, D., Zhou, X., Hu, D., Zhou, H., Bai, L., Liu, Z. & Ouyang, W. Sepfusion: Finding optimal fusion structures for visual sound separation. *Proceedings Of The AAAI Conference On Artificial Intelligence.* **36**, 3544-3552 (2022)

[11] Colonel, J. & Reiss, J. Reverse engineering of a recording mix with differentiable digital signal processing. *The Journal Of The Acoustical Society Of America.* **150**, 608-619 (2021)

[12] Gupta, R., He, J., Ranjan, R., Gan, W., Klein, F., Schneiderwind, C. & Välimäki, V. Augmented/mixed reality audio for hearables: sensing, control, and rendering. *IEEE Signal Processing Magazine.* **39**, 63-89 (2022)

[13] Srinivasamurthy, A., Gulati, S., Repetto, R. & Serra, X. Saraga: Open Datasets for Research on Indian Art Music. *Empirical Musicology Review.* **16**, 85-98 (2021)

[14] Sheeja, J. & Sankaragomathi, B. CNN-QTLBO: an optimal blind source separation and blind dereverberation scheme using lightweight CNN-QTLBO and PCDP-LDA for speech mixtures. *Signal, Image And Video Processing.* **16**, 1323-1331 (2022)

[15] Patel, P., Ray, A., Thakkar, K., Sheth, K. & Mankad, S. Karaoke Generation from songs: recent trends and opportunities. *Asia-Pacific Signal And Information Processing Association Annual Summit And Conference (APSIPA ASC).* pp. 1238-1246 (2022)

[16] Zhang, T., Zhang, Y., Sun, H. & Shan, H. Parkinson disease detection using energy direction features based on EMD from voice signal. *Biocybernetics And Biomedical Engineering.* **41**, 127-141 (2021)

[17] Luo, H., Du, J., Yang, P., Shi, Y., Liu, Z., Yang, D. & Wang, Z. Human–Machine Interaction via Dual Modes of Voice and Gesture Enabled by Triboelectric Nanogenerator and Machine Learning. *ACS Applied Materials & Interfaces.* **15**, 17009-17018 (2023)

[18] Guo, Y., Mustafaoglu, Z. & Koundal, D. Spam detection using bidirectional transformers and machine learning classifier algorithms. *Journal Of Computational And Cognitive Engineering.* **2**, 5-9 (2023)

[19] Tang, T., Li, Z., Cheng, Y., Xu, K., Xie, H., Wang, X. & Ou, J. Single-step growth of p-type 1D Se/2D GeSe x O y heterostructures for optoelectronic NO2 gas sensing at room temperature. *Journal Of Materials Chemistry A.* **11**, 6361-6374 (2023)

[20] Hu, W., Zhang, H., Sang, W., Anna, S. & Yuan, S. Surface-wave dispersion curves extraction method from ambient noise based on U-net++ and density clustering algorithm. *Journal Of Applied Geophysics.* **213**, 248-257 (2023)

[21] Yang, Z., Ren, Y., Wu, Z., Zeng, M., Xu, J. & Yang, Y. DC-FUDA: Improving deep clustering via fully unsupervised domain adaptation. *Neurocomputing.* **526**, 109-120 (2023)

[22] Reza, S., Seyyedsalehi, S. & Seyyedsalehi, S. Modified deep attractor neural networks for variability compensation in recognition tasks. *Computers And Electrical Engineering.* **99**, 1077-1092 (2022)

Edited by: Zhengyi Chai
Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City
Received: Nov 23, 2023
Accepted: May 6, 2024