Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 3627113, 11 pages
https://doi.org/10.1155/2022/3627113

Research Article

Violin Teaching Improvement Strategy in the Context of Intelligent Internet of Things

Yifeng Zhang

Taizhou University, Taizhou, Zhejiang 318000, China

Correspondence should be addressed to Yifeng Zhang; zhangyifeng@tzc.edu.cn

Received 16 May 2022; Revised 5 June 2022; Accepted 9 June 2022; Published 6 July 2022

Academic Editor: Naeem Jan

With the progress of science and technology, the construction of intelligent teaching is also developing and making progress. Intelligent teaching is a kind of teaching mode with the help of emerging information technology such as the Internet of things, cloud computing, and big data to build a diversified classroom. The construction of intelligent teaching is one of the important parts in the field of education informatization, and it is also an active field. The research in this area is beneficial to China's growth of information education. Online education has progressed significantly in the context of the intelligent Internet of Things. The focus of this paper is on violin online education. In the course of the research, it was discovered that violin teachers have little control over their students' after-school practice results, and that parents of violin students have little time to supervise their children's practice at any given moment. To solve this problem, the mixed speech composed of speech and violin accompaniment is selected as the research object. Audio signal analysis and neural network algorithm are used for analysis and research. The time domain and frequency domain characteristics of various audio signals are analyzed, including sparsity of speech signals and repeatability of music signals. The logarithmic power spectrum of audio signal is selected as the characteristic parameter, the sample is preprocessed and the feature is extracted, and the sound source separation model based on deep neural network is realized. The PESQ index was used to evaluate the separation results after using the audio source separation model to separate the mixed speech consisted of speech and violin accompaniment. By comparing the PESQ evaluation results of 40, 50 and 60 iterations, the model converges when the number of iterations is 50. The comparison between the model and L-MMSE algorithm shows that the performance of the model is better in speech source separation. The system will serve for the automatic evaluation of violin teachers' homework and the remote practice supervision of teachers and parents.

1. Introduction

With the development of society, the service industry accounts for an increasing proportion in all industries of the country and companies also require young people to have more soft power in the process of recruitment. It is a personal plus for job seekers to have a special talent in music. In the process of educating their children, they pay special attention to quality education. The entertainment atmosphere of modern society is increasing day by day. Various singing competitions and talent shows on the media also promote the enthusiasm of young people to learn music. There are more and more music education training institutions in the market, and people's access to music education has become more abundant.

In the field of music, the study of various instruments is an important component. The violin is the most important instrument in the modern orchestra, which is widely spread all over the world. At the same time, the violin plays an extremely important role in all Musical Instruments, so it is called "after Musical Instruments." Violin sounds like human voice, it is suitable for the expression of tenderness, enthusiasm, happiness, solemn, and other feelings. There are many advantages to learning the violin. Among many musical instruments, the violin is small in size and light in weight, which is suitable for carrying around and playing outside [1]. The violin needs both linkage, effectively promote the balanced development of the left and right brain; often, playing the violin can improve one's own music accomplishment and aesthetic ability, and at the time of the violin playing, one makes oneself immersed in a beautiful violin performance, can edify their sentiment, reduce stress and depression, and give yourself a happy day.

With the development of the Internet, the transmission cost of information has dropped sharply. The education industry can be highly informationized due to its knowledge attributes, so online education is developing rapidly. In the first half of 2020, due to the impact of the epidemic, all primary and secondary schools and institutions of higher learning could not reopen normally, and online education has become an important basis for continuous school suspension due to its noncontact nature [2]. With so many benefits, online education is developing rapidly in China, with many excellent products emerging in the field of music education. Traditional education requires space and small class teaching, while the Internet nature of online education can greatly reduce the cost of education, so that the cost of education is reduced and even free online courses appear. The wide audience of Internet products makes the concentration of the industry increase rapidly, excellent teaching materials and teachers can be spread quickly and widely, and the quality of education that people get is also higher than the traditional education model [3]. With the development of artificial intelligence technology, all sorts of intelligent functions are also embedded in music education software, such as music practice recommendations, intelligent music-level rating, the intelligent software and connect with electronic violin, realize the combination of software and hardware, can according to the practitioners of violin key intelligent error correction, intelligent with spectrum, and greatly improve the learners' learning enthusiasm and efficiency. With the popularity of smart phones, a large number of mobile software for music education has come into our sight [4]. Through market research, we found that most of the musical instrument education software in the market focus on providing musical instrument teaching videos for learners, or real-time transmission of the practitioner's finger-pointing and instrument pronunciation to the instrument teacher through the video function of the mobile phone, so as to realize the accompaniment and remote guidance for musical instrument learning.

Although there are so many musical instrument-teaching products on the market, we also found some market gaps in the market survey. Children generally need 1–2 hours to practice the violin every day, but in daily life, parents are busy with work, so it is difficult to spend a large period of time to supervise their children's practice. Most of the time, children need to practice alone at home, and the quality and length of practice are difficult to guarantee. Moreover, the instrument teacher's time is limited, and the children can finish the after-school exercise homework with quality and quantity guaranteed. Parents are not professional instrument teachers, and it is difficult to evaluate their children's practice results quantitatively.

Music is a collection of distinct frequency components at different periods in terms of signal processing. Even if the identical piece of music is played by different instruments, there will be significant variances in hearing, as well as distinct time and frequency domain features. A very important research direction is to figure out how to utilize a computer to recognize these traits and to analyses and recognize music signals. This not only helps people to understand the essence of music more deeply but also can be applied in music teaching and searching. Based on this, people can analyze and process the massive music resources on the network and reduce the noise of audio to obtain better audio quality. For the field of audio processing, is the focus of all parties, there are many foreign research institutes and colleges on the field of audio processing research, and achieved some results [5–10]. It can also facilitate information retrieval of music. Thus, greatly facilitated the dissemination and communication of music. The sound source separation model of a deep neural network is utilized to separate mixed speech, and the speech quality evaluation (PESQ) index is employed to evaluate the separation results in order to achieve sound source separation. Then, using a deep neural network-based sound source separation model, the sound source of the mixed audio of speech and violin is separated.

The following is the paper's organization paragraph: In Section 2, the related work is provided. The suggested work's approaches are examined in Section 3. The trials and results are discussed in Section 4. Finally, the research job is completed in Section 5.

2. Related Works

2.1. Research Status of Intelligent Internet of Things. Artificial intelligence of things (AIoT) involves information processing, artificial intelligence, Internet of Things, fog computing, edge computing, cloud computing, and many other technologies and is multilearning. The product of intersection and integration of science and technology has been widely used in smart city, smart home, smart medical treatment, intelligent transportation, intelligent manufacturing, and other scenarios [5]. However, the development of the Intelligent Internet of Things is still in its early stage, facing many problems and challenges that need to be solved, such as architecture, security and trust management, heterogeneous data fusion and processing, heterogeneous network fusion, and complex event processing collaboration. However, the wide interest, participation, and support from industry and academia makes us confident about the future development of intelligent IoT [6]. It is believed that in the near future, the intelligent Internet of Things field will be born with strong innovative and influential landmark results, innovative application mode, and new related-technology solutions, etc., to promote the artificial intelligence Internet of Things continue to mature.

At present, AIoT implementation, especially commercial implementation, is mostly based on 4G technology. Compared with 4G, 5G will bring great changes in communication and data exchange. Its high speed and low delay are expected to further expand and extend the application scenarios of AIoT, such as real-time application scenarios (remote control, remote intelligent medical treatment, real-time monitoring, etc.,) that require higher bandwidth and time limit [7]. However, challenges such as energy consumption, multiscene/cross-domain service collaboration, interoperability, and user-orientedness must be addressed before 5G can be used in AIoT. Javaid discussed the inevitability of incorporating artificial intelligence technology into the Internet of Things based on 5G network architecture, as well as the impact of AI on 5G Internet of Things dynamic spectrum management, heterogeneous device integration, interoperability, and energy savings. Wang proposes a next-generation Internet connectivity framework called 5G Smart IoT. Figure 1 illustrates the AIoT model.

The architecture makes full use of the latest communication technologies as well as AI technologies such as big data mining, deep learning, and reinforcement learning to help intelligently process Iot big data and attempt to optimize communication channels. Patel et al. roposed that AIoT model should be constructed from the perspective of engineering application based on the scientific theoretical knowledge related to AI and IoT [8]. Hu designed a set of home network communication protocols for the needs of home sensing and control, and based on this, designed a shared home network construction framework. Lin et al. has developed a household robot that can be remotely controlled using a smartphone [9]. Adiono has proposed a smart home software application based on the Internet of Things, capable of automatically intelligent control of home appliances based on operational commands or behaviors, and protecting communication security and privacy through encryption.

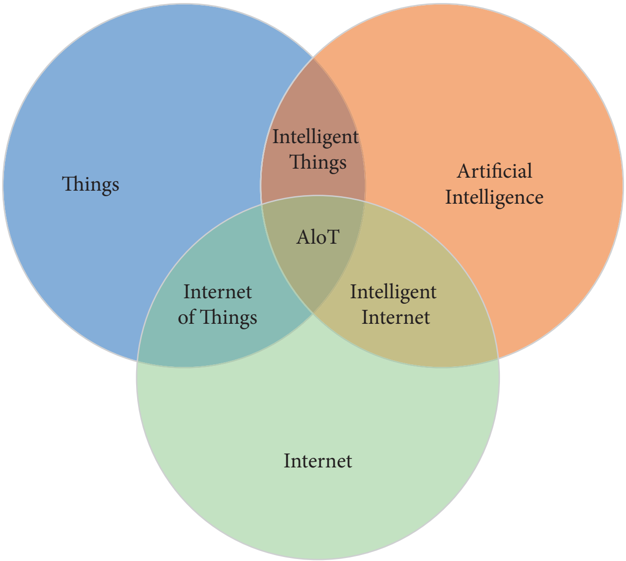

Figure 1: AIoT is the fusion of IoT and AI.

2.2. Research Status of Violin Teaching.

With the rapid development of social economy in our country, on the basis of the materials that are rich, people having higher spiritual pursuit, especially in recent years, and the state is advocating quality education unceasingly thorough popular feeling, prompting more parents to choose to let children have access to art education, especially in recent years, increasing the number of people to learn the violin, but the violin teachers are scarce. The traditional one-to-one model cannot meet the huge demand of violin market, which promotes the development of violin collective class in China [10]. However, due to the violin class collective development time being relatively late in our country, an imperfect teaching system, teachers and parents' lack of understanding of collective violin lessons, and various problems in the practice process, collective study of violin class teaching to develop violin education universally, and respond to a nation's call, cultivation of all-round development of talent has a profound significance.

With the development of the times, especially the popularization of Internet, multimedia, and mobile phone, mobile terminals and the cross application of science and technology in other fields, the teaching of violin collective class should closely keep up with the times and serve for our violin teaching. First of all, one should make full use of multimedia in our teaching processes, such as in class for the children to find related-audio video data section of this course, let the children know as much as possible about their music from an image understanding, can also share with the children violin-playing videos to enhance the children's interest, and to join in the process of teaching and learning violin-content-related pictures, and videos. Or after class, a WeChat group can be established, where children can send practice videos to the group, and teachers can make timely comments, so as to enhance the efficiency of children's violin practice and the frequency of communication with teachers [11]. Second, science and technology in other fields can be flexibly applied to teaching, such as the combination of recording major and violin teaching. First, computer music arrangement function can be used to create and compose music that students are learning to play with pleasant accompaniment. Second, after students learn several pieces of music, teachers can organize students to go into the recording studio to record, and collective class students can adopt a variety of ways to play music in undivided play, which can not only stimulate students' great interest but also learn to listen to evaluate their own voice and performance in the recording process [12]. Teachers can post students' recording results on various short video platforms once they have been recorded, which can boost students' confidence, promote violin art, and help the violin flourish and spread. Finally, the Internet and mobile terminals are employed to break traditional teaching's space and time limits. At present, there are many apps available for teachers to teach violin online. Teachers can connect students from different places on the Internet at the same time for online group class interaction. By using the new and interesting operations in the App, students' enthusiasm in learning violin can be improved and their sense of participation strengthened [13]. At present, there are many musical instrument education products on the Internet, among which the most famous are Knger, Violin Bar, violin coach, and practice together. They are all mobile applications and each has its own characteristics:

(1) Knger is a main instrument in the teaching of music education software, it needs for different levels of

users, the free online video teaching instrument, its characteristic is to have a lot of instruments teaching video and song resources, and has carried on the system of teaching resources classification, make students of different levels choose according to their own situation different classes and process to learn [14]. Kinger also provides a music-level rating test, but the content of the test focuses on music theory knowledge, listening to music, and recognizing music.

(2) Violin Coach is a violin education software that focuses on game teaching. It also provides free video courses and music library. Its feature is that it integrates a series of games of listening to music and recognizing music in the application. From the most basic music theory, users can learn violin even if they do not know the stuff. The software can obtain the user's playing process by connecting the intelligent instrument, but the cost is high because the intelligent instrument hardware specified by the software needs to be purchased. It is mainly aimed at elementary students, students can play and learn according to the key tips in the process of practice; this way can make the students playing and learning very interesting.

(3) To play the violin, and to have both web applications and mobile app to play the violin, please provide the free score and teaching video, and provide the music-culture practice, listening practice, such as tools, compared with the former two apps. To play the violin, the characteristic is to construct a perfect online education community, in the online community. There are many famous teachers, teacher ancillary practice provides rich resources, and students performance audio can be posted online to invite teachers to comment on it, forming online interactive teaching [15]. Intelligent violin to play the violin, a song to practice can be connected, through the process of intelligent monitoring user violinist, rating and to play, but playing process can only follow the melody fixed rhythm, according to the screen display of notes being played, at the same time due to the intelligent guitar and intelligent violin also specifying the hardware for software, cost is higher.

(4) With practice a large, medium, and violin practice scoring procedures, it can practice audio collection, and then for practitioners, feedback where wrong, pitch, rhythm, speed deviation, let the student get the practice feedback in time, improve the learning efficiency, but the practice together requires practitioners to score beat synchronization in the app, once the score of the practitioners and software Out of sync will affect the rating of the app, which connects to an electric violin to track the user's performance.

2.3. Research Status of Sound Source Separation. Mixed audio source separation has become a hot topic in the field of audio signal processing in recent years, thanks to the rapid development of modern music signal-processing technology and computer technology. Monophonic mixed audio source separation has gotten a lot of attention in recent decades as a prominent instance of audio signal processing. For the field of audio processing, is the focus of all parties, there are many foreign research institutes and colleges on the field of audio processing research, and achieved some results [16]. The Digital Music Center of Queen Mary University of London and the Music and Acoustics Computer of Stanford University in the United States have made outstanding research in this field. There are also some international competitions or events in the direction of audio signal processing, such as the famous competition. This competition provides a unified data source as a test for many cutting-edge technologies and algorithms in the field of music information retrieval. It is also a fair and reliable evaluation platform in the field of music signal processing.

Pendor used nonnegative matrix decomposition separation technology to study the establishment of repetitive structure of music and how to extract background music. Although NMF separation technology may split the music spectrum used for research into numerous fundamental audio files, if there are multiple musical instruments in the audio, time-frequency aliasing will occur when the different musical instruments play. Ni adopts the two-dimensional sparse matrix nonnegative matrix decomposition model extended from the traditional NMF algorithm and adopts the empirical mode decomposition algorithm to preprocess the mixed music signals, and clustering the audio signals to make them become the instrument audio source [17]. Because the prior information of audio harmonic spectrum is easily constructed from sine model, sine model is often used as the fitting model of audio harmonic signal. Ren constructs three different harmonic spectrum peaks in the instrument audio structure fitted by sinusoidal model and uses filters to deal with overlapping harmonics to achieve the purpose of instrument audio separation. Guo uses the method of least mean square frequency estimation to realize the separation of mixed audio sources based on sinusoidal model. Duan et al. proposed a monophonic speech source separation algorithm based on interframe correlation and fundamental frequency state [18]. Based on the prior information of the music signal, Xiao used principal component analysis to fit the probability model of the musical instrument's mood, and then synthesized the audio signal by stacking harmonics matching the model.

Many machine-learning techniques have been used to the direction of audio source separation and have yielded positive results. Zou deduces the fundamental frequency and harmonic parameters of different audio frequencies using a Bayesian probability model, extracts characteristics, and uses the extracted features to differentiate instrument sources. The existing vector quantization algorithm is improved by using a large amount of prior histogram of training data to achieve speech separation. The pattern-driven approach uses some statistical methods to simulate potential source signals [19]. Although pattern-based techniques provide high-quality isolation of speech and music output, computational complexity is a major disadvantage of these approaches. Liu proposed a hybrid PCA-VQ model based on k-means clustering. The goal of PCA is to map data to linear mappings in lower-dimensional Spaces, so as to maximize data changes in the new space.

As an efficient algorithm built by imitating the neural structure of human brain, neural network algorithm has been widely used in many fields, such as image recognition, and also used for mixed-audio source separation. Ma introduces the development of a noise reduction module based on neural network. However, it is necessary to adapt to the noise with mismatched speech in the test. Bagaam studies a blind separation algorithm combining empirical mode decomposition and mutual information maximization for underdetermined speech music signal separation. Tzinis et al. proposed an improved supervised single channel signal separation method based on deep recursive neural network (DRNN) model [20]. The separation ability of neural network is improved. The fusion of neural network algorithm and traditional audio source separation algorithm is also the research direction of audio source separation. Sachdev combined the traditional NMF algorithm with the deep neural network algorithm, and used NMF to initialize DNN for single-channel speech source separation.

3. Algorithm Design

3.1. Overall Framework. In this paper, we study that the module is a part of the violin online education platform. The platform's goal is to make full use of the Internet connection properties. Low cost makes the violin practitioners at home get good coaches of online teaching and guidance. For learners to provide high-quality video self-study resource, coaches, parents, and the violin hand, all the activities are transferred to online. Through online, the automatic operation of the platform can reduce the teaching cost of coaches, improve the supervision efficiency of parents, and fully mobilize the enthusiasm of violin players to practice.

In order to achieve the goal of audio source separation of mixed audio, it is necessary to analyze the characteristics of mixed audio of human voice and various musical instruments, and find out the characteristics that can distinguish different kinds of audio. The selection of characteristic parameters is very important and will directly affect the result of acoustic source separation. How to select the appropriate characteristic parameters is worth studying? In order to select the characteristic parameters, the time domain and frequency domain characteristics of all kinds of audio should be analyzed first to find out the different characteristics that can separate all kinds of audio sources, and then use these characteristics to separate different audio sources. The overall frame of audio source separation is shown in Figure 2, and the time domain and frequency domain features of audio signals are extracted.

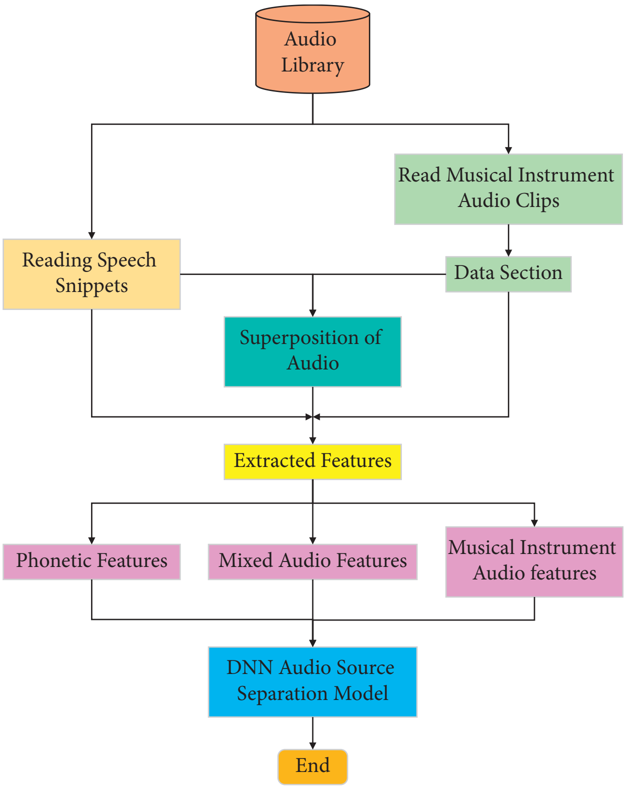

Figure 2: Frame diagram of audio source separation model.

The musical signal's temporal properties are determined by the instrument being performed. Distinct instruments make different sounds, and varying instruments produce different audio durations in the time domain. Moreover, the energy distribution of audio signals is also different in different time periods. The four periods are collectively called ADSR. When a musical instrument is played, the first thing it goes through is the initiation stage, which means the beginning of a sound. The energy will rise sharply at the beginning of a sound, and then it will fall back, which is the attenuation period. After the attenuation period, it enters the stationary period, and the amplitude lasts almost unchanged for some time, and then gradually decreases until it disappears. The initiation stage of a violin is obvious, which shows that the energy will rise sharply at the beginning of a note, while for the same note, the energy peak value of a violin is the same amplitude value [21]. It is worth mentioning that the time domain waveform of violin audios and the inch domain waveform of violin audios can be seen that the change of violin amplitude is more gentle than violin. This shows that a violin has a better continuity of sound, that is, the vibrations produced by striking the keys of a violin are more regular than those produced by rubbing the string with the bow.

In addition to the time domain characteristics, speech signals also have some frequency domain characteristics. From the perspective of human subjective feelings, the frequency range felt by human ears is between 20 and 20 kHz, and the sound in the frequency band gives people the feeling of no *N*. For human hearing, the frequency band between 150 and 250 Hz is the low-frequency sensitive part of human voice. The measure of 4–6 kHz is the middle and high frequency part of human ear, which is also the most sensitive part of human ear [22]. About 10 kHz is the high-frequency sensitive part of human voice. In general, the higher the frequency of the signal, the clearer the sound. Its basic principle is to add the fast Fourier transform of window function. This can be used as a technique for rough processing of musical signals. At present, in the field of deep learning, more and more rough processing data are directly input into neural network, and the neural network model can learn and analyze the characteristics of sample data by itself. For human's own perception system, the sound of a long period of time is also divided into small segments for analysis, and the SIFT process is similar to human's perception of sound to a large extent.

3.2. Data Preprocessing. As mentioned above, this paper will study the sound source separation of speech audio and instrumental audio. In order to analyze the audio characteristics of voice and instrument audio, a certain number of music fragments are required as data sets, that is, samples. The quality and quantity of these samples have great influence on the result of subsequent feature extraction and model training. Here, the selection, slicing, audio superposition, and feature extraction of voice signal and music signal are described. The realization of slicing, audio superposition, and feature extraction is mainly introduced to pave the way for the following part of model realization.

The data set of the voice part comes from TIMIT sound library, while the music part is violin music downloaded from the Internet. The second part is the data set used for separation of mixed audio of multiple Musical Instruments. Since this paper studies instrument separation of violin, violin music downloaded from the Internet is used.

The data set of voice and audio comes from TIMIT Database, which is a 3-second speech fragment. It is suitable for the following work of audio overlay and feature extraction [23], but used to separate instruments violin audio and audio downloads from the Internet works violinist audio and the audio average length in 1 hour or so. Some for two hours, so the length of the audio file is usually the size of 1–2 g, and so not suitable to be used directly as input audio and directly with audio function, etc. Therefore, it is necessary to slice violin audio and violin audio. In this paper, each piece of music is cut into 9 seconds, and the small pieces are named according to the song name and serial number, and the slice pieces of different pieces are stored in different folders.

The supervised network approach is used, and the matching pure audio is necessary, because the research object of this paper is mixed speech and mixed musical instrument audio; so you need to superimpose the two. Since the stacking ratio of the two kinds of audio is similar to the SNR, the SNR is used to refer to the stacking ratio of the two kinds of audio in this paper. In this paper, mixed voice and mixed instrument audio are needed for sound source separation. As supervised neural network algorithm is used for model training, a large number of mixed audio and corresponding pure audio data pairs at each SNR are needed to make the trained model have good generalization ability. On the Internet, audio resources that meet such conditions are hard to find. In order to facilitate the research, the voice fragments in time library and violin fragments are directly superimposed here to obtain the mixed voice audio. Add the violin fragment and violin fragment to get the mixed instrument audio.

The previously obtained audio of mixed voice and mixed instrument, as well as the corresponding pure audio data, are only original audio data and cannot be input as samples. Audio features need to be extracted from them for the following model training process. The final output audio energy of mixed audio signal is the effect of the superposition of the energy of audio components, and the difference in time domain is not obvious, and different instruments play audio differently in the frequency domain. Logarithmic power spectrum is selected as the feature. From the time domain waveform of music signal, we can know that music signal is nonstationary signal, and the effect of Fourier transform processing nonstationary signal is not very good. Signals that are very different in the time domain have very uniform spectra, making it difficult to distinguish these nonstationary signals. In the field of audio signal processing, the short-time Fourier transform can be used to deal with nonstationary signals. STFT uses windowing function to decompose the whole time-domain process into multiple short time segments, so that each segment can be regarded as stationary signals, and then Fourier transform is used to obtain frequency domain and extract frequency domain features.

3.3. Sound Source Separation Model. The previous slicing, audio overlay, and feature extraction of original voice and instrument audio are all intended to provide feature data for the training of the following audio source separation model. Therefore, the specific content of sound source separation model based on deep neural network is introduced here. It should be noted that the sound source separation model based on deep neural network has designed a five-layer model.

The process of deep neural network (DNN) for music separation is to use artificial neural network to train music signals, so that the neural network can learn the spectrum characteristics of each audio component in the mixed audio. Finally, the network has the ability to output the mixed audio components separately. Depth generally refers to the number of hidden-layer neural network that is greater than or equal to three neural network, and the depth of the most widely used neural network model is the BP neural network. BP neural network is used in the training process, and there is the need to constantly seek out the error between the output signal and the desired output signal, back to the former level network and the network weight coefficient of each layer in the ongoing correction until the error is less than a given threshold. This is also known as the back-propagation process.

BP neural network generally consists of input and output model, weight optimization, error model, and activation function. The following are the meanings and calculation formulas of each model:

(1) Node output model.

The node output model formula of the hidden layer is shown in equation (1):

$$H_j = f\left(\sum_{i=1}^{M} w_{ij}x_i - \theta_j\right). \tag{1}$$

The output model formula of nodes in the output layer is shown in equation (2):

$$Y_k = f\left(\sum_{j=1}^{N} t_{jk}H_j - \theta_k\right). \tag{2}$$

(2) The activation function

The activation function, also known as the transfer function, reflects the impulse intensity of the lower input to the upper node. Usually, the activation function is nonlinear, monotonic, and differentiable; so you take an s-type function and the range is (0, 1), which is the function. Tanh function is also a common activation function. Different from function, tanh function has a mean value of 0, so the actual effect of tanh function is better than that of function. Its mathematical formula is shown in equation (3):

$$\tanh(x) = 2\text{Sigmoid}(2x) - 1. \tag{3}$$

(3) Error model:

BP neural network will continuously work out the error between the output signal and the expected output signal in the learning process, and feed back to the front-level network. The error calculation formula is shown in equation (4):

$$e_p = \frac{1}{2}\sum_{i=1}^{N}\left(t_{pi} - o_{pi}\right)^2. \tag{4}$$

(4) Weight optimization:

When the BP neural network calculates the error between the output signal and the expected output signal 3, the front-layer networks at all levels update the weight values according to the established rules. In this process, the weight coefficients of each layer in the network are constantly revised, which is called the self-learning model of BP neural network. This paper adopts weight optimization of driving quantity, and its expression is shown in equation (5):

$$w_{ij}(\text{new}) = \eta\phi_i o_j + \alpha w_{ij}(\text{old}). \tag{5}$$

The sound source separation model of deep neural network in this paper is shown in Figure 3, which consists of input layer, hidden layer (three layers), and output layer. The number of neurons in each layer is 1419, 2048, 2048, 2048, and 129, respectively. The specific process is that the input vector passes through the input layer, passes through three hidden layers in turn, carries out feedforward calculation with the weight matrix of each layer, and finally reaches the output layer to obtain the output vector. Then the self-learning model of BP neural network can minimize the loss function by modifying the weight coefficients of each layer in the network. After the training meets the conditions, the iteration is stopped and the training result, which is the weight coefficient of each layer, is returned. The trained model will be used to separate the mixed audio.

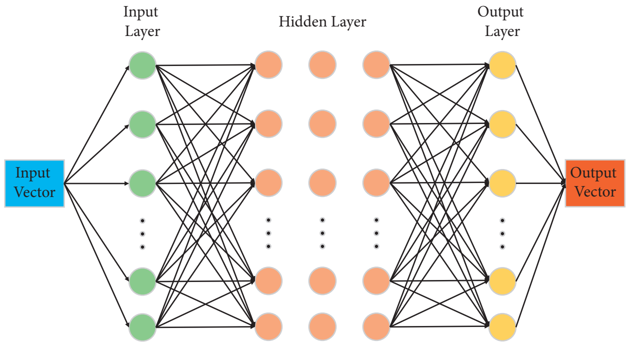

FIGURE 3: Five-layer deep neural network sound source separation model.

4. Experiments

4.1. Experimental Data Set. For the separation of mixed speech audio. The mixed audio composed of voice and violin accompaniment needs to be obtained first. The data set of the voice part comes from TIMIT Database, and the data of TIMIT Database are all speech fragments ranging from 1 second to several seconds. A 3-second speech segment is chosen here to assist the subsequent process of audio overlay and feature extraction. The violin accompaniment audio is the violin solo audio acquired from the Internet, which is normally more than half an hour long, resulting in a huge file, but the audio overlay operation requires the voice and the violin accompaniment audio to be the same length. Therefore, the violin accompaniment audio needs to be preprocessed and cut into 3-second segments; then comes the audio superposition stage, in which the voice clip and violin accompaniment clip of equal length are superimposed according to six different signal-to-noise ratios (−5, 0, 5, 10, 15, 20).

The training process of the model is as follows: first, the speech fragment and the violin accompaniment audio fragment are obtained from the audio library respectively. For the violin accompaniment audio fragment, slice processing is also required. The audio is then superimposed on the equal length of the speech clip and the violin accompaniment audio clip to obtain the mixed speech. Then, feature extraction is carried out for mixed speech audio, speech fragment, and violin accompaniment audio, respectively. The work of feature extraction is described in the previous section, including frame segmentation, short-time Fourier transform, and logarithmic power spectrum extraction, etc. In this frame segmentation process, the length of each frame is 30 ms and the frame shift is 15 ms. Logarithmic power spectrum is 129 dimensional data. Then comes the model training process, whose input is the logarithmic power spectrum characteristics of mixed speech, speech fragment, and violin accompaniment audio. After the model training, the speech separation model based on deep neural network is obtained. After the model training, the next step is to apply the model to sound source separation, which is different from the model training stage in two aspects. One difference is that this is an unsupervised neural network forward propagation process, so there is no need to input the logarithmic power spectrum characteristics of speech clips and violin accompaniment audio. Another difference is that the output requires waveform reconstruction, that is, the separated audio component features need to be converted into a subjectively measured time domain waveform file. The process of waveform reconstruction includes exponential operation, inverse Fourier transform, and waveform superposition.

4.2. Evaluation Indicators. Based on the work of the previous part, the audio files that can be listened to by people can be obtained after the separation of audio sources of mixed speech. The next step is to evaluate the effect of speech separation, which is related to the quality of the speech separation model, and the separated voice quality directly affects the user experience, so it is necessary to accurately evaluate the voice quality; therefore, the evaluation of the effect of speech separation is a very important issue. Generally speaking, there are two evaluation methods: subjective evaluation and objective evaluation. Subjective evaluation mainly reflects people's subjective feelings on the speech effect. The specific realization process is to let the listener listen to the original speech and the separated speech, respectively, and then classify the separated speech and the original speech according to the similarity degree, and express the separation effect according to the given grade or score difference. Objective evaluation is to use mathematical methods to directly calculate the similarity between the original speech and the separated speech. It can be evaluated based on input–output speech or directly evaluated on the output speech. At present, mature, and practical objective evaluation criteria include PSQM, AD/MNB, PAMS, and PESQ.

4.3. Experimental Results and Performance Analysis. We analyzed the convergence performance of the neural network across the training curve in order to compare the model's convergence intuitively. The BP neural network's loss function value diagram is shown in Figure 4. As can be seen from Figure 4, the BP neural network can converge within 120 rounds and maintain good recognition performance in the verification set.

Figure 5 shows the training accuracy of neural network. It can be seen from Figure 5 that the error function in Figure 4 has similar results, and the neural network can achieve recognition accuracy of more than 90% within 120 training rounds.

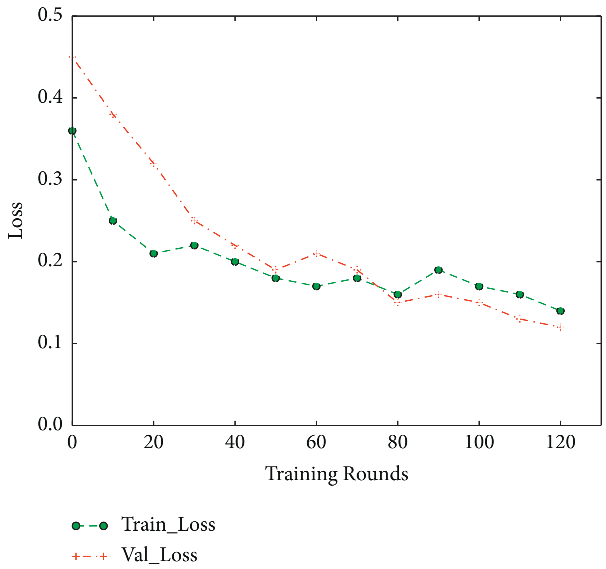

Figure 4: Neural network training loss function diagram.

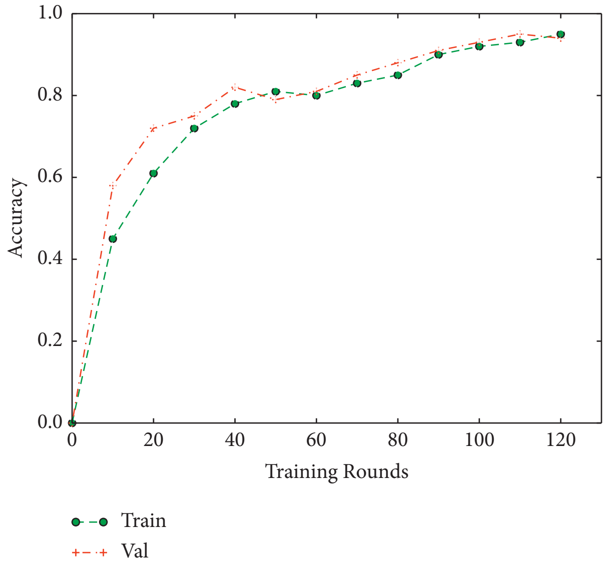

Figure 5: Neural network training accuracy diagram.

The voice and violin accompaniment mix audio on the depth is based on neural network model of hybrid model training voice source separation model, and use of the trained model has been carried out through the audio source separation, and this can be got from subjective listening to wav format audio files, and uses the perception of speech quality evaluation (PESQ) evaluated from the separation results in this paper. The following is the analysis of the experimental results:

Table 1 shows the PESQ evaluation results of voice and violin audio obtained after acoustic source separation of original mixed audio and model. Since voice and violin accompaniment are mixed according to six different SNR, the results here show the PESQ evaluation value of different

Table 1: PESQ evaluation of the three kinds of audio after speech separation.

SNR (dB)	PESQ_union	PESQ_voice	PESQ_violin
−5	1.64	2.69	1.57
0	2.17	2.94	1.59
5	2.43	3.18	1.73
10	2.72	3.40	1.45
15	3.05	3.49	1.31
20	3.26	3.81	1.54
Avg.	2.54	3.36	1.57

audios under six different SNR. As can be seen from the table, the average PESQ evaluation value of the original mixed audio is 2.54 at the six SNR, while the average PESQ evaluation value of the separated audio and violin audio is 3.36 and 1.57, respectively.

In order to more intuitively show the difference of the evaluation results of the three kinds of audio, the PESQ evaluation bar chart of the three kinds of audio are drawn here, as shown in Figure 6. As can be seen from the bar chart of PESQ evaluation of the three kinds of audio in Figure 6, for the original mixed audio and the separated audio, the PESQ evaluation value also rises in turn with the increase of SNR from −5 to 20, and the increase is large. The PESQ values of the original mixed audio and the separated audio were 3.72 and 3.69, respectively, when the SNR was 20 dB. The PESQ value of violin audio obtained after separation basically remained stable between 1.26 and 1.34, and reached the maximum value of 1.38 in 0 dB. This indicates that in the process of the SNR increasing from −5 to 20, the proportion of the voice audio in the 104 audio is increasing, so the energy of the separated voice audio is greater, so the evaluation result will be better, but the improvement of the SNR has little influence on the violin accompaniment audio.

The number of iterations in order to test the training model has an effect on the separation result. Here, 40, 50, and 60 iterations are used to train the speech source separation model. Then, the PESQ evaluation value of the voice and audio obtained after sound source separation under various SNR is shown in Table 2. As can be seen from Table 2, when the number of iterations is 40, 50, and 60 during the training model, the average value of PESQ evaluation of voice and audio obtained after audio source separation under various SNR is 3.18, 3.32, and 3.27, respectively. It can be known that the PESQ evaluation values of the three iterations are roughly similar.

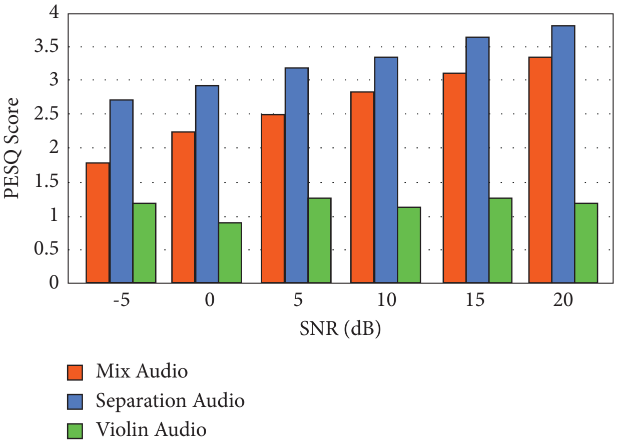

Figure 6: PESQ evaluation bar chart of the three kinds of audio after speech separation.

Table 2: PESQ evaluation of speech separation models under different training times.

SNR (dB)	PESQ_voice_40	PESQ_voice_50	PESQ_voice_60
−5	2.58	2.73	2.60
0	2.94	2.91	2.88
5	3.13	3.25	3.23
10	3.26	3.46	3.35
15	3.45	3.58	3.57
20	3.62	3.75	3.73
Avg.	3.18	3.32	3.27

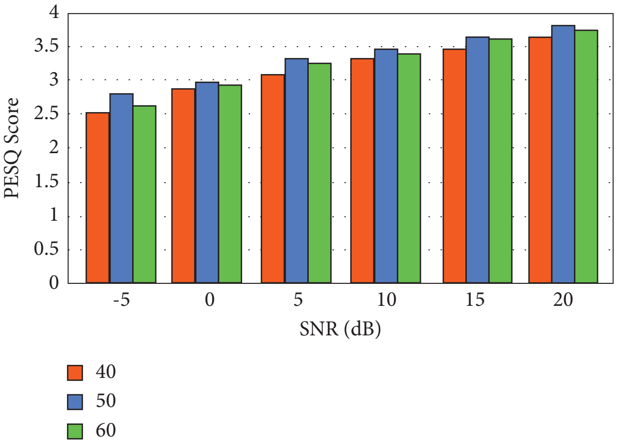

Figure 7: PESQ evaluation bar chart of speech separation model under different training times.

In order to more intuitively show the influence of different iterations on the result of audio source separation, the PESQ evaluation bar chart of the speech separation model under different training times is drawn, as shown in Figure 7. As can be seen from Figure 7, under six kinds of SNR, the PESQ evaluation value of the voice and audio with a single iteration number keeps increasing, while under the same SNR, the PESQ evaluation value of the iteration number of 40 times is significantly smaller than that of the iteration number of 50. Moreover, the PESQ rating value of 50 iterations was slightly greater than the PESQ rating value of 60 iterations. This demonstrates that the model is optimal when the number of iterations is 50. Therefore, considering the time factor, the number of iterations is 50, which is appropriate for the sound source separation model based on deep neural network.

In order to analyze the performance of sound source separation model based on deep neural network, the optimal modified least-mean-square log-amplitude spectrum estimation named L-MMSE algorithm is used for performance comparison. For mixed audio, L-MMSE algorithm also uses TIMIT database, while nonspeech part uses additive white

Table 3: PESQF evaluation of DNNF and L-MMSE.

SNR (dB)	DNN	L-MMSE
−5	2.69	1.49
0	2.91	1.84
5	3.18	2.28
10	3.37	2.65
15	3.65	3.02
20	3.74	3.35
Avg.	3.26	2.68

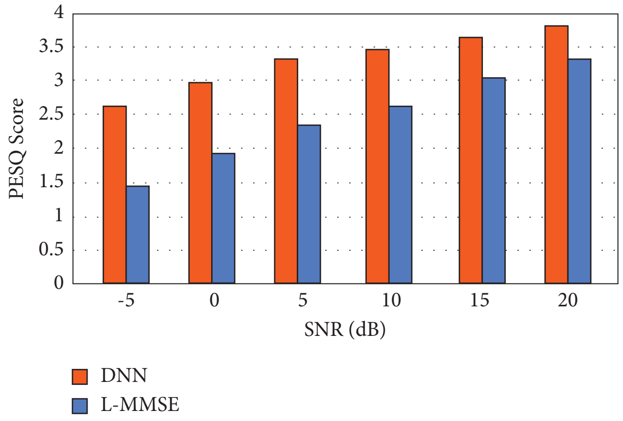

Figure 8: PESQ evaluation of DNN and L-MMSE.

Gaussian noise and other four kinds of noise. The separation results were evaluated by PESQ evaluation index, and the experimental results were shown in Table 3. As can be seen from Table 3, in the sound source separation model based on deep neural network used in this paper, the average value of PESQ evaluation of voice and audio obtained after sound source separation is 3.26, while the average value of PESQ evaluation of voice and audio obtained by L-MMSE is 2.68.

To illustrate this problem more intuitively, the PESQ evaluation bar chart of FDNN and L-MMSE is drawn here, as shown in Figure 8. As shown in Figure 8, the PESQ assessment value of speech and audio achieved by DNN is higher than that obtained by the L-MMSE technique for six different SNR. This shows that the sound source separation model based on deep neural network has better performance in speech source separation.

5. Conclusions

In the context of intelligent Internet of Things, online music teaching is becoming more and more common. However, in the field of online violin education, there is still a lack of a complete supervised learning system. Aiming at this problem, this paper chooses the mixed instrument audio composed of voice and violin as the research object. The sparsity of speech signals and multiple repetitiveness of music signals is investigated in the time domain and frequency domain of diverse audio sources. Also included are the sample pretreatment and feature extraction operations, such as training set selection, simple drying, slicing, audio overlay, feature extraction, and other procedures. Developing a deep neural network-based sound source separation model. A sound source separation model based on deep neural network is used to separate mixed speech, and the evaluation of perceptual speech quality is introduced. The PESQ results of original mixed audio, separated audio and violin audio at different SNR are analyzed. The results show that the model has a good separation effect for mixed audio. This method can provide supervision and scoring functions for violin online learning system.

Data Availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The author declares that he has no conflicts of interest.

References

[1] S. A. Gholson, “Proximal positioning: a strategy of practice in violin pedagogy,” *Journal of Research in Music Education*, vol. 46, no. 4, pp. 535–545, 1998.

[2] T. Zelig, “A direct approach to pre-school violin teaching,” *Music Educators Journal*, vol. 53, no. 5, pp. 44–46, 1967.

[3] J. M. Geringer, M. L. Allen, R. B. MacLeod, and L. Scott, “Using a prescreening rubric for all-state violin selection: influences of performance and teaching experience,” *UPDATE: Applications of Research in Music Education*, vol. 28, no. 1, pp. 41–46, 2009.

[4] A. Creech and S. Hallam, “Interpersonal interaction within the violin teaching studio: the influence of interpersonal dynamics on outcomes for teachers,” *Psychology of Music*, vol. 38, no. 4, pp. 403–421, 2010.

[5] L. Xiao, X. Wan, X. Lu, Y. Zhang, and D. Wu, “IoT security techniques based on machine learning: how do IoT devices use AI to enhance security?” *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 41–49, 2018.

[6] W. Sun, J. Liu, and Y. Yue, “AI-enhanced offloading in edge computing: when machine learning meets industrial IoT,” *IEEE Network*, vol. 33, no. 5, pp. 68–74, 2019.

[7] F. Fei Tao, Y. Ying Zuo, L. D. Li Da Xu, and Z. Lin, “IoT-based intelligent perception and access of manufacturing resource toward cloud manufacturing,” *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1547–1557, 2014.

[8] P. Patel, M. Intizar Ali, and A. Sheth, “On using the intelligent edge for IoT analytics,” *IEEE Intelligent Systems*, vol. 32, no. 5, pp. 64–69, 2017.

[9] S. Lin, B. Zheng, G. C. Alexandropoulos, M. Wen, F. Chen, and S. Mumtaz, “Adaptive transmission for reconfigurable intelligent surface-assisted OFDM wireless communications,” *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2653–2665, 2020.

[10] D. Göktürk, “Current status of string teacher education at university music teacher training schools in Turkey,” *International Journal of Music Education*, vol. 28, no. 2, pp. 176–192, 2010.

[11] Y. Ahmet Kerim ACAR, A. Ş Sakin, and A. K. Acar, “Educators’ views on online/distance violin education at Covid-19 outbreak term,” *Journal for the interdisciplinary art and education*, vol. 1, no. 1, pp. 1–19, 2020.

[12] H. Rothenberg, "The new methodical approach to violin teaching," *American String Teacher*, vol. 10, no. 2, pp. 8–10, 1960.

[13] J. Du, C. Jiang, Z. Han, H. Zhang, S. Mumtaz, and Y. Ren, "Contract mechanism and performance analysis for data transaction in mobile social networks," *IEEE Transactions on Network Science and Engineering*, vol. 6, no. 2, pp. 103–115, 2019.

[14] S. Leong, "Navigating the emerging futures in music education," *Journal of Music, Technology & Education*, no. (2-3), pp. 233–243, 2012.

[15] P. Webster, "Historical perspectives on technology and music," *Music Educators Journal*, vol. 89, no. 1, pp. 38–43, 2002.

[16] N. Souviraa-Labastie, A. Olivero, E. Vincent, and F. Bimbot, "Multi-channel audio source separation using multiple deformed references," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1775–1787, 2015.

[17] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Transactions on Audio Speech and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2013.

[18] W. Duan, J. Gu, M. Wen, G. Zhang, Y. Ji, and S. Mumtaz, "Emerging technologies for 5G-IoV networks: applications, trends and opportunities," *IEEE Network*, vol. 34, no. 5, pp. 283–289, 2020.

[19] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[20] E. Tzinis, Z. Wang, X. Jiang, and P. Smaragdis, "Compute and memory efficient universal sound source separation," *Journal of Signal Processing Systems*, vol. 94, no. 2, pp. 245–259, 2022.

[21] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio Speech and Language Processing*, vol. 14, no. 1, pp. 191–199, 2006.

[22] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[23] M. T. S. Al-Kaltakchi, R. R. O. Al-Nima, M. A. M. Abdullah, and H. N. Abdullah, "Thorough evaluation of TIMIT database speaker identification performance under noise with and without the G.712 type handset," *International Journal of Speech Technology*, vol. 22, no. 3, pp. 851–863, 2019.