Folk music has a unique musical beauty, but limited by the skill proficiency of players, quality of instruments, etc., it is difficult to guarantee a stable and high-quality experience of folk music. Sound enhancement is a forensic audio enhancement service that uses nondestructive methods to enhance audio recordings while preserving voice clarity. This is important so that the fact-finder can assess the events described in the documentary evidence. The music sound enhancement has the function of beautifying the sound color and improving the output sound quality. The principle of the sound enhancement is mainly to carry on the special reprocessing to the left and right sound channel signal, enlarge the sound field, produce the extremely realistic positioning sound effect around the listener, and bring the real auditory experience [1]. Literature [2] uses DNN sound source enhancement technology to improve high music sound effects, but the range of sound source intensity enhancement for folk music instruments is limited, and the realization is difficult. The DNN source enhancement architecture is utilized to estimate the latent parameters in the T-F mask processing output signals’ continuous probability density function (PDF). Literature [3] uses the excitation loudspeaker in the OLED panel for sound enhancement. The principle behind OLED panel speaker technology is that an OLED panel works as a speaker diaphragm by adding an oscillator called an exciter to the back of the OLED panel. With the ability to play back both the sound and the video in the same place, the technology allows listeners to experience the panel directly and clearly while also enhancing the interactive experience. This method has a good effect on audio enhancement in video, but it is not good for pure audio processing. In addition, there are certain limitations in the sound enhancement methods that use front and rear dimensional 3D enhancement, base-missing harmonic control bass enhancement, sound source localization, and sparse Bayesian learning, which cannot achieve ideal effects for the sound enhancement of folk music [4–7].

The Internet of Things is a kind of network that connects any terminal to the Internet through radio-frequency identification, wired and wireless data transmission, etc., to realize data exchange. In essence, the Internet of Things is a new omnidirectional and multiangle communication network with communication network as the carrier, and its overall architecture consists of the perception layer, the network layer, and the application layer from bottom to top [8]. It can reduce the signal interference in the signal
transmission process. Genetic algorithm has good iteration and can improve the sound enhancement [9].

The Internet of Things (IoT) allows diverse physical items to share information over a 6G wireless communication network. IoDs get access to the main spectrum by acting as an orthogonal frequency-division multiplexing (OFDM) relay using energy gathered from a radio-frequency (RF) transmission received. IoT network is powered by UAVs that uses orthogonal frequency-division multiplexing (OFDM). Two ground nodes (GNs) are supplied with electricity by two UAVs in the proposed network through a down link WPT. In order to resolve the suggested optimization issue, we use sequential convex programming (SCP) approach. Simulation findings demonstrate that our suggested design provides a higher total average transmission rate than the benchmark methods [10, 11].

The main motivation of the work is discussed as follows. This paper will study the folk music audio enhancement method combining the Internet of Things and GA. Further, the postprocessing algorithm of the wavelet packet decomposition method is combined with Wiener filter to reduce noise after obtaining folk music audio by IoT. And, the output error of the left and right sound channels is reduced by using HRTF.

2. Folk Music Audio Enhancement Method Combining Internet of Things and GA

2.1. Internet of Things Technology to Access Music Audio

In order to improve the efficiency of enhancing the sound effects of folk music, this paper uses sound sensors to collect the occurrence of each folk music playing instrument and uses wireless Internet of Things for data transmission. The sensor measures noise levels effectively at frequencies ranging from 3 kHz to 6 kHz. The human ear can detect noises in this frequency range. Acoustic pressure sensors, pressure microphones, high amplitude pressure microphones, prepolarized condenser microphones, probe microphones, condenser microphones, and prepolarized free-field condenser microphones are some of the sensors used in the sound enhancement process. Multiple audio sensor nodes with association are divided into clusters, and one sensor node within each cluster is selected as the cluster head node based on certain algorithmic mechanisms, which is used to manage and coordinate the work of its cluster member nodes, collect node information, and perform specific processing and forwarding functions such as music signal denoising [12–14]. The cluster head node in the lower level network can become a cluster member node in the higher level network and continue to cluster up to the highest level. The highest level cluster head node communicates directly with the base station [15].

Due to the different parts distribution and sound modes of different folk music instruments, there are differences in the intensity of music signals collected. In order to ensure that the final output music is in a relatively harmonious state, it is necessary to conduct intensity balance processing through the audio analyzer before the sound enhancement [16–20]. After the signal is picked up by the sound sensors on the different instruments being played, it needs to be processed by an audio analyzer, which uses a CMOS chip to divide the audio spectrum into seven bands: 63 Hz, 160 Hz, 400 Hz, 1 kHz, 2.5 kHz, 6.25 kHz, and 16 kHz. The auditory spectrum, which ranges in frequency from 20 Hz to 20,000 Hz, is the range at which humans can hear. The audio spectrum ranges from 20 Hz to 20,000 Hz and may be efficiently divided into seven distinct frequency bands, with each band having a particular effect on the overall sound.

The complementary metal-oxide semiconductor (CMOS) chip, which stores configuration and other data for the hard drive, is battery-operated. CMOS chips often offer CMOS memory and real-time clock (RTC) in microcomputers and microcontrollers. The seven frequencies detected provide a DC band amplitude for the multiplexed output. No additional components are used as a response to the selection filter. Only one of the chip’s resistors and capacitors needs to be switched off to select the clock oscillator frequency on the chip. The filter center frequency tracks the clock oscillator frequency [21, 22]. It is often described as the geometric mean or the arithmetic mean of the lower and higher cutoff frequencies of a band-passing or band-stopping system. The center frequency of a filter or channel is a measurement of the frequency that is in the middle of the upper and lower cutoff frequencies. The chip power supply provides the best performance for 5 volts. The multiplexer allows multiple inputs by resetting the control gate, but the multiplexer has only 2 pins to read out. The read speed of the multiplexer is also output by the decay time control, which decays by about 10% per read. The music signal after the initial equalization processing by the audio analyzer is transmitted to the processing center through the Internet of Things for the actual ethnic sound enhancement processing [23–25].

The real-time message transmission protocol (RTMP) is used to transmit music signal, so as to improve the timeliness of music sound effect enhancement. RTMP is an application layer protocol based on TCP. It is designed for real-time communication, mainly for audio communication between Flash/AIR platform and server that supports RTMP. Internally, ActionScript 3.0 is the main programming language used by AIR, and it shares a codebase with the rendering engine of Flash Player. A TCP-based technology called RTMP provides for low-latency communication while still maintaining persistent connections. It divides streams into segments, and their sizes are constantly negotiated between the client and server in order to deliver streams smoothly and transfer as much information as feasible. RTMP meets the real-time requirements of this design, and the open source RTMDDump toolkit can be used to reduce the development effort [23]. RTMDDump is a free software project aimed at the development of an RTMP toolkit. RTMDDump, Rtmppsv, and Rtmpsock are the three applications included in the package. Similar to regular Flash video player clients, RTMDDump can connect to RTMP servers, capture the network stream, and save it to a file. It allows for the construction of instructions utilizing connection and authentication details previously collected by rtmppsv from the RTMP server.

To stream music over the RTMP, there are four steps: shake hands, establish a connection, establish a stream, and
send. After the handshake is successful, it is necessary to realize the network connection of the Internet of Things in the establishment of connection stage, then complete the construction of network flow between the C/S structures of the Internet of Things in the establishment of stream stage, and complete the transmission of audio data in the transmission stage. In the handshake stage, the server and the music uploader need to send three fixed-size data blocks, respectively [26–30]. The brief process is as follows [31].

1. At the beginning of the handshake, the upload end sends C0 and C1 data blocks, and the server end sends S0 and S1 data blocks to the upper transmission end after receiving C0 or C1.
2. After the uploader and server receive S0 and S1, C0 and C1, respectively, they send C2 and S2 to the opposite end.
3. When C2 and S2 are received, the handshake process is completed. The RTMP handshake process is shown in Figure 1 [32].

The process of establishing the connection phase is as follows:

(1) The uploader sends the “Create Stream” message to the server.
(2) After receiving the command “Create Stream,” the server sends the “result” message to the top end to inform the state of the stream.

Sending process:

(1) The upload side transmits the “Send” command to the server side.
(2) The server sends the message of setting block size after receiving the command.
(3) The server sends the “stream start” control message to the upstream end, including the music stream ID.
(4) After the successful execution of the command, the server sends the “response status” message to the upload end to inform the execution status of the command at the upload end.

After using the Internet of Things to obtain folk music signals, in order to achieve better music sound effects, it is necessary to denoise the music signals to improve the quality of audio signals.

2.2. Music Noise Reduction Processing. Signal collected by sound sensors in the Internet of Things will contain unavoidable noise during performance. In this paper, the postprocessing algorithm of wavelet packet decomposition combined with Wiener filtering is used to suppress noise in music. The Wiener filter may be used to remove the noise from a distorted signal and estimate the relevant underlying signal. The minimal mean square error (MMSE) estimator article provides a more statistical explanation of the idea as the Wiener filter is based on a statistical methodology. The Wiener filtering achieves the best trade-off between noise smoothing and inverse filtering. At the same time, it reverses the blurring and eliminates the additional noise. The Wiener filtering has the best mean square error performance.

The first step of wavelet packet decomposition is to select the appropriate wavelet packet base. The choice of wavelet packet basis is not unique; as long as the function satisfies the wavelet packet condition in the wavelet packet basis, the appropriate wavelet packet basis is very important for the decomposition of the wavelet packet. The principle of selecting the wavelet packet basis is that the higher the coincidence degree between the wavelet packet basis and the useful signal, the better. Since the time-domain medium and small wave packet bases and signals are convolved, if the waveform of wavelet packet bases and useful signals are close, the features of useful signals will also be amplified, while those with different shapes will be suppressed, which is conducive to signal denoising. Noise is generally nonstationary, with time-varying and mutagenicity. It is proved that the wavelet packet basis of Daubechies (DB) series is a compactly supported orthogonal wavelet packet basis, which is suitable for processing abrupt signals. Daubechies wavelets are frequently employed to address a variety of issues, including the self-similarity characteristics of a signal, fractal issues, and signal discontinuities. Therefore, the wavelet...
packet base of Daubechies series is selected in this study. Using the DB2-DBLL wavelet packet base of DB waveform, the music signal with noise is decomposed into four layers of wavelet packet for denoising [33]. Using mean square error and signal-to-noise ratio to judge, it is found that DB4 is the best of 10 wavelet packet basis, so DB4 in Daubechies series is selected as the wavelet packet basis, and the number of decomposes layers in the wavelet packet decomposition of suppressing music noise is set as 3. For a given signal \( F(x) \), its orthogonal wavelet decomposition formula is as follows [34]:

\[
M_{j-1}F(x) = M_jF(x) + N_jF(x),
\]

\[
M_jF(x) = \sum_n p_n^j \cdot \varphi_{j,n}(x),
\]

\[
N_jF(x) = \sum_n q_n^j \cdot \psi_{j,n}(x).
\] (1)

In the above equation, \( M \) and \( N \) are the frequency bands of the approximate part and the detail part in the multi-resolution analysis, respectively. \( p_n^j \) and \( q_n^j \) are coefficients, which can be determined, recursively.

The wavelet packet reconstruction formula for the music signal as follows [35]:

\[
t_j^{1,n} = \sum_k \left[ h_{l-2k} q_k^{j+1,2n} + g_{l-2k} q_k^{j+1,2n+1} \right].
\] (2)

In the above equation, \( t_j^{1,n} \) is the signal reconstructed by wavelet packet; \( h_{l-2k} \) and \( g_{l-2k} \) are the filter coefficients. Mallat algorithm can implement wavelet packet decomposition and wavelet packet reconstruction as filter, low-pass filter is equivalent to scale function, and high-pass filter is equivalent to wavelet function. The Mallat algorithm inverts a single-scale orthogonal expansion into a multiscale orthonormal expansion. The generalization of the coefficients is not necessary for the Mallat method for some function’s Fourier coefficients. Wavelet packet decomposition filter is to get the first layer coefficient of wavelet packet decomposition by low-pass filtering and high-pass filtering, respectively, and the next layer coefficient is obtained by low-pass filtering, high-pass filtering, and low sampling of the upper layer coefficient until the set number of decomposition layers. The wavelet packet reconstruction filter realizes the inverse process equivalent to the wavelet packet decomposition filter; that is, the coefficients of the lowest layer are sampled up and then summed up through the corresponding low-pass reconstruction filter and high-pass reconstruction filter to get the coefficients of the upper layer until the original signal of the uppermost layer is reconstructed.

The prior signal-to-noise ratio of the current frame audio signal can be obtained from the prior signal-to-noise ratio of the previous frame and the posterior signal-to-noise ratio of the current frame. After obtaining the prior signal-to-noise ratio \( \xi(k) \) of the voice signal of the current frame, the following equation can be used to obtain Wiener filter’s transfer function for the audio signal in the current frame [36]:

\[
S(k) = \frac{\bar{\xi}(k)}{1 + \bar{\xi}(k)} = 1 - \frac{1}{\gamma(k)}
\] (3)

In the above formula, \( \xi(k) \) and \( \gamma(k) \) represent prior SNR and posterior SNR, respectively. The ratio between the clean signal and the noisy signal's power is known as the a priori SNR, whereas the posteriori SNR is defined as the ratio of the observed noisy signal’s squared magnitude to the noise power. Both SNRs are calculated for each frequency bin. The transfer function of Wiener filter is as follows:

\[
S_i(k) = \frac{\hat{\xi}_i(k)}{\hat{\xi}_i(k) + 1}.
\] (4)

The output of Wiener filter can be calculated as follows:

\[
B(k) = S_i(k)A_i(k).
\] (5)

After denoising the audio signals transmitted and collected by the Internet of Things in accordance with the above process, GA is used to enhance the sound effects of folk music.

2.3. GA to Enhance Music Sound Effects. The enhancement processing of folk music sound effect is mainly realized according to the principle of spatial hearing characteristic of human ear. Therefore, this design will use genetic algorithm to improve the HRTF, so as to achieve sound efficiency. The head related transfer function (HRTF) defines how an ear receives sound from a sound source. A sound wave is created when a sound is generated and travels across space in all directions. The sound wave emitted by the sound source reaches the ears after scattering and reflection through the head, auricle, trunk, etc., which changes the sound and enables the human ear to locate it. This physical process can be regarded as a linear time-invariant (LTI) filtering system, whose characteristics can be fully described by the frequency domain transmission function of the system. HRTF is the frequency domain transfer function of the acoustic filtering system. In the case of free field, HRTF is defined as [37]:

\[
H_L = H_L(r, \theta, \varphi, \omega, a) = \frac{P_L(r, \theta, \varphi, \omega, a)}{P_0(r, \omega)},
\]

\[
H_R = H_L(r, \theta, \varphi, \omega, a) = \frac{P_R(r, \theta, \varphi, \omega, a)}{P_0(r, \omega)}.
\] (6)

In the above formula, \( P_L \) and \( P_R \) are the complex sound pressures generated by the sound source in the listener’s left and right ears, respectively. \( P_0 \) is the complex sound pressure at the center of the head when the head does not exist. \( H_L \) and \( H_R \) are functions of the horizontal azimuth of the sound source \( \theta \), elevation \( \varphi \), the distance from the center of the sound source to the head \( r \), and the angular frequency of the sound wave \( \omega \). In addition, the size and shape of the head and auricle and trunk of different people are different; so strictly speaking, the HRTF of each person is different; that is to say, the HRTF is a physical quantity with personalized characteristics. HRTFs are usually measured in an anechoic room to reduce the impact of early reflections and
reverberation on the recorded response. Interpolation is used to generate HRTFs at arbitrary positions of. HRTFs are recorded at modest increments, such as 15° or 30° in the horizontal plane. In the formula, \( a \) represents the parameter with personalized characteristics.

The morphology of spectrum peaks or valleys in the HRTF amplitude spectrum can be described by three parameters, namely, the center frequency, the height of the peak or the depth of the valley, and the width of the peak or valley. These three parameters are used to uniquely determine a filter, so a second-order notch filter can be used to control the synthesis of a single spectrum peak or spectrum valley, which can better approximate the spectrum peak and spectrum valley actually measured. The shape positions of spectrum peaks and valleys vary significantly between individuals, which is the main source of errors in the use of nonpersonalized HRTF for sound enhancement synthesis. Using genetic algorithm to improve the HRTF can reduce the error of sound enhancement synthesis.

In this study, multiple continuous variables of HRTF are encoded by genetic encoding in the form of real encoding, and \( N \) dimensional real vectors are used to represent \( N \) real variables. The most classic method of selecting operations is roulette selection. It uses the proportion of fitness to determine the size and proportion of the central angle of the fan on the roulette. If \( N \) individuals are selected as parents, the roulette wheel needs to be rotated \( N \) times, which is equivalent to doing \( N \) times of uniform probability distribution sampling. Using small selection pressure in the early stages of genetic search makes the search area wider and directs the search location to the promising neighborhood. In the late stage of search, high selection pressure is applied to accelerate the convergence of the algorithm to the optimal solution of the current covered neighborhood [38, 39]. According to the HRTF optimized by genetic algorithm, the sound effect enhancement of folk music can be realized by controlling the intensity difference between the left and right sound channels during the audio output. At this point, the method combining the Internet of Things and GA is realized to enhance the sound effect of folk music and improve the auditory experience of folk music. A genetic algorithm is a type of search-based algorithm used in machine learning to solve optimization issues. This method is significant because it quickly resolves challenging issues that would otherwise require a lot of effort.

3. Sound Enhancement Method Testing

The sound enhancement method of folk music combining IT and GA is proposed above. According to the requirement of the sound display effect when folk music is performed, this section will test the studied sound enhancement method.

3.1. Sound Effect Enhancement Test Content. The actual effect of the method is tested from two aspects: the spatial sense of music and the anticrosstalk effect of the music signal after the processing of the ethnic music sound enhancement method. The term “spatial music” refers to music in which the listener’s attention is drawn primarily to the listener’s location and movement in relation to the sound sources’ positions and movements. It could include a single mobile sound source or several simultaneous fixed or mobile sound events occurring in various places. The easiest technique to avoid crosstalk is to take advantage of the parallelism that causes it by tightly connecting the return path to the ground to your high-speed signals. Because the return path is same in amplitude but opposite in direction, the fields cancel out and decrease crosstalk. In order to make the test results credible, the sound enhancement method based on is chosen as a comparative reference. Through the form of comparative test, the feasibility of the sound enhancement method proposed in this paper is visually verified.

3.2. Test Preparation and Process. Five students from a music school majoring in folk instruments are invited to participate in this test, and all students will play famous music excerpts from the corresponding folk instruments. The folk music played by the students on different instruments will be used as the test data for this experiment to test the actual effect of the sound enhancement method. In order to avoid interference from the outside environment and equipment, the equipment will be tested and noise interference from the outside environment will be eliminated as much as possible before the recording of the students’ instrumental pieces. 5 students will play the pieces as required, and pickups will be attached to the vocal positions of the playing instruments to capture the instrumental pieces. After the folk music was performed, the clips were intercepted using music editing software. To facilitate the subsequent analysis of the test data, the 30 s length of the music was intercepted from the more obvious segments of the music waveform and numbered. All the music for the test was only intercepted in length, without any other processing operations.

The experimental group for comparison test is the sound enhancement method designed in this paper, and the sound enhancement method based on sparse Bayesian learning is used as the reference group. The two groups of sound enhancement methods were used to enhance the sound of the input music clips. The music output after the two groups of sound enhancement methods is played back using the audio, and the output signals of the left and right channels of the audio are collected using a sound acquisition device. On the premise of ensuring the clarity of the audio output music, the reverberation waveform of the enhanced music signal is collected, and the reverberation delay and reverberation time of the output music signal and other parameters are counted.

When conducting the anticrosstalk test, the input signal of one channel is used to test the output signal of the other channel separately. By comparing the magnitude of the crosstalk rejection of the two channels, the anticrosstalk ability of the enhancement method is judged. In accordance with the above design of the sound enhancement method of comparison test content, the sound enhancement method to test summarizes all test data analysis content to obtain the final experimental results.
3.3. Experimental Results. The reverberation test after sound enhancement using actual ethnic music performance clips is shown in Figure 2, the upper part of the signal is the audio reverberation after the experimental group method, and the lower part is the audio reverberation after the reference group method. Each unit of the horizontal coordinate in the figure is 1.25 s. Reverberation is the continuation of sound after the source of the sound has ended. It is the consequence of several reflected waves that the brain can interpret as a continuous sound. On the other hand, an echo happens when a sound pulse is audible twice. Reverberation effects are frequently employed in studios to give sounds depth. The pitch of a sound is unaffected by reverberation, but the apparent spectral structure is changed. The sound bouncing in a large speaker is an illustration of reverberation.

Comparative analysis of the upper and lower parts of the signal in Figure 2 shows that, for the audio signal processed by the experimental group method, when the reverberation is established after 65 ms of the end of the input performance clip, the entire reflection sound still exists until about 5 s or even 6.5 s, thus achieving the most significant extension of the reverberation time, to ensure the purpose of the reflection density. Considering the reference group method of processing the audio signal in the end of the performance clip 74 ms after the establishment of reverberation, the reflection sound only lasted until about 3 s, and the reflection density is also rapidly declining and cannot achieve a better sound enhancement effect.

Figure 3 shows the application of two sound enhancement methods and the output speaker crosstalk suppression response curve of the left and right channels. The horizontal axis of the figure is the frequency of the speaker output signal, while the vertical axis indicates the response curve signal amplitude. The greater the amplitude of the channel crosstalk suppression response curve, the worse the channel crosstalk suppression effect and the more violent fluctuations in the
channel crosstalk suppression response curve, representing the more serious channel crosstalk problem.

Separate analysis of Figure 3, the application of the experimental group sound enhancement method after processing, the left and right channels, respectively, on the other channel of the maximum amplitude of interference is only 45 dB; and the reference group method after processing, the left channel on the right channel of the sound channel crosstalk suppression effect is slightly better than the right channel on the left channel of the sound channel crosstalk suppression effect, but the application of the method, the left and right channels respectively on the other channel of the crosstalk effect is poor.

Summarizing the above experimental data analysis, the ethnic music sound enhancement method combining IoT and GA proposed in this paper can most significantly extend the reverberation time and has the advantages of improving the output spatial sense and optimizing the anticrosstalk effect. The ratio of each voice part of the processed audio sound is more coordinated and even, and the overall output sound is remarkable and the actual experience is better.

4. Conclusion

The application of sound effect enhancement methods is promising, but constrained by the economic cost and processing capacity, there are currently common problems of poor output and poor user experience. To this end, this paper proposes an ethnic music sound enhancement method combining IoT and GA. In the future, in-depth research will be conducted on further improving the processing efficiency and sound field detail adjustment of sound enhancement.

Abbreviations:

HRTF: Head related transfer function
GA: Genetic algorithm
DNN: Deep neural network
OLED: Organic light-emitting diodes
CMOS: Complementary metal-oxide semiconductor
RTC: Real-time clock
DC: Direct current
RTMP: Real-time messaging protocol
TCP: Transmission control protocol
C/S structures: Client/server structure
ID: Identification
DB: Daubechies
SNR: Signal-to-noise ratio
LTI: Linear time-invariant
IoT: Internet of Things.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Shanxi Tourism and Performing Arts Industry Research Project, a philosophical and social science planning project in Shanxi Province (2020YY285).

References


