Hindawi
Complexity
Volume 2021, Article ID 7442877, 11 pages
https://doi.org/10.1155/2021/7442877

Research Article

Research on Segmentation Experience of Music Signal Improved Based on Maximization of Negative Entropy

Qin Yao,[1] **Zhencong Li**,[1] **and Wanzhi Ma**[2]

[1]*School of Music and Dance, Ningxia Normal University, Guyuan, Ningxia 756000, China*
[2]*Department of Educational and Culture Contents Development, Woosuk University, Jeonju 55338, Republic of Korea*

Correspondence should be addressed to Wanzhi Ma; 997443418@stu.woosuk.ac.kr

Received 23 April 2021; Revised 8 May 2021; Accepted 14 May 2021; Published 25 May 2021

Academic Editor: Zhihan Lv

Copyright © 2021 Qin Yao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the rapid growth of digital music today, due to the complexity of the music itself, the ambiguity of the definition of music category, and the limited understanding of the characteristics of human auditory perception, the research on topics related to automatic segmentation of music is still in its infancy, while automatic music is still in its infancy. Segmentation is a prerequisite for fast and effective retrieval of music resources, and its potential application needs are huge. Therefore, topics related to automatic music segmentation have important research value. This paper studies an improved algorithm based on negative entropy maximization for well-posed speech and music separation. Aiming at the problem that the separation performance of the negative entropy maximization method depends on the selection of the initial matrix, the Newton downhill method is used instead of the Newton iteration method as the optimization algorithm to find the optimal matrix. By changing the descending factor, the objective function shows a downward trend, and the dependence of the algorithm on the initial value is reduced. The simulation experimental results show that the algorithm can separate the source signal well under different initial values. The average iteration time of the improved algorithm is reduced by 26.2%, the number of iterations is reduced by 69.4%, and the iteration time and the number of iterations are both small. Fluctuations within the range better solve the problem of sensitivity to the initial value. Experiments have proved that the new objective function can significantly improve the separation performance of neural networks. Compared with the existing music separation methods, the method in this paper shows excellent performance in both accompaniment and singing in separated music.

1. Introduction

Music is the most common form of artistic expression in daily life, which greatly meets people's spiritual and cultural needs and enriches people's leisure life. People relax and enrich their lives by enjoying music. With the development of digital music, the threshold of music creation is getting lower and lower. As a kind of audio signal, music signal is widely spread through the convenient Internet. With copyright permission, people can download all kinds of music on the Internet. Therefore, the amount of music audio data is getting larger and larger, and the requirements for retrieval tasks are getting higher and higher [1, 2]. However, many mainstream music search engines are still based on simple text retrieval, that is, manually labeled song names, singers, years, and so on. If retrieval can be performed based on the content information of the music signal itself (such as melody, rhythm, harmony, timbre, intensity, speed, mode, and musical style) and these features can be automatically identified, this has meaning for retrieval efficiency and user experience major [3–5].

The key technology of automatic music segmentation has important research value. The index structure established based on the results of the automatic segmentation will further improve the performance of the music retrieval system [6]. In addition, the automatic music segmentation system also helps to establish an objective theoretical system of music analysis in addition to the subjective way of human perception and intuition and reduces human prejudices and prejudices [7]. The music style segmentation system can be

used to identify the works of a specific composer by training the segmenter, to help determine the copyright of unknown musical works, and to determine the main characteristics that distinguish different genres. By comparing with the "objective" features obtained by the computer segmenter, the segmentation results will also support research on the concept of human music similarity in sociology and psychology and the process of music group formation. The segmenter can also automatically analyze and segment the records added to a large database. Based on the analysis and segmentation of music content, the music recommendation system can be used to find popular or high and low music works in a massive music database and recommend lesser-known works according to personal preferences [8]. This kind of personalized recommendation is expected to weaken the strong trend of popular music and better search for massive music resources. After training, the segmenter can segment personal music collections according to emotions and scenes and can automatically select suitable records in different situations such as driving, meeting customers, and cleaning. Similarity analysis can also be used to monitor the distribution of various types of records. Using the results of music segmentation, the automatic music transcription system can also identify different styles of sound effects as corresponding notes [9].

In this paper, an algorithm that combines negative entropy maximization and Newton's downhill method is adopted, and the downhill factor makes the objective function have a descending property. The simulation experimental results show that the algorithm can separate the speech signal and music signal well under different initial values. Observing the experimental results of 30 sets of random initial matrices, the average iteration time of the improved algorithm is reduced by 26.2% and the number of iterations is reduced. The iteration time and the number of iterations fluctuate within a small range, which better solves the problem of sensitivity to the initial value. Experiments have proved that the method in this paper can significantly improve the separation performance of neural networks. Compared with the existing music separation methods, the method in this paper has excellent performance in separating the accompaniment and singing voice in music. At the same time, the method in this paper has excellent performance in separating music. It is less affected by the separated signal and has strong universality and generalization performance.

2. Related Work

In classical theory, the short-time Fourier transform is used to analyze the signal, and the frequency amplitude is approximated by the coefficient of the harmonic function [10]. This usually does not adequately represent the music signal, because the music signal is not only a mixture of multiple instruments playing the same pitch (fundamental frequency), each instrument has a specific range of overtones (the collection of these overtones is called timbre), all musical instruments have a frequency distribution that is much more ambiguous than a single sine, for specific musical instruments or different players, the frequency distribution has certain fluctuations, and the singer's voice is often mixed, so the harmonic function is used [11]. To represent the signal, quite a lot of coefficients are needed. Using wavelet function, Gabor function and other time-frequency analysis methods can better describe each musical instrument or describe different aspects of music, namely, timbre. Because the time-frequency resolution of wavelet transform is adapted to each signal, the signal can be represented more effectively [12].

In order to compare music in order to effectively extract features, a specific representation method is required. Due to the differences in the capabilities of different wavelet transforms, the music signal representation method suitable for one feature extraction is not so sufficient when describing other features, so each feature needs to be represented differently. Sparse component transformation is a method that can fully describe a variety of features so far. For example, the DIRAC base can describe the random noise in the signal, the DCT can describe the frequency characteristics of the entire time interval, and the wavelet packet can be used to describe the short-term and long-term events of the signal, such as the phenomenon at the beginning of a note and the long-term events [13]. Through experiments and analysis, it is necessary to find a set of dictionary functions that can effectively represent different characteristics of music signals [14].

The segmenter uses the idea of template matching to create a template for each audio type, then calculates the feature vector of the actual audio frame, and uses the feature vector to match the template vector (usually calculating their distance in the vector space) to identify the audio type. In the music clustering system developed by scholars from the Australian Institute of Artificial Intelligence, the type judgment method of template matching is adopted, the matching is performed by calculating the Euler distance between the template vector and the feature vector, and the retrieval system ARS also uses a template-based audio retrieval algorithm [15–17].

Since the first application of auditory scene analysis to the separation of voice and music, the separation of voice and music has introduced methods such as fundamental frequency analysis, time-frequency analysis tools, and blind source separation [18]. Related scholars have simulated how the human auditory system can distinguish a sound from a mixed sound and determine which parts of the spectrum come from the same channel of information according to the endpoint information, frequency changes, and overtones of different frequency ranges and form the same signal based on these characteristics [19]. Researchers have proposed a system for separating piano accompaniment and singing, using the existing piano accompaniment score or overtone trajectory as prior knowledge and using a linear combination of sinusoids with time-varying frequency, amplitude, and phase to simulate piano accompaniment and singing [20]. The source signal can be obtained by obtaining the coefficients of these linear combinations. Related scholars use blind source separation algorithms to separate speech and music signals in the actual environment [21]. The limited

filter length and nonlinear sensor noise of the hybrid model in the theoretical algorithm make the algorithm limited in practical applications. Assuming that the number of source signals and the number of sensors are fixed, the frequency domain blind source separation algorithm without any prior knowledge is used to separate the signals, and the separated signals are divided into dominant ones according to the relative power.

Related scholars added short-term continuity and sparsity constraints to the nonnegative matrix factorization to achieve the separation of mixed music signals [22]. The basic idea of the paper is to decompose the amplitude spectrum of the input signal to obtain the sum of a series of vectors. On the contrary, when the decomposition vector is known, the source signal can be recovered by solving the coefficients. The square of the gain difference between adjacent frames is used as the cost function of short-term continuity, and the nonzero gain is used as the cost function of sparsity [23]. The parameters of each signal are obtained by minimizing the reconstruction error between the input spectrum and the model obtained by NMF training. Compared with independent component analysis and nonnegative matrix factorization methods, NMF with restricted conditions can get a better separation effect. Among them, short-term continuity is more effective in detecting high-pitched music signals. Related scholars have proposed a semiblind separation of speech and music based on sparsity and continuity; they used sparsity and continuity constraints to optimize dictionary coefficients, used the dictionary to represent the power spectral density of each source signal, and mixed them through a nonlinear function [24–32]. The power spectrum of the signal is mapped to the dictionary space, and finally, the source signal is reconstructed using an adaptive Wiener filter and spectral subtraction.

3. Music Feature Analysis and Musical Note Modeling

3.1. Analysis of Music Features. The tone has four characteristics of pitch, value, intensity, and timbre. These four characteristics correspond to the vibration frequency, duration, vibration amplitude, and frequency spectrum distribution of the musical instrument, respectively. Pitch is a perceptual attribute of sound. Pitch can be quantified as frequency, which depends on the speed at which sound waves vibrate the air, and has almost nothing to do with the strength or amplitude of the wave. In other words, a "high" tone means a very fast oscillation, and a "low" tone corresponds to a slower oscillation. Since the vibration of the sounding body is usually composed of a set of waveforms with different frequencies and different amplitudes, it is stipulated that the lowest vibration frequency in this group of compound vibrations is the fundamental tone, and the others are all overtones, where the fundamental tone determines the pitch. In the production of musical instruments, each key or string of the musical instrument corresponds to a different fundamental tone. Therefore, a reference tone must be drawn up first. On this basis, the remaining notes are calculated according to the temperament used. Temperament is the scientific basis for the quantitative characterization of musical notes. The schematic diagram of note time value cutting is shown in Figure 1.

The pronunciation time of the pronunciation body is related to the vibration it produces. The vibration stops and the pronunciation stops. In the field of music, the beat is used to describe the sound value. The beat does not have a fixed length, but it is closely related to the style of the music and the duration of the performance. Beat is the basic unit of rhythm. Any music has a rhythm. The notes of different pitch values are combined into bars, and then each bar is connected in series to form a rhythm. Because the rhythm of each music is unique, rhythm research is also very helpful for song identification.

Sound intensity is the subjective perception of sound pressure by the human ear. It is defined as a kind of auditory attribute; according to this attribute, the sound can be sorted from quiet to noisy. Sound intensity is also related to psychological factors, which means that loudness and amplitude are not exactly proportional to each other. In the field of music research, if the sound frequency of the musical instrument does not change, the strength of the sound of the musical instrument depends only on the amplitude of the musical instrument's own vibration.

The timbre belongs to the auditory sensory characteristics of the human ear and is mainly determined by the frequency spectrum of the sound. According to the American Standards Association's definition of timbre, the difference in sound quality other than pitch and intensity is called timbre. After analyzing the sound containing the same spectrum components, it can be known that the timbre is to a large extent related to the amplitude variation characteristics of the overtones in the compound vibration at the beginning and the end of the vibration. In addition, timbre can also distinguish different types of sound production and help the human ear distinguish different instruments in the same category such as oboe and clarinet.

3.2. Signal Preprocessing. After analyzing the four major characteristics of music and mastering the key acoustic characteristics of note modeling, the relevant parameters of the notes are extracted based on these acoustic characteristics. The discrete signal after sampling and quantization must be preprocessed before being used for data analysis.

3.2.1. Preemphasis. According to the string vibration equation, it can be seen that the standing wave generated by the string vibration is mixed with many high-frequency overtones, and its power spectrum decreases with the increase of frequency. This causes the signal to have a large low-frequency signal-to-noise ratio and a high-frequency signal-to-noise ratio. In addition, the signal exhibits low-pass filtering characteristics during transmission, which makes high-frequency transmission very difficult. In order to solve the problem of high-frequency transmission, it is necessary to emphasize the high-frequency signal

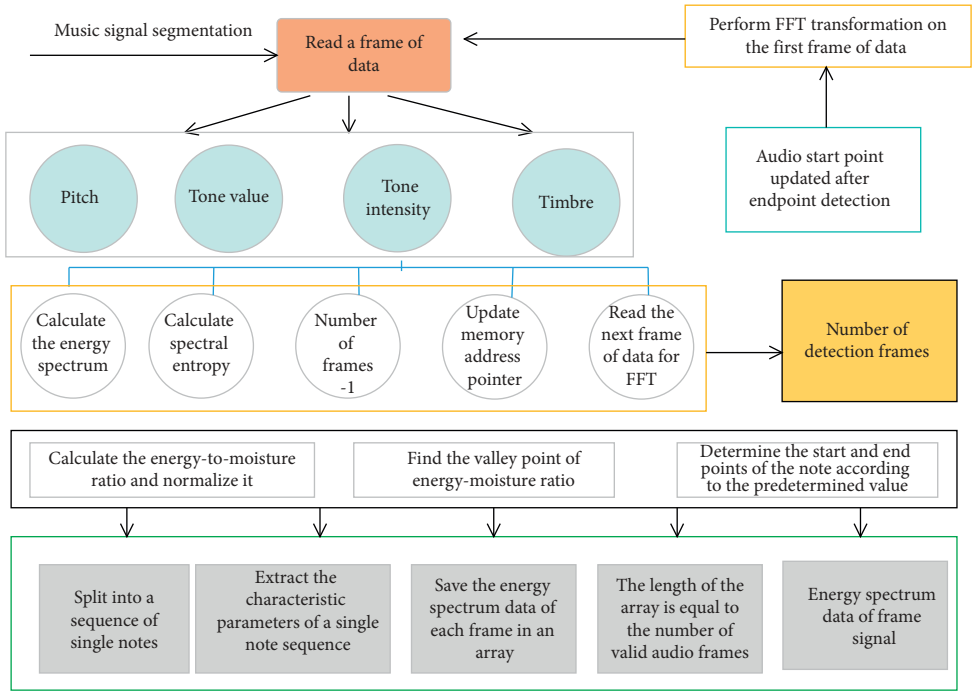

Figure 1: Schematic diagram of note time value cutting.

component to generate a modulation index that is more equal to the transmission spectrum, that is, to compensate the high-frequency component of the input signal. This processing method is preemphasis. The paper uses frequency domain technology for preemphasis, and the original signal is calibrated and filtered before subsequent processing. The transfer function of the preemphasis filter is as follows:

$$H(z) = 1 + \frac{\alpha}{z}. \tag{1}$$

In the formula, α is the preemphasis coefficient.

3.2.2. Windowing and Framing. In the field of signal analysis, according to the characteristics of inertia, it can be considered that the distribution of nonstationary signals in a relatively short period of time does not change with time, so the steady-state method can be used to analyze nonstationary signals. The audio signal is a typical nonstationary signal. Before analyzing and processing it, it first needs to be aligned for time-domain framing. The framing is realized by a movable window of limited length. In order to ensure the continuity of the voice, there must be a certain overlap between each frame of data when the window is moved, and the number of samples moved each time is the frame shift.

3.2.3. Endpoint Detection. Endpoint detection is to determine the starting point and ending point of a valid voice from the audio file. Only when the starting and ending points of the valid audio are found, the subsequent signal analysis is meaningful. The significance of signal endpoint detection is that it can reduce the amount of data processing for note recognition in the embedded system, which is mainly manifested in the following two aspects.

On one hand, it can reduce the amount of blank voice signal transmission inside the system and reduce the computing load of the processor. This is of great significance to the real-time recognition of signals; on the other hand, it can filter out noise signals that do not contain effective information. If the signals to be recognized are mixed with noise, it will not only cause waste of memory resources but also disrupt the recognition process to a certain extent.

3.3. Establishment of Mathematical Model of Musical Notes. The purpose of mathematical modeling is to find a corresponding relationship, under which the corresponding mathematical form of quantity and quantity can be realized, and the maximum matching accuracy between two physical quantities can be achieved through the corresponding relationship. Mathematical modeling of musical notes is to find the correspondence between note names and waveforms. Through the analysis of the four characteristics of music, the paper extracts the time-domain envelope and frequency spectrum parameters of the note signal and performs parameter fitting in Matlab according to the frequency domain parameters. The specific process is shown in Figure 2. Figure 3 is a model based on the attenuation law of the note envelope.

The envelope function of the note time-domain contains the characteristics of the note value and intensity; the analysis of the note spectrum is mainly to study the fundamental tone and overtone of the note. Therefore, the mathematical representation of musical notes is as follows:

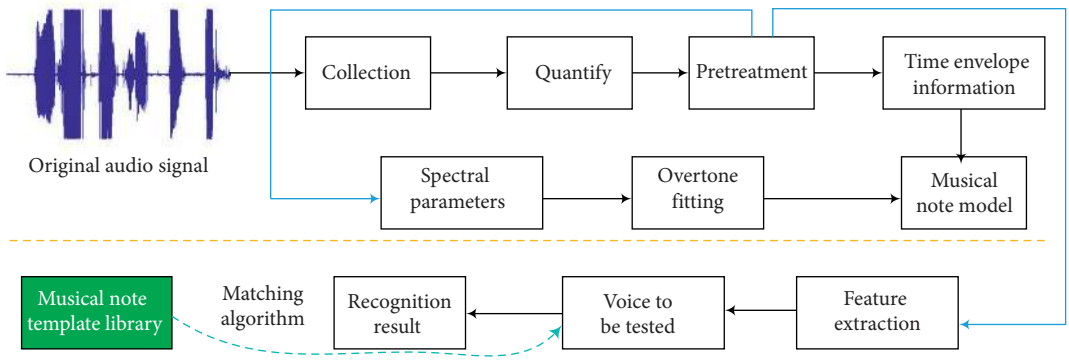

Figure 2: Flowchart of musical note modeling.

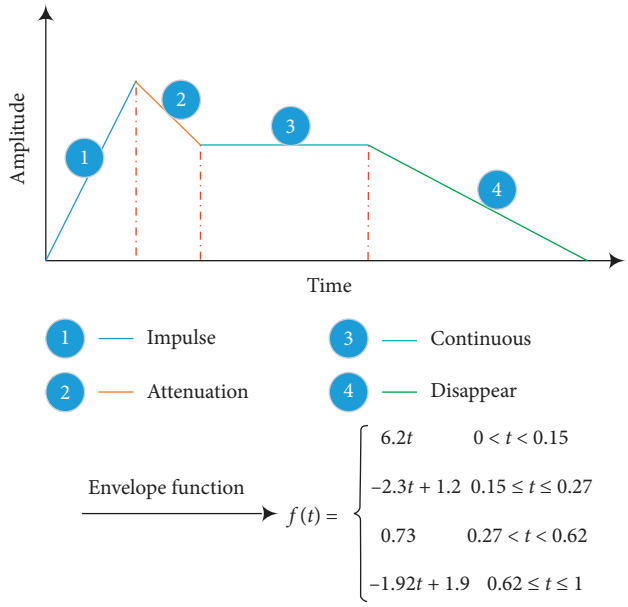

Figure 3: ADSR envelope characteristics.

$$Y = \prod \cos(2\pi n f t) \bullet A_{nf} \bullet E_t. \tag{2}$$

In the formula, f is the fundamental frequency corresponding to the key, nf represents the n-th octave of the fundamental frequency, t represents the duration of the note, A_{nf} is the amplitude fitting function of each frequency, and Et is the note time-domain envelope decay function.

Based on the establishment of the single-note model, the continuous-note signal is the superimposition of the single-note signal in the time domain. According to the mathematical model of the single-note signal, the mathematical representation of the continuous-note signal can be derived; namely,

$$Y = \prod_{i=1}^{k} \prod_{t_{i-1}}^{t_i} \cos(2\pi n f_i t) \bullet A_{nf_i} \bullet E_{t-1}. \tag{3}$$

In the formula, k means containing k single notes, $[t_{i-1}, t_i]$ means the duration of the i note, and f_i means the fundamental frequency of the i note.

4. Improved Speech and Music Signal Separation Based on Negative Entropy Maximization

4.1. Negative Entropy Maximization Method. The most commonly used method of blind source separation is independent component analysis, which mainly uses the independence between signals to separate signals. If the components of the source signal vector are independent of each other, the source signal vector is subjected to matrix transformation. The individual components are also independent of each other. The essence of independent component analysis is the process of separating statistically independent source signals from the mixed signal, which is basically obtained by maximizing or minimizing the objective function. The distribution of the sum of multiple independent random variables tends to be Gaussian; that is, the Gaussianness of the sum of the variables is stronger than the Gaussianness of each variable. We consider a component of $Y(t)yi(t) = wX(t)$, w is a column vector of the separation matrix W and requires w that maximizes the non-Gaussianity of $yi(t)$; then, we separate a component from the observed signal. Commonly used non-Gaussian measures include kurtosis and negative entropy. For zero-mean signals, kurtosis is its fourth-order statistic, which is defined as

$$\text{kurt}(y-1) = E\left(y^4\right) - 3E\left(y^3\right)^2. \tag{4}$$

According to the value of kurtosis, the signal can be divided into three categories according to Gaussian. When the kurtosis is equal to zero, it is a Gaussian signal; when the kurtosis is greater than zero, it is a super-Gaussian signal, and when the kurtosis is less than zero, it is a sub-Gaussian signal. The value of negative entropy is greater than or equal to zero. When the variable obeys the Gaussian distribution, the negative entropy is zero. The kurtosis can be used to approximate negative entropy, but kurtosis is sensitive to singular values. For this reason, an approximation method for negative entropy is proposed:

$$J(y_i) = E[G(y_{i-1})] - E[G(v-1)]^2. \tag{5}$$

In the formula, y_i and v are output variables and Gaussian random variables with zero-mean and unit

variance, respectively, and G is a nonsquare nonlinear function. According to the mixed signal, the nonlinear function can be divided into the following three types:

$$G_1(y) = \frac{1}{a_1}\log_2 \cosh(a_1 y) 0 < a_1 < 3,$$
$$G_2(y) = \exp\left(0.5 \cdot y^2\right), \tag{6}$$
$$G_3(y) = 0.25 \times (y-1)^3.$$

Among them, $G_1(y)$ is suitable for mixing sub-Gaussian signals and super-Gaussian signals, $G_2(y)$ is suitable for mixing super-Gaussian signals, and $G_3(y)$ is suitable for mixing sub-Gaussian signals. The application is based on the Gaussian nature of the signal. We choose a suitable nonlinear function.

This paper uses the method of unsupervised learning. As long as there is no change or small change in W during the two iterations, it can be considered as convergent and an independent component is separated. We use the above steps to extract multiple independent components and iterate out the separation matrix components W_1, W_2, ..., W_n in turn. When a new independent component is extracted in each iteration, the newly obtained W_i is separated from the previously obtained i. The matrix components are orthogonalized to ensure that the newly obtained vector is different from the convergence direction of the calculated vector. The orthogonalization method is as follows:

$$W_{i+1} = W_i - \prod_{j=0}^{i-1} W_j W_{i-1}^{T-1} W_j,$$
$$W_{i+1} = W_i \bullet \left(W_i^{T+1} W_{i-1}\right)^{-(1/2)}. \tag{7}$$

The negative entropy maximization algorithm has the following characteristics: (1) The advantage of the Newton iteration method in the algorithm is the fast convergence speed, which is generally quadratic convergence. (2) The parameters in each iteration process are obtained through the results of the previous step. (3) Only one independent component is extracted after each iteration of the algorithm until convergence, so if you are interested in a component in the mixed signal and have sufficient prior knowledge, you can quickly extract the required component, thereby reducing the calculation of the amount. (4) According to the Gaussianness of the signal, there are three kinds of nonlinear functions that can iterate out independent components. Choosing a suitable nonlinear function can improve the algorithm performance.

4.2. Improved Blind Separation Algorithm for Initial Value Sensitivity. Negative entropy is an important non-Gaussian measurement method. Maximizing negative entropy maximizes the non-Gaussian nature of random variables, thereby making the output components independent of each other. The negative entropy maximization algorithm takes negative entropy as the objective function and the Newton iteration method as the optimization algorithm. Aiming at the problem of sensitivity to initial value selection in Newton's iteration method, this paper replaces Newton's iteration method with Newton's downhill method. By changing the downhill factor, the objective function is in a downward trend and the algorithm's dependence on the initial value is reduced.

The separation of speech and music signals based on the maximization of negative entropy uses negative entropy as the objective function and the Newton iteration method as the optimization algorithm. There are two main problems with the Newton iteration method: large amount of calculation and sensitivity to initial value. Both need to calculate the derivative, which increases the amount of calculation, and when the initial value is too far from the root, the iteration will not converge. Therefore, modifying the Newton iteration method is a way to improve the performance of the algorithm.

The Newton iteration method selects the initial value of the iteration more strictly. If the initial value is not well selected, it may cause nonconvergence. To ensure that the initial value converges in a larger range, the deformed Newton downhill method of the Newton iteration method can be used. The current calculation result and the calculation result of the previous step are processed as a weighted average, and the average value is used as the new approximate value. The process is as follows:

$$x_{k+1} = (1-\lambda)x_k - \lambda \overline{x}_{k-1}. \tag{8}$$

Here, λ is called the downhill factor. Introducing the Newton downhill method in the separation of speech and music signals based on the maximization of negative entropy, the iterative process can be obtained as

$$W(k+1) = W(k-1) - \lambda \bullet \frac{\beta W(k-1) + E\left[W x^T x g(k)^T\right]}{\beta I - E\left[W x^T g'(k)^T\right]}. \tag{9}$$

In order to avoid singular weights, the denominator of formula (9) is not zero. Among them,

$$\beta = E\left[x x^T g W(k)^T\right]. \tag{10}$$

$W(k)$ is the separation matrix component obtained in the previous iteration. The above formula introduces a downhill factor, and the current value and the separation matrix component obtained in the previous step are weighted and averaged to make the calculation result more stable.

4.3. Algorithm Implementation. The flowchart of the blind separation of speech and music signals based on the maximization of negative entropy is shown in Figure 4.

(1) Mix n channels of voice and music signals into m channels of mixed signals, and each channel has a length of N.

(2) Perform zero-mean and whitening preprocessing on the mixed signal, so that the components of the

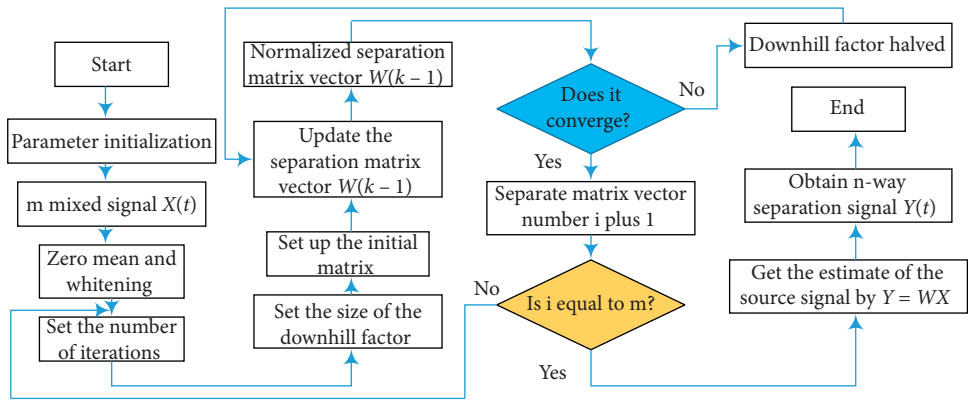

Figure 4: Algorithm flowchart.

whitened signal are independent of each other and meet the condition that the mean value is zero.

(3) Let the number of iterations $k = 0$ and the descending factor $\lambda = 1$, and set the initial matrix randomly.

(4) Update $W(k+1)$ and normalize.

(5) The algorithm converges, and an independent component is estimated; otherwise, the algorithm does not converge, $k = k + 1$, and λ is halved.

(6) Obtain m separation matrix components, and each time a separation matrix component is obtained, it is orthogonalized with the previously obtained components.

(7) Calculate the estimation of the source signal $Y = WX$ from the separation matrix W, and analyze the algorithm performance.

5. Experimental Results and Analysis

5.1. Experimental Data Settings. The music data in the experiment use the music data set MIR-1K released by Hsu Lab. The data set consists of 110 Chinese songs edited into 1000 pieces of music, each piece of music is 4 s–12 s, using 18 kHz to save the accompaniment and singing in the left and right channels of the WAVE file, and the singing part is recorded by amateurs. In order to facilitate neural network training, the music fragments in MIR-1K are divided into audio with a length of 2 s to ensure that the length of the training input data is consistent.

The pure accompaniment and singing voice used in the experiment are the audio stored separately in the left and right channels, and the mixed music used in the experiment is the single-channel audio mixed with the pure accompaniment and singing voice of the left and right channels in the above audio file at 0 d*B*. The hidden layers of the neural network are 3 layers of standard LSTM and 1 layer of bidirectional LSTM. The number of hidden cells in each layer is 128. The training data are the frequency spectrum of audio. 128 points are selected as one frame, and the overlapped half-frames are used as short-time Fourier. For transformation, the timestep is set to 2, the optimizer selects Adam, and batch_size is set to 100. The training process is about 1 hour.

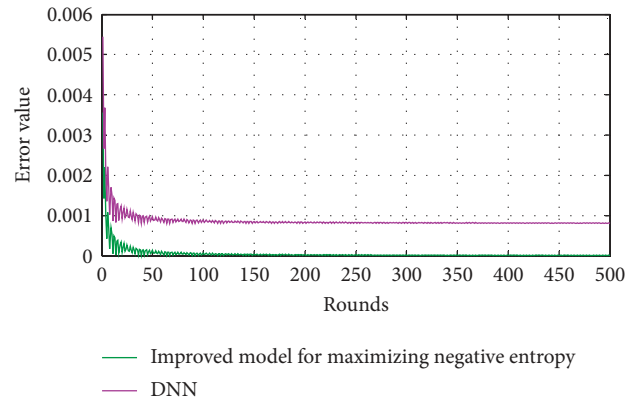

Figure 5: Convergence curve comparison.

5.2. Experimental Results. Convergence speed, as an important evaluation index of the neural network model, is an important criterion for measuring a network model [33–35]. Figure 5 shows the convergence curve of the DNN-based speech separation model and the improved negative entropy maximization model used in this paper. It can be seen from Figure 5 that the algorithm in this paper is superior to the DNN-based model in terms of convergence speed. Although the convergence speed of the two types of models has decreased, the error value still maintains a continuous decrease. When the training reaches about 50 rounds, the convergence speed of the model gradually stagnates, and the error value reaches the limit.

The basic structure of the model in this paper uses the LSTM network, which can make full use of the correlation between the previous and next frames of the spectrum during training. The *c* SA based on discriminative training is used as the training objective function of the model in this paper. It has more advantages than the traditional neural network using the mean square error function (MSE) as the objective function. It can distinguish the difference between different source signals. In the process, a faster convergence rate can be achieved. We choose a piece of music randomly from MI-1K, and its time-domain waveform is shown in Figure 6.

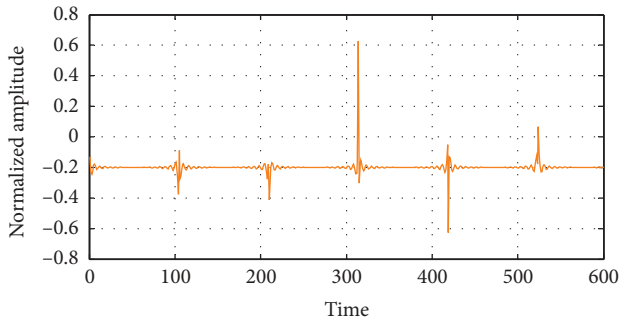

Figure 6: Waveform of music clip.

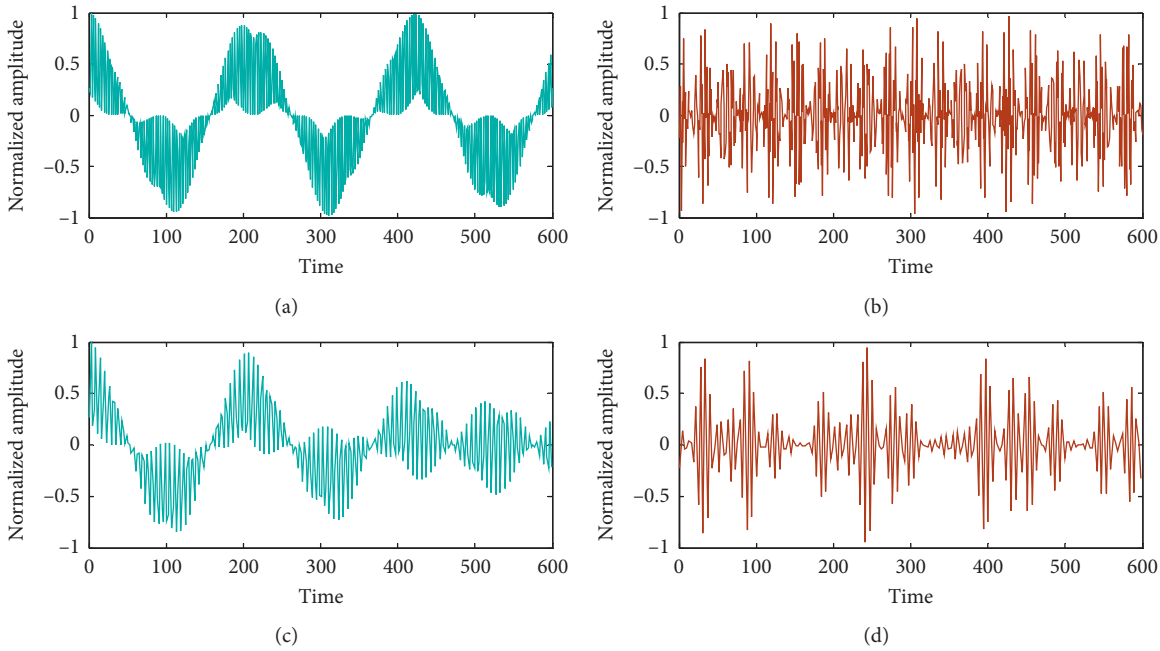

Figure 7: Waveform comparison before and after music segment separation. (a) Original accompaniment waveform. (b) Original singing waveform. (c) Accompaniment waveform after separation. (d) Singing waveform after separation.

It can be seen from Figure 7 that the separated accompaniment and singing have clear waveforms, which are basically the same as the original and pure accompaniment and singing waveforms. Figure 7(a) is compared with Figure 7(c). The separated accompaniment has a slight reduction in amplitude, but it can basically be ignored. According to the actual sound effects, the reduction in amplitude will not affect the actual information expression of the accompaniment. That is, the separated accompaniment has the same melody, rhythm, and pitch as the original accompaniment, and the amplitude reduction will only reduce the loudness of the sound. It can be seen from Figures 7(b) and 7(d) that the separated singing voice has a clear waveform structure, and the peak position and amplitude are consistent with the original singing voice. In the first second of the separated singing voice, there is a slight fluctuation in amplitude, and the original singing voice is basically 0 in this 1 s waveform. According to the relationship between the time-domain waveform and the sound effect, it shows that the separated singing voice appeared during this period of time. Noise interference is also mutually confirmed by the partial distortion of the separated singing frequency spectrum. At the same time, it also shows that when the method of this paper separates the accompaniment and singing, the separation result will produce noise interference for the audio of the silent section.

We use the method in this paper and traditional existing algorithms to separate 1000 pieces of music in the MIR-1K data set, use the blind source separation tool to evaluate and compare, and calculate the global average GSAR (Global SAR). The separation result is shown in Figure 8. It can be seen from Figure 8 that the method in this paper is superior to the separation method based on DNN in the separation index GSAR. This shows that the network used in this paper is more suitable for separating the singing voice in music than the DNN-based network model.

Randomly we select 100 pieces of music with weaker rhythm and 100 pieces of music with stronger rhythm from

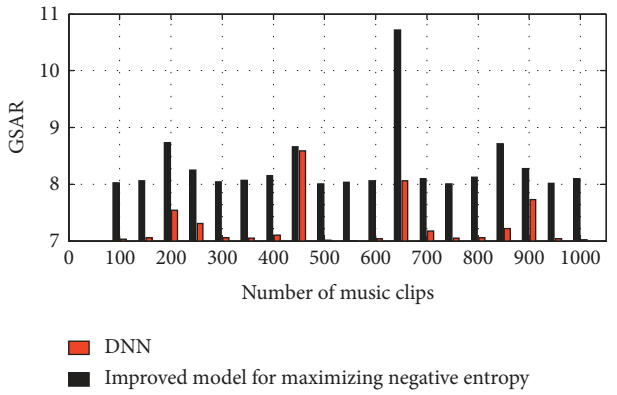

Figure 8: GSAR comparison of the separation results of music fragments in the MIR-1K data set.

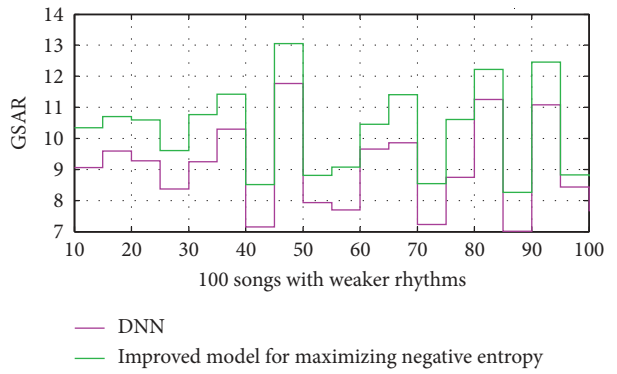

Figure 9: The separation results of 100 pieces of music with weaker rhythms in MIR-1K.

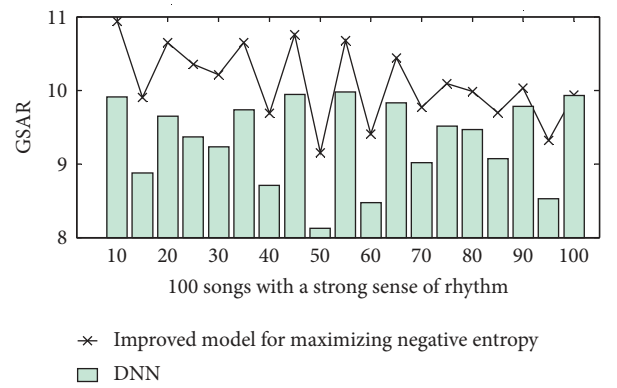

Figure 10: 100 pieces of music separation results with strong rhythm in MIR-1K.

MIR-1K and use the traditional method and the method of this paper to separate, and the separation results are averaged as shown in Figures 9 and 10.

It can be seen from Figures 9 and 10 that the algorithm has a great difference in separation performance for music with a stronger sense of rhythm and weaker music. For music with a weak sense of rhythm, the separation effect of the DNN separation algorithm is relatively poor. This is because the DNN algorithm needs to be modeled according to the melody and beat of the music when separating, and the weaker rhythm music does not have a clear beat, so it is impossible to establish a clear beat model, which leads to the poor separation effect of the DNN algorithm in this type of music. Compared with traditional algorithms, for this kind of music with weaker rhythm, it can be seen from the figure that the method in this paper still maintains a better separation effect.

For music with a strong sense of rhythm, traditional algorithms have a better separation effect when separating accompaniment and singing. For this type of music, the separation effect of the method in this paper is not much different from that of weaker rhythm music. There is no huge change in the separation performance due to the strength of the rhythm, which shows that the method in this paper is different from the traditional nonneural network algorithm in separating music. The method in this paper is less dependent on music samples, and it has good separation performance whether it is music with a strong or weak rhythm. At the same time, it also shows that the method in this paper has good generalization performance and is suitable for separating different types of music.

6. Conclusion

Negative entropy is an important non-Gaussian measurement method. Maximizing negative entropy maximizes the non-Gaussian nature of random variables, thereby making the output components independent of each other. The negative entropy maximization algorithm takes negative entropy as the objective function and the Newton iteration method as the optimization algorithm. Aiming at the problem that the Newton iteration method is sensitive to the initial value selection, the Newton descending method is used instead of the Newton iteration method, and the objective function is changed by changing the descending factor. The downward trend reduces the dependence of the algorithm on the initial value. The experimental results show that the algorithm can separate the source signal well under different initial values. The average iteration time of the improved algorithm is reduced by 26.2% compared with that before the improvement, the number of iterations is reduced by 69.4%, and the iteration time and the number of iterations are both relatively low. Fluctuations in a small range better solve the problem of sensitivity to the initial value. The separation effect under multiple sets of different mixing matrices shows that the separation effect has nothing to do with the mixing matrix. The results show that the new objective function can significantly improve the separation performance of the neural network. Compared with the existing music separation methods, the method in this paper shows excellent performance in both accompaniment and singing in the separation of music. The study of the overall feature space of music involves the level of music understanding. At this time, how to combine the relevant theories of musicology to extract essential features and better express

the structure and information in music has become the key to solving the problem. Through the analysis of a large number of music data samples, combined with music theory and subjective evaluation methods, we compare the mapping relationship between various basis functions or dictionary functions and the overall characteristics of music structure, music style, and emotional connotation and determine basis functions or dictionaries. The feature subspace corresponding to the function is a useful research idea in the future.

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during the current study.

Consent

Informed consent was obtained from all individual participants included in the study references.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] J. P. Bello, "Measuring structural similarity in music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2013–2025, 2011.

[2] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: an overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.

[3] D. Hu, F. Nie, and X. Li, "Deep binary reconstruction for cross-modal hashing," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 973–985, 2019.

[4] M. Pearce and G. Wiggins, "Improved methods for statistical modelling of monophonic music," *Journal of New Music Research*, vol. 33, no. 4, pp. 367–385, 2004.

[5] J. Torres-Sánchez, F. López-Granados, and J. M. Peña, "An automatic object-based method for optimal thresholding in UAV images: application for vegetation detection in herbaceous crops," *Computers and Electronics in Agriculture*, vol. 114, pp. 43–52, 2015.

[6] F. Bimbot, E. Deruty, G. Sargent, and E. Vincent, "System & contrast," *Music Perception*, vol. 33, no. 5, pp. 631–661, 2016.

[7] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[8] G. Song, D. Wang, and X. Tan, "Deep memory network for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1261–1275, 2019.

[9] J.-F. Paiement, Y. Grandvalet, and S. Bengio, "Predictive models for music," *Connection Science*, vol. 21, no. 2-3, pp. 253–272, 2009.

[10] T. Liu, Q. Miao, K. Tian, J. Song, Y. Yang, and Y. Qi, "SCTMS: superpixel based color topographic map segmentation method," *Journal of Visual Communication and Image Representation*, vol. 35, pp. 78–90, 2016.

[11] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali, "Music in our ears: the biological bases of musical timbre perception," *PLoS Computational Biology*, vol. 8, no. 11, pp. 1–16, 2012.

[12] G. Grindlay and D. P. W. Ellis, "Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1159–1169, 2011.

[13] E. Yu, J. Sun, J. Li, X. Chang, X.-H. Han, and A. G. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1276–1288, 2019.

[14] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "A sticky HDP-HMM with application to speaker diarization," *Annals of Applied Statistics*, vol. 5, no. 2A, pp. 1020–1056, 2011.

[15] H. Turić, H. Dujmić, and V. Papić, "Two-stage segmentation of aerial images for search and rescue," *Information Technology and Control*, vol. 39, no. 2, pp. 138–145, 2010.

[16] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1229–1240, 2014.

[17] Y. Zhang and Q. Yang, "An overview of multi-task learning," *National Science Review*, vol. 5, no. 1, pp. 30–43, 2018.

[18] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 128–141, 2018.

[19] M. J. Johnson and A. S. Willsky, "Bayesian nonparametric hidden semi-Markov models," *Journal of Machine Learning Research*, vol. 14, pp. 673–701, 2013.

[20] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 707–710, 2007.

[21] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 318–326, 2008.

[22] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Trans. Audio Speech Lang. Process,* vol. 18, no. 8, pp. 2121–2133, 2010.

[23] Y. Hu, L. Zheng, Y. Yang, and Y. Huang, "Twitter100k: a real-world dataset for weakly supervised cross-media retrieval," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 927–938, 2018.

[24] S. Cherla, H. Purwins, and M. Marchini, "Automatic phrase continuation from guitar and bass guitar melodies," *Computer Music Journal*, vol. 37, no. 3, pp. 68–81, 2013.

[25] L. Stankovic, "ISAR image analysis and recovery with unavailable or heavily corrupted data," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 3, pp. 2093–2106, 2015.

[26] V. Arora and L. Behera, "Multiple F0 estimation and source clustering of polyphonic music audio using plca and HMRFs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 278–287, 2015.

[27] J. Yang, Z. Zhang, W. Mao, and Y. Yang, "Identification and micro-motion parameter estimation of non-cooperative UAV targets," *Physical Communication*, vol. 46, Article ID 101314, 2021.

[28] B. Yang, X. Cheng, D. Dai, T. Olofsson, H. Li, and A. Meier, "Real-time and contactless measurements of thermal discomfort based on human poses for energy efficient control of buildings," *Building and Environment*, vol. 162, Article ID 106284, 2019.

[29] W. Wei, E. S. L. Ho, K. D. McCay et al., "Assessing facial symmetry and attractiveness using augmented reality," *Pattern Analysis and Applications*, pp. 1–17, 2021.
[30] J. Qian, X. Cheng, B. Yang et al., "Vision-based contactless pose estimation for human thermal discomfort," *Atmosphere*, vol. 11, no. 4, p. 376, 2020.
[31] W. Wang, T. Tang, F. Xia, Z. Gong, Z. Chen, and H. Liu, "Collaborative filtering with network representation learning for citation recommendation," *IEEE Transactions on Big Data*, vol. 20, no. 1, pp. 1–151, 2020.
[32] J. Lee and S. K. Yoo, "Design of user-customized negative emotion classifier based on feature selection using physiological signal sensors," *Sensors*, vol. 18, no. 12, p. 4253, 2018.
[33] T. T. Erguzel and B. Stün, "The development of A fuzzy logic model-based suicide risk assessment tool," *The Journal of Neurobehavioral Sciences*, vol. 7, no. 3, pp. 156–163, 2021.
[34] E. Olcay, C. Schöttl, A. Schttl, M. A. Zaggl, and B. Lohmann, "An agent-based model of an online collaboration community by using fuzzy logic," *IFAC-PapersOnLine*, vol. 52, no. 13, pp. 665–670, 2019.
[35] J. A. Meda-Campaña, "On the estimation and control of nonlinear systems with parametric uncertainties and noisy outputs," *IEEE Access*, vol. 6, pp. 31968–31973, 2018.