Percussion Big Data Mining and Modeling Method Based on Deep Neural Network Model

Xi Song
Department of Music, College of Arts, Xiamen University, Xiamen 361005, China

***Abstract*—In order to improve the analysis effector percussion waveform, this paper studies the percussion big data mining and modeling method based on the deep neural network model. Aiming at the problem of the high sampling rate of Analog to Digital Converter (ADC) when the wideband frequency-hopping Linear Frequency Modulation (LFM) percussion waveform is sampled by Nyquist, this paper proposes a method of under sampling, and conducts a simple theoretical analysis. When the signal-to-noise ratio is 35dB, the frequency measurement error is close to 1MHz, which can meet the requirements of frequency measurement accuracy. When the signal-to-noise ratio is higher than 35dB, the frequency measurement error gradually decreases and eventually stabilizes, with a frequency measurement accuracy of around 30 kHz. Due to the low environmental interference in the sound wave recognition of percussion instruments and the close distance between the hardware equipment and the percussion instruments in this paper, the recognition results of the model in this paper have high accuracy Compared with existing methods, this article is more reliable in identifying percussion sound waves. From the data, it can be seen that the method proposed in this article has better performance in waveform recognition in impact big data mining models.**

Keywords*—Deep neural network; percussion; big data; mining; modeling

I. Introduction

All musical instruments generate and propagate sound waves. Sound waves can be simulated and form echoes through frequency hopping signals. Therefore, to extract effective information from instrument performances, one can start with frequency hopping signals and propose signal processing methods that can be applied to instrument performance information data mining. This article takes big data mining of percussion as an example for research, first analyzing the relevant research on the performance characteristics of percussion instruments.

In many percussion instruments, the same timbre can be played in different hitting positions, such as bell rings, bass drum bangs, and so on. When the player hits these sounds, he usually chooses a relatively convenient hitting position to complete the performance according to the preceding and following phrases among the many hitting positions. In some percussion works, by carefully arranging the striking position, the body shape can be changed to achieve the purpose of displaying the musical image [1].

There is relatively little research on extracting effective information from instrument performance, so this article proposes an effective music information data mining method based on practical needs. This paper studies and improves the percussion big data mining and modeling method based on the deep neural network model combined with the robot simulation technology, explores the research effect of percussion, and effectively improves the performance of percussion [2].

A method of undersampling is proposed to address the issue of high ADC sampling rate during Nyquist sampling of broadband frequency hopping LFM percussion waveforms, and a simple theoretical analysis is conducted. At the same time, for the frequency ambiguity caused by undersampling, two commonly used frequency deblurring methods, the Chinese remainder theorem and time-frequency analysis, were introduced, and the implementation complexity of these two methods was analyzed. A method based on multi-channel frequency partition decomposition blurring was proposed [3].

When performing percussion works, in certain sections with complex rhythmic changes or compound beats, the performer will subconsciously use their head, torso, and other limbs to strike the rhythm, prompting the audience to follow the logic of rhythm division. When fingers strike a paragraph in music, the shape of the fingers can guide the audience's understanding of the musical phrase. The performer will design the finger shape during or after striking while ensuring the timbre. While ensuring the beauty of the striking form, integrate it with the underlying emotions of the music. The innovation of this article lies in proposing a deblurring method based on multi-channel frequency division, which greatly reduces the implementation complexity. Finally, the fast frequency measurement is achieved through linear interpolation zero crossing frequency measurement method, improving the extraction effect of vocal waveform

Section I of this article mainly introduces the background and current situation, leading to the research content of this article. The following is the relevant work section, which mainly summarizes the existing research work in Section II, raises the existing research problems, and proposes improvement strategies for this article. Section III is the algorithm model section, which proposes the improved algorithm and model of this article, conducts experimental research is presented in Section IV, and finally summarizes the research content of this article in Section V.

II. Related Work

Non-deep learning algorithms consider audio characteristics and search for different feature representations of accompaniment and singing in songs, separating

*Corresponding Author

accompaniment and singing. This type of method relies on long-term accumulated audio knowledge to identify differences between the two, but typically finds distinguishable features with long cycles, high difficulty, and may not be universally applicable to all types of songs. Non-deep learning separation techniques mainly include matrix factorization and acoustic features. Non-negative matrix factorization (NMF) and robust principal component analysis (RPCA) are two typical matrix factorization methods used for vocal separation. From the perspective of acoustic features, propose a method for calculating auditory scene analysis based on pitch inference and accompaniment repetition. Due to the involvement of multiple disciplines in the fields of audio and computer science, the accompaniment and vocal frequency spectra in songs are intertwined and intertwined. Currently, non-deep learning algorithms for vocal separation have made some progress, but there are still problems with mixed vocal/accompaniment and low separation quality [4]. Deep learning algorithms mainly use deep, high semantic, and highly distinguishable features automatically learned by neural networks to separate and predict the time-frequency spectrum of accompaniment/singing, and finally reconstruct the accompaniment and singing signals. This type of algorithm mainly relies on the selection of neural networks. Suitable neural networks can learn and capture features that distinguish between the two, thereby predicting time-frequency spectra that are closer to the original accompaniment/singing. Deep learning algorithms include two categories: modeling in the frequency domain and modeling in the time domain [5]. Reference [6] focuses on deep learning algorithms and therefore provides a detailed introduction to the frequency domain and time domain models of deep learning. Frequency domain model: Due to the significant performance of neural networks on images and the fact that the frequency domain has more exploitable information compared to the time domain, existing algorithms focus on modeling time-frequency spectra in the frequency domain, known as frequency domain models. The main idea is to transform the song from time domain to frequency domain through short-time Fourier transform, input the time-frequency spectrum of the song into a neural network, and the network predicts the time-frequency spectrum of the accompaniment and singing voice. Finally, the phase approximation of the original song is used instead of the accompaniment and singing phase, and the time-frequency spectra of the accompaniment and singing are combined with the original song phase spectrum to reconstruct the time-domain signals of the accompaniment and singing. The separation performance of frequency domain models depends on the selection of neural networks. A network structure with rich structure and the ability to capture and learn comprehensive features can predict high-precision accompaniment/singing time-frequency spectra. In the reconstruction phase, the frequency domain model approximates the separated signal phase using the original song phase, without modeling the phase, which is currently a factor that restricts the quality of separation [7].

Time domain model: Modeling in the time domain refers to using time-domain signals as input and directly putting them into a neural network for training. The network outputs separated time-domain signals of accompaniment and singing. Directly modeling in the time domain avoids the problem of phase distortion in the frequency domain model. The study in [8] attempted to model in the time domain and achieved good separation results. However, due to the high sampling rate of audio signals, the one-dimensional signal in the time domain is very large, resulting in excessive input to the neural network. Whether the network can adapt to the large input size and learn abstract features such as time and space reasonably is a challenge to the network separation performance. Therefore, there is still some research and exploration space for the time domain model.

The main separation idea of the frequency domain model is to use the time-frequency spectrum after short-time Fourier transform as the network input, utilize the advantage of neural network automatic feature learning, capture high semantic features that can distinguish accompaniment and singing, and predict the mask matrix (composed of numbers between 0 and 1) of accompaniment and singing signals. Then, based on the original song time-frequency spectrum and the predicted mask, the time-frequency spectrum of accompaniment/singing is obtained. Finally, by combining the phase reconstruction of the original song, the time-domain signals of the accompaniment and singing voice are obtained [9]. The quality of frequency domain model separation depends on the accuracy of the time-frequency spectrum predicted by the network, and the network structure and learned features determine the quality of separation [10].

At present, the neural networks used in frequency domain models have transitioned from basic neural networks (such as RNN, LSTM, CNN) to structurally rich and multi-level neural networks (such as U-Net, SH-4Stack). With the continuous enrichment and diversity of network structures, the learned features have also been continuously improved. However, the common feature of advanced neural networks used for monaural vocal separation today is that the network structure is serial, and after multiple downsampling, some information will be lost. Moreover, upsampling cannot restore the original information feature appearance, and the defects in the feature learning process result in low amplitude accuracy of the predicted time-frequency spectrum [11].

Music originates from rhythm, and rhythm is also the most basic element of music. When our ancestors in ancient times, based on the relationship between the heart and the pulse, rhythm instinctively evolved into a form of music. Rhythm is more important to music today than ever before. In the use of music, rhythm is more important than melody, harmony, and pitch. Rhythm without specific pitch can make the listener understand the content, but pitch without rhythm can only be called accent [12]. Rhythm is very important in musical elements, and accent is irreplaceable in the rhythm system. After the accent is played well, it will produce the corresponding rhythm. The accent is actually the power generated by the rhythm, and the rhythm is the vitality of the rhythm. Simply playing the rhythm without the change of accent, even if the music played is correct, it cannot make the listener dance with the music. The playing of the accent produces the rhythm, and the existence of the rhythm makes the rhythm have vitality. If the rhythm has life, the music will create a magical power for the listener to enjoy it [13].

With the improvement of productivity and manufacturing level, the development of music goes hand in hand with it. Under the fierce market competition, many professional musical instrument craftsmen have created a guild system while working hard to produce excellent works. The appearance of guilds is to protect the interests of fellow handicraftsmen from being infringed by outsiders, in order to prevent the competition of foreign handicraftsmen and limit the competition between local handicraftsmen in the same industry, a civil organization established by urban handicraftsmen [14]. Guilds have both positive and negative effects. The various regulations issued by the guild have improved the production level of the musical instrument manufacturing industry to a certain extent, but also restricted free competition, the number of employees, the mass production of commodities, and the application of new production tools [15]. The various rules of the guild also make the shapes of musical instruments appear to be similar. The same type of musical instrument, although made by different craftsmen, has almost the same dimensions. In order to meet the market demand of the music industry, instrument manufacturers need more manpower for expanded reproduction [16]. Due to the high difficulty of processing musical instruments, many complex processing procedures still require manual operations and the skilled craftsmen of the processors. Therefore, the way for many musical instrument craftsmen to expand reproduction is not the training system, but the apprenticeship system [17]. Many apprentices need to practice in the workshop for several years, and then take over the mantle of the master and continue to make musical instruments. They don't have time to practice their musical instruments, and they have little experience in musical performances. They know the structure and workmanship of musical instruments well, but they don't understand music. Their duty is to produce instruments of the same level as the Master, pursuing more exquisite craftsmanship and production methods, rather than surpassing or innovating. It is precisely out of respect for the guild system, respect for traditions and a strong sense of responsibility for inheritance that many craftsmen have created the phenomenon of “inheritance” that is unique to musical instruments and is difficult to break [18].

Previous studies have shown that measuring the frequency of music signals in music data mining can cause spectrum aliasing, leading to frequency ambiguity. Therefore, it is necessary to deblur the sampled signals in order to obtain the true frequency of the signals. The core of frequency measurement methods under undersampling conditions is frequency deblurring, which involves undersampling broadband analog signals to obtain digital signals, and then using deblurring algorithms to recover the frequency of the digital signals to obtain the frequency of the original signal. The commonly used deblurring algorithms are the Chinese remainder theorem, time-frequency analysis, and compressive sensing. This article proposes a new deblurring algorithm based on multi-channel frequency band division. On this basis, the method of linear interpolation zero crossing frequency measurement is used to achieve fast frequency measurement of broadband frequency hopping signals. This method greatly reduces the system complexity while reducing the ADC sampling rate, and does not introduce additional deblurring errors. Finally, fast frequency measurement was achieved through linear interpolation zero crossing frequency measurement method.

III. Research Method

Percussion instruments have various forms of performance, and tapping with different parts can also emit audio signals with different characteristics. Feature mining can promote the development of smart music and is of great significance in helping performers discover deficiencies in performance in a timely manner.

Due to the fact that the Nyquist sampling theorem cannot be satisfied when using time-domain undersampling technology to sample the measured acoustic signal, using classical frequency estimation methods at this time will result in spectral aliasing. For undersampled sample sequences, in order to obtain their frequency estimation without ambiguity, a feasible algorithm needs to be used to perform frequency deblurring on the undersampled sequence. Usually, methods such as the Chinese remainder theorem, time-frequency analysis, and compressive sensing can be used to de fuzzify the frequency of the test signal. These methods can indeed achieve good results in their respective application fields, but they have high computational complexity and cannot be used as a universal method for de fuzzifying broadband frequency hopping signals under undersampling conditions. Therefore, it is necessary to propose an undersampling frequency deblurring method with low computational complexity and suitable for broadband frequency hopping signals.

The model in this article collects percussion signals, so in the actual collection process, the terminal hardware device will be connected to the collection device. The device that collects sound waves is very close to the percussion, and the volume and tone of the percussion sound are relatively high, which can be accurately collected by the terminal device. Therefore, the channel loss in the collection of sound channel signals can be ignored.

According to the Nyquist sampling theorem, the sampling rate of the ADC should be at least twice or greater than the Nyquist sampling rate.

A. The basic Theory of Under Sampling

Under sampling is defined as digitizing percussion waveforms at a sampling frequency lower than the Nyquist sampling rate. The following under sampling analysis is carried out through the single carrier frequency percussion waveform.

The input tone percussion waveform can be expressed as in [19]:

$$x(t) = \sin\left(\omega_0 t + \varphi\right) \tag{1}$$

Among them, ω_0 is the real frequency of the single-carrier percussion waveform, and φ is the initial phase of the single-carrier percussion waveform. According to the Fourier transform formula, it can be known that its spectrum is:

$$X(\omega)=\int_{-\infty}^{\infty} sin(\omega_0 t+\varphi)e^{-j\omega t}dt$$
$$=\pi\delta(\omega_0-\omega)e^{-j\varphi}+\pi\delta(\omega_0+\omega)e^{-j\varphi} \quad (2)$$

According to the relevant theory of digital percussion waveform processing, it can be known that the time domain sampling will cause the periodic extension of the spectrum, and the spectrum $X_s(\omega)$ of the percussion waveform after sampling and the spectrum $X(\omega)$ of the percussion waveform before sampling satisfy:

$$X_s(\omega)=\frac{1}{T_s}\sum_{n=-\infty}^{+\infty}X(\omega-n\omega_s) \quad (3)$$

Among them, ω_s is the sampling frequency, T_s is the sampling period. Therefore, when the percussion waveform x(t) is sampled at a fixed sampling rate Ω_s, the spectrum of the digital percussion waveform after sampling is [20]:

$$X_s(\omega)=\frac{1}{T_s}\sum_{n=-\infty}^{+\infty}\left[\pi\delta(\omega_0-\omega-n\Omega_s)e^{-j\varphi}+\pi\delta(\omega_0+\omega-n\Omega_s)e^{-j\varphi}\right] \quad (4)$$

It can be seen from the above formula that under the condition of under sampling, the real angular frequency ω_0 of the percussion waveform can be obtained by calculating according to the fuzzy angular frequency ω measured by the spectrum of the percussion waveform, the sampling frequency Ω_s and the number of ambiguities n relative to the sampling frequency. Therefore, the real percussion waveform frequency under under-sampling condition is obtained. The expression is as follows:

$$\omega_0=n\Omega_s+\omega, f_0=nf_s+f \quad (5)$$

Among them, f is the fuzzy frequency measured according to the percussion waveform spectrum, f_s is the sampling frequency of the percussion waveform, and f_0 is the real percussion waveform frequency. The sampling rate of the ADC must not be less than the Nyquist sampling rate.

B. Ambiguous Understanding of Chinese Remainder Theorem

The Chinese remainder theorem, as an outstanding achievement in ancient Chinese mathematics, embodies the wisdom of our ancestors and has made significant contributions in many modern research fields. This section will introduce the algorithm principle of the Chinese remainder theorem and further expand it. Simultaneously utilizing the Chinese remainder theorem for frequency analysis to achieve the goal of resolving ambiguity

The Chinese remainder theorem is an important theorem in number theory. Its content can be described as: $m_1,m_2,\mathrm{L}\ ,m_L$ is assumed as a positive integer that is relatively prime, and is defined as:

$$M_i=m/m_i, 1\le i\le L \quad (6)$$

Among them, there is $m=m_1m_2\mathrm{L}\ m_L$, then for any integer $r_1,r_2,\mathrm{L}\ ,r_L$, the following first-order congruential equations must have a solution [21],

$$\begin{cases} X\equiv r_1\ mod\ m_1 \\ X\equiv r_2\ mod\ m_2 \\ \quad \mathrm{M} \\ X\equiv r_L\ mod\ m_L \end{cases} \quad (7)$$

Furthermore, the solution to the system of equations is

$$X=\sum_{i=1}^{L}\overline{M_i}M_i r_i\ mod\ m \quad (8)$$

Among them, $\overline{M_i}$ is the inverse of M_i to modulo m_i, and it satisfies the following relation:

$$\overline{M_i}M_i\equiv 1 mod\ m_i, 1\le i\le L \quad (9)$$

If a and b are assumed to be given arbitrary positive integers, they can be decomposed into the form of division with remainder, which is expressed as follows [22]:

$$a=bq_1+r_1, 0<r_1<b$$
$$b=r_1q_2+r_2, 0<r_2<r_1$$
$$\cdots$$
$$r_{n-2}=r_{n-1}q_n+r_n, 0<r_n<r_{n-1}$$
$$r_{n-1}=r_nq_{n+1}+r_{n+1}, r_{n+1}=0 \quad (10)$$

Among them, $q_1,q_2,\mathrm{L}\ q_n,q_{n+1}$ and $r_1,r_2,\mathrm{L}\ r_n,r_{n+1}$ are arbitrary integers obtained. Because every division with remainder will reduce the remainder by at least 1, and b is a finite positive integer, in order to obtain an equation with a remainder of 0, at most b divisions with remainder can be performed. At this time, there is $r_{n+1}=0$. According to the Euclidean algorithm, the greatest common divisor of a and b is the last remainder that is not 0 in Eq. (10), that is r_n. Therefore, the following expression can be obtained [23]:

$$gcd(a,b)=r_n \quad (11)$$

In Eq. (11), $gcd(\cdot)$ represents the greatest common divisor. In the process of solving the greatest common divisor, the coefficients generated by the solution are collected by extending the Euclidean algorithm. Then, after backward operation, the integers x and y can be found to satisfy the following equation:

$$ax+by=gcd(a,b) \tag{12}$$

According to Eq. (6), it is easy to know that M_i and m_i are relatively prime, that is $gcd\left(M_i,m_i\right)=1$. According to Bezuo's theorem, there must be integers x_i and y_i such that the following equation holds [24]:

$$M_i x_i + m_i y_i = gcd\left(M_i,m_i\right), 1 \le i \le L \tag{13}$$

By extending the Euclidean algorithm, the parameters x_i and y_i can be obtained, and the modular inverse $\overline{M_i} = x_i$ can be obtained. In the radar system, the percussion waveform can usually be expressed as a single-frequency complex exponential form, and the percussion waveform expression is:

$$s(t)=Aexp\left(j2\pi f_0 t\right)+\omega(t) \tag{14}$$

Among them, the amplitude and frequency of the percussion waveform are represented by A and f_0, respectively, and the additive noise is represented by $\omega(t)$. If the additive noise $\omega(t)$ is assumed to be Gaussian white noise with zero mean and variance σ^2, the signal-to-noise ratio (SNR) satisfies the following equation [24]:

$$\rho = A^2 / \sigma^2 \tag{15}$$

Among them, ρ represents the signal-to-noise ratio of the single-frequency complex percussion waveform with additive noise. From a fixed time, the percussion waveform is sampled at the sampling rate of f_s. If the sampling time is assumed to be T, the length of the sample sequence after sampling is N, and the following relationship is satisfied:

$$N = Tf_s \tag{16}$$

Eq. (14) and Eq. (16) are combined to further obtain the time domain expression of the sample sequence after sampling:

$$s(n)=Aexp\left(j2\pi f_0 \cdot n / f_s\right)+\omega\left(n / f_s\right), 0 \le n < N \tag{17}$$

If the sampling rate f_s satisfies the Nyquist sampling theorem, there is $f_s \ge 2 f_0$. At this point, the sample sequence is subjected to N-point DFT analysis, which can be obtained The Spectrum of sample sequence:

$$S(k)=DFT(s(n)), 0 \le k < N \tag{18}$$

The spectrum S(k) of the sample sequence is subjected to spectral peak search, and the index position k_p corresponding to the peak spectral line satisfies the following equation:

$$k_p = arg\,max_{0 \le k < N}\{|S(k)|\} \tag{19}$$

Then, the real frequency f_0 of the percussion waveform can be obtained according to the following formula.

$$f_0 = k_p \cdot \Delta f \tag{20}$$

Among them, $\Delta f = f_s / N$ represents the spectral resolution of the DFT.

However, in a practical environment, the percussion waveform frequency f_0 can be taken very large. In this case, if the sampling rate f_s satisfies the Nyquist sampling theorem, the value of f_s will be very large, which requires high requirements for ADC devices and high cost, which is difficult to achieve in some special occasions. At this time, the under-sampling scheme should be considered, and the sampling rate f_s does not satisfy the Nyquist sampling theorem, that is,

$$f_s < 2 f_0 \tag{21}$$

In this case, the frequency estimation value of the original percussion waveform cannot be directly obtained by using the DFT frequency estimation method. At this time, according to the periodicity of the DFT spectrogram, the obtained frequency estimate is actually the frequency remainder (or aliasing frequency) f_r, which satisfies the following equation:

$$f_r = f_0 \bmod f_s \tag{22}$$

Considering that the Chinese remainder theorem uses the system of congruence equations to solve, the method of multi-channel under sampling can be used. According to the Eq. (21), the sampling frequency $f_{s1} \sim f_{sL}$ is selected to perform L-channel under sampling on the percussion waveform respectively. At the same time, the DFT analysis is performed on the sample sequence of each channel, and the index position $k_{p1} \sim k_{pL}$ corresponding to the spectral peak is obtained by using the Eq. (19), then the frequency remainder of each channel can be obtained to satisfy the following equation:

$$f_{ri} = k_{pi} \cdot \Delta f, 1 \le i \le L \tag{23}$$

According to Eq. (22), the system of congruence equations can be obtained, and the expression is as follows:

$$\begin{cases} f_0 = n_1 f_{s1} + f_{r1} \\ f_0 = n_2 f_{s2} + f_{r2} \\ \quad \text{M} \\ f_0 = n_L f_{sL} + f_{rL} \end{cases} \tag{24}$$

Among them, n_1, n_2, L, n_L is the fuzzy multiple. Obviously, the Chinese remainder theorem can be used to solve the equation system (24), so as to obtain the percussion waveform frequency f_0. When the signal-to-noise ratio is high enough, the index position $k_{p1} \sim k_{pL}$ corresponding to the spectral peak can be directly obtained by spectral peak search according to the DFT spectrogram.

Through the above analysis, in order to complete the fuzzy-free estimation of the frequency of percussion waveforms, at least two percussion waveforms are required. Therefore, Fig. 1 presents a dual-rate defuzzification structure based on the remainder theorem.

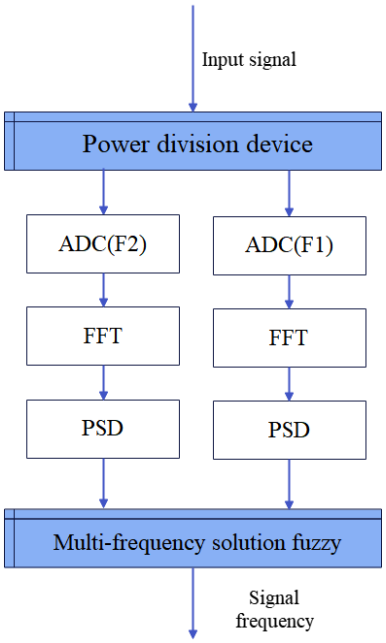


Fig. 1. Double-rate defuzzification structure based on remainder theorem.

C. Defuzzification based on Time-Frequency Analysis

The so-called time-frequency analysis is to use the joint representation of the time domain and the frequency domain to obtain an accurate description of its local characteristics.

Usually, the time-frequency analysis is performed after the real percussion waveform is converted into an analytical percussion waveform. If s(t) is assumed to be a non-stationary real percussion waveform, its corresponding analytical percussion waveform s(t) is expressed as:

$$z(t) = s(t) + jH(s(t)) \tag{25}$$

Among them, H(s(t)) represents the Hilbert transform of the real percussion waveform s(t). For the real percussion waveform, its Fourier transform satisfies the characteristic of conjugate symmetry, and the positive and negative frequency components contain the same information. Moreover, the advantage of analyzing the percussion waveform is that the negative frequency components with residual information are removed, and only the positive frequency components are retained, which will not cause information loss.

For the real percussion waveform s(t), its energy density is $|s(t)|^2$, then the total energy of the percussion waveform is expressed as:

$$E = \int_{-\infty}^{\infty} |s(t)|^2 \, dt \tag{26}$$

If the energy of the percussion waveform is limited, without loss of generality, E=1 can be set. It can be known from formula (26) that the energy density at any time point can be accurately calculated according to the real percussion waveform s(t). Therefore, the time resolution is infinite, whereas the frequency resolution is zero.

The first-order distance of the energy density is expressed as the time center $\langle t \rangle$ of the energy distribution of the percussion waveform, which satisfies the following relationship:

$$\langle t \rangle = \int_{-\infty}^{\infty} t \, |s(t)|^2 \, dt \tag{27}$$

The second moment of the energy density is expressed as the duration T^2 of the percussion waveform, and its expression is as follows:

$$T^2 = \int_{-\infty}^{\infty} (t - \langle t \rangle)^2 \, |s(t)|^2 \, dt \tag{28}$$

Among them, the percussion waveform time width T is the square root of the duration.

If the spectrum of the real percussion waveform s(t) is assumed to be $S(\omega)$, its energy density is $|S(\omega)|^2$, and the expression of the total energy of the percussion waveform is as follows:

$$E = \frac{1}{2\pi} \int_{-\infty}^{\infty} |S(\omega)|^2 \, d\omega \tag{29}$$

According to Parseval's theorem, the total energy of the percussion waveform in the time domain is equal to the total energy of the percussion waveform in the frequency domain. Therefore, there is E=1. It can be known from formula (29) that the energy density of any frequency point can be accurately calculated according to $S(\omega)$. Therefore, the frequency resolution is infinite, whereas the time resolution is zero. It is similar to the definition of time center and time width, and the expressions of frequency center and bandwidth are:

$$\langle \omega \rangle = \int_{-\infty}^{\infty} \omega \, |S(\omega)|^2 \, d\omega \tag{30}$$

$$B=\sqrt{\int_{-\infty}^{\infty}(\omega-\langle\omega\rangle)^2\,|S(\omega)|^2\,d\omega} \tag{31}$$

Generally speaking, both the time center $\langle t\rangle$ and the frequency center $\langle\omega\rangle$ of the energy distribution of the percussion waveform can be set to 0, so Eq. (28) and (31) can be further simplified into the following forms:

$$T^2=\int_{-\infty}^{\infty}t^2\,|s(t)|^2\,dt \tag{32}$$

$$\begin{aligned}B^2&=\int_{-\infty}^{\infty}\omega^2\,|S(\omega)|^2\,d\omega\\&=\int_{-\infty}^{\infty}\left(-j\frac{d}{dt}\right)^2|s(t)|^2\,dt\\&=\int_{-\infty}^{\infty}\left|\frac{ds(t)}{dt}\right|^2dt\end{aligned} \tag{33}$$

When there is $|t|\to\infty$, there is $\sqrt{t}s(t)\to 0$, the product of Eq. (32) and Eq. (33) satisfies the following relation:

$$\begin{aligned}T^2B^2&=\int_{-\infty}^{\infty}\left|\frac{ds(t)}{dt}\right|^2dt\cdot\int_{-\infty}^{\infty}t^2\,|s(t)|^2\,dt\\&\geq\left|\int_{-\infty}^{\infty}\frac{ds(t)}{dt}\cdot ts^*(t)dt\right|^2\\&=\left|\frac{1}{2}\left[ts^2(t)\Big|_{-\infty}^{\infty}-\int_{-\infty}^{\infty}|s(t)|^2\,dt\right]\right|^2\\&=\frac{1}{4}\left|\int_{-\infty}^{\infty}|s(t)|^2dt\right|^2=\frac{1}{4}\end{aligned} \tag{34}$$

Therefore, the following relationship can be further obtained:

$$TB\geq\frac{1}{2} \tag{35}$$

Eq. (35) is called the uncertainty principle. It shows that for any percussion waveform s(t) or window function h(t) with limited energy, the time resolution and frequency resolution are contradictory, and it is impossible to obtain ideal time resolution and frequency resolution at the same time.

The algorithm model of STFT(short-time Fourier transform) can be obtained as shown in Fig. 2. Choose a time-frequency localized window function, assuming that the analysis window function g (t) is stationary (pseudo stationary) within a short time interval, move the window function so that f (t) g (t) is a stationary signal at different finite time widths, and calculate the power spectrum at different times. The short-time Fourier transform uses a fixed window function, and once the window function is determined, its shape no longer changes, and the resolution of the short-time Fourier transform is also determined. If you want to change the resolution, you need to reselect the window function. Short time Fourier transform can still be used to analyze segmented stationary signals or approximately stationary signals, but for non-stationary signals, when the signal changes dramatically, the window function is required to have a high time resolution; When the waveform changes relatively smoothly, mainly for low-frequency signals, a window function with high frequency resolution is required. Short time Fourier transform cannot meet the requirements of frequency and time resolution. The window width is set to N, and the number of FFT(Fourier Transform) points is also set to N. Then, a series of continuous digital knock waveforms are input from the outside. The percussion waveform is transformed into a digital sequence of length N after passing through the data sorting module. Then, through the windowing filtering processing module, the N components of the digital sequence are respectively weighted and sent to the FFT module in sections. After frequency domain analysis, the mathematical expression of STFT is obtained. The STFT algorithm can continuously analyze the spectrum of the sampled data and output real-time analysis results.

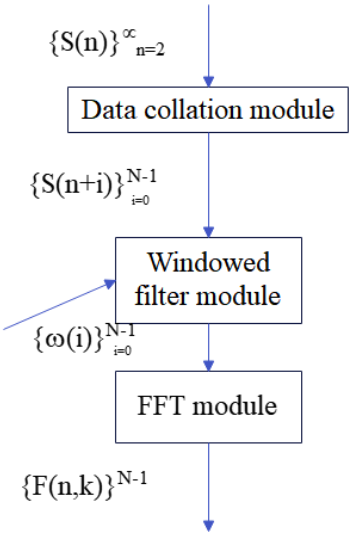


Fig. 2. STFT algorithm model.

By analyzing the algorithm model shown in Fig. 2, the mathematical expression of STFT can be obtained:

$$F(n,k)=\sum_{i=0}^{N-1}s(n+i)\omega(i)e^{-j\frac{2\pi}{N}ki} \tag{36}$$

Among them, n is the time point, and satisfies $n=mL\leq N;k$ is the channel number, and satisfies $k=0,1,\mathrm{L}\ ,N-1$. L is the number of sliding points of the time window, $\{\omega(i)\}_{i=0}^{N-1}$ is the window function, and the window width is N, which is mainly used to reduce the side lobes of the filter, thereby reducing the occurrence of spectral leakage and inter-spectral interference. F(n,k) represents the frequency domain analysis result of the kth channel at time n, that is, the frequency distribution of the percussion waveform in the time window.

The STFT algorithm can be combined with the under-sampling algorithm, so that the under sampled RF broadband percussion waveform can be directly de-blurred, so as to realize the frequency estimation of the original RF broadband percussion waveform. The FFT of the sampling sequence in the function window is the output result of the STFT. Taking the remainder theorem under sampling method as an example, the block diagram of the STFT channelization structure under the condition of under sampling is given as shown in Fig. 3.

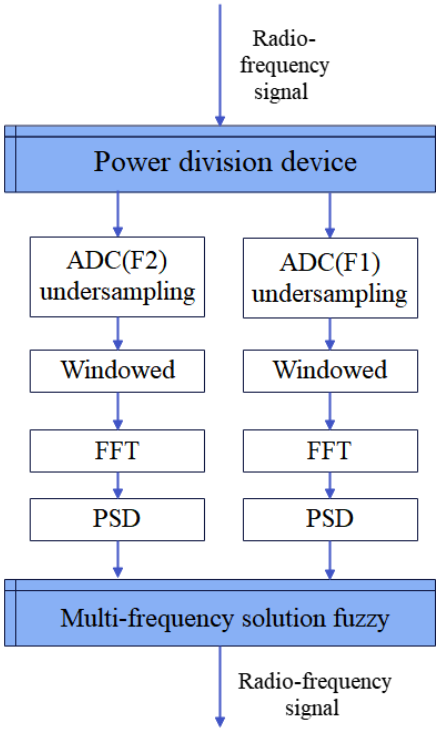


Fig. 3. Block diagram of STFT channelization structure under sampling condition.

Through the above analysis, it can be further obtained that the flow of STFT channelization under the condition of under-sampling is shown in Fig. 4.

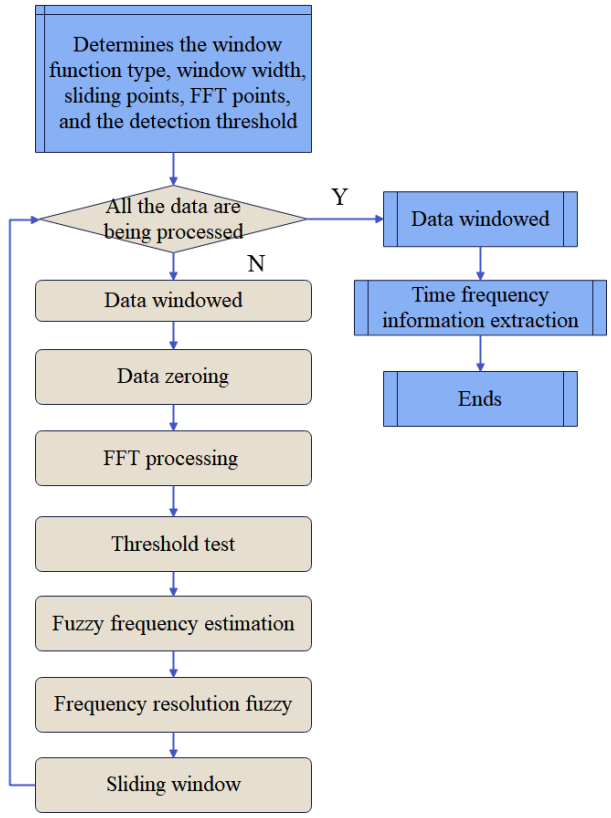


Fig. 4. Flow chart of STFT channelization under sampling condition.

D. Multi-Channel Frequency Division Defuzzification

If it is assumed that the frequency hopping range of the wideband frequency-hopping LFM percussion waveform is $f_1 \sim f_2$, the bandwidth of the LFM percussion waveform is B_s, and the following relationship is satisfied:

$$\begin{cases} B_s = f_1 \\ B_s = f_2 - f_1 \end{cases} \tag{37}$$

The ADC sampling rate is selected as f_s, and it satisfies the following relationship:

$$2B_s < f_s < 2(f_2 - f_1) \tag{38}$$

If the reference frequency is set to f_0 and the number of channels is set to M, the RF analog percussion waveforms of M channels can be down-converted to the same IF frequency hopping range through M different local oscillator percussion waveforms, and are divided into N intermediate frequency sub-bands, as shown in Fig. 5.

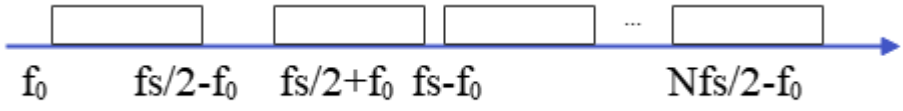


Fig. 5. IF sub-band division.

It is easy to see from Fig. 5 that the range of each IF sub-band can be expressed as:

$$\frac{(k-1)f_s}{2}+f_0 \sim \frac{kf_s}{2}-f_0, k=1,2,\ldots N \tag{39}$$

Therefore, the frequency hopping range of the IF sub-band can be expressed as:

$$f_0 \sim \frac{Nf_s}{2}-f_0 \tag{40}$$

Thus, the expressions of the intermediate frequency hopping bandwidth B_1 and the bandwidth B_2 of each sub-band are as follows:

$$\begin{cases} B_1=\frac{Nf_s}{2}-2f_0 \\ B_2=\frac{f_s}{2}-2f_0 \end{cases} \tag{41}$$

It is easy to know that when there is N=2, it is the easiest to comprehensively analyze the subsequent over-threshold detection results and frequency measurement results. When N increases gradually, the complexity of frequency deblurring will increase, but the sampling rate f_s of ADC can be reduced lower. Therefore, after the M channels are down-converted from the radio frequency band to the intermediate frequency band, each channel can be divided into N sub-bands with an interval of $2f_0$. According to the Nyquist sampling theorem, if the IF percussion waveform of a certain channel is directly frequency measured, the types of frequency ambiguity that will appear include: First, the frequency ambiguity between N sub-bands, which is caused by spectrum folding. The second is the self-ambiguous frequency band of the channel itself, and its range can be expressed as:

$$\frac{kf_s}{2}-f_0 \sim \frac{kf_s}{2}+f_0, k=1,2,\ldots N-1 \tag{42}$$

The IF frequency hopping bandwidth B_1 is the same as the frequency hopping bandwidth of each channel in the radio frequency band. The RF frequency bands of the M channels are divided, as shown in Fig. 6.

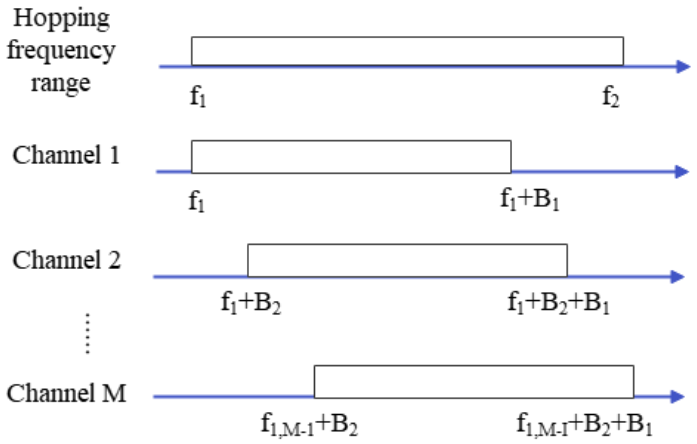


Fig. 6. Division of RF frequency bands for each channel.

As can be seen from Fig. 6, the starting point of the frequency band of the i-th channel is represented by $f_{s,i}$, and it satisfies the following expression:

$$f_{s,i}=f_{s,i-1}+B_2, i=2,3,\ldots, j-1, j+1,\ldots M \tag{43}$$

Among them, the starting point of the frequency band of the 1st channel is $f_{s,1}=f_1$, and the end point of the frequency band of the i-th channel is denoted by $f_{e,i}$, and the following expressions are satisfied:

$$f_{e,i}=f_{s,i}+B_1, i=1,2,\ldots,M \tag{44}$$

When the channel number M is selected, the total frequency band of the channel must completely cover the frequency hopping range, that is, $f_{s,M}+B_1 \geq f_2$. The condition for frequency deblurring is that the overlapping frequency band bandwidth between M channels does not exceed $f_s/2$. In order to satisfy this condition, the j-th channel is reserved here. For different situations, it is necessary to design the frequency band starting point of the channel to satisfy the frequency de-ambiguity condition. If the mid-frequency band of any channel is divided into only two sub-bands, namely N=2, the condition for frequency de-ambiguity must be established at this time. Therefore, the j-th channel may not exist.

The flow chart of multi-channel frequency division defuzzification is shown in Fig. 7.

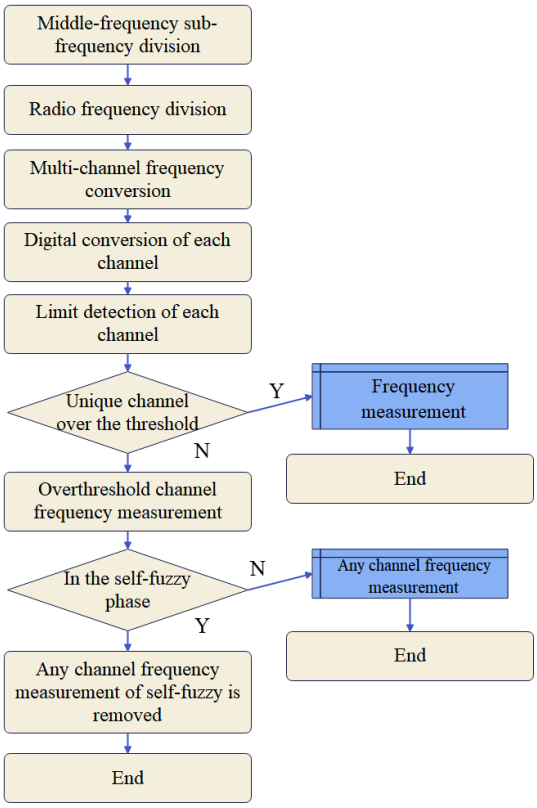


Fig. 7. Flowchart of the realization of multi-channel frequency division defuzzification.

E. Linear Interpolation Zero-Crossing Frequency Measurement

The zero-crossing frequency measurement method can directly measure the frequency through the percussion waveform time series waveform, that is, calculate the frequency of the percussion waveform by measuring the time interval of the zero point. It has the advantages of simple principle, small calculation amount, and fast operation speed.

If it is assumed that the percussion waveform to be tested is a point-frequency percussion waveform with zero initial phase, its expression is:

$$x(t)=\cos(2\pi ft) \tag{45}$$

If the sampling rate of the percussion waveform is f_s, the expression of the discrete percussion waveform after sampling is:

$$x[n]=x(nT_s)=\cos(2\pi fnT_s) \tag{46}$$

It is easy for us to know that the position where the zero point appears is:

$$2\pi fnT_s=\frac{\pi}{2}+k\pi, k\in N \tag{47}$$

$$\frac{f}{f_s}=\frac{2k+1}{4n}, n=1,2,3\ldots$$

That is, when is satisfied, the zero position of the simulated percussion waveform can be sampled. Therefore, when the sampled percussion waveform does not contain the zero position, it is necessary to determine the zero point by the method of depreciation.

The linear interpolation method is used to measure the frequency, and Fig. 8 is a partial enlarged view of the cross position of the left and right of the zero point. A and B are two zeros, and l_{AB} is the interval between zeros. Therefore, if l_{AB} can be obtained, the period of the percussion waveform is $T=l_{AB}$ and the frequency is $f=1/T$. The connection line at the positive and negative intersection of the zero point can be regarded as a straight line, that is, the method of linear measurement, there are:

$$\frac{x_2}{x_1}=\frac{y_2}{y_1} \tag{48}$$

$$\frac{x_4}{x_3}=\frac{y_4}{y_3} \tag{49}$$

According to the sampling theory, the percussion waveform sampling period is $T_s=x_1+x_2=x_3+x_4$. Combining Eq. (48) and Eq. (49), the following formula is obtained:

$$x_2=\frac{y_2}{y_1+y_2}T_s \tag{50}$$

$$x_3=\frac{y_3}{y_3+y_4}T_s \tag{51}$$

Therefore, the zero-point interval $l_{AB}=x_2+nT_s+x_3$ is obtained. Among them, n is the number of discrete points between the two points of C,D.

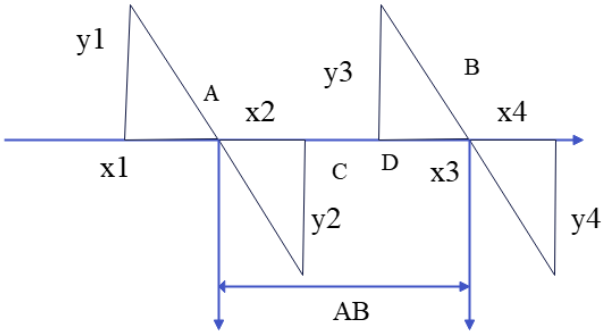


Fig. 8. Schematic diagram of linear interpolation geometry of zero point position.

IV. Model Experimental Results

A. Test Model

This article uses the algorithm model in the third part to extract the features of percussion sound signals, mine and analyze the percussion sound signals, and combine the STFT algorithm with undersampling algorithm. This enables direct deblurring of undersampled RF broadband percussion waveforms, thus achieving frequency estimation of the original RF broadband percussion waveforms, and inputting them into the system as recognizable data. It can provide reliable reference for intelligent recognition of percussion waveforms and virtual simulation of percussion in the future

After the music data is input into the feature selection model, the corresponding feature information is obtained through the long and short-term memory network. Then, it is input into the attention calculation module to analyze the feature distribution of each data block. The attention calculation layer is composed of a two-layer neural network (Fig. 9), and the structure is shown in Fig. 10.

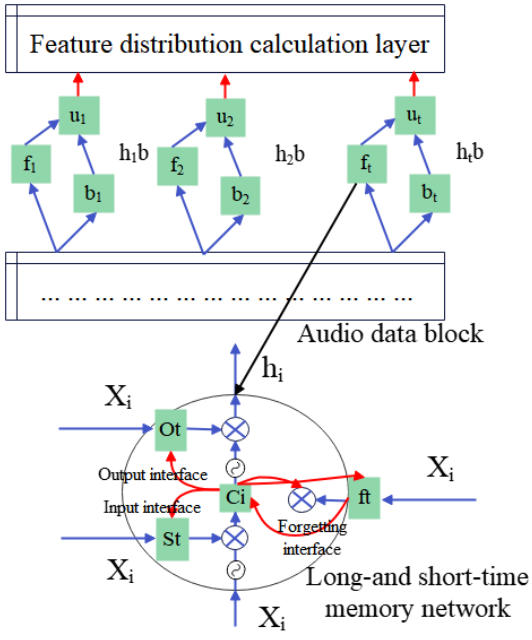
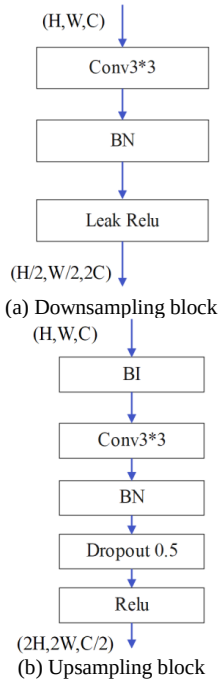
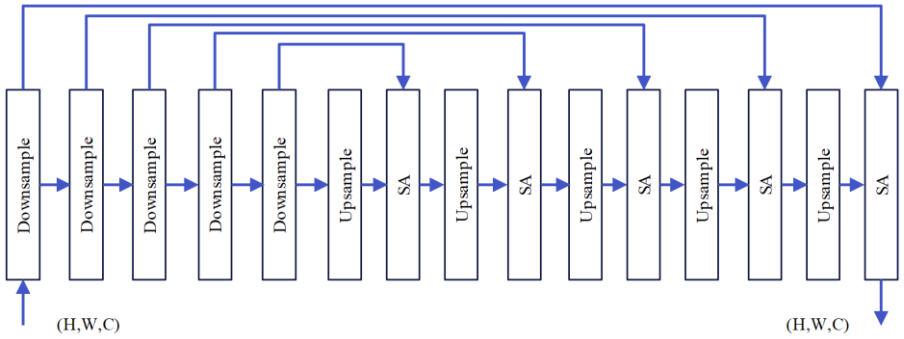


Fig. 9. Modeling of percussion big data mining based on deep neural network.

B. Analysis of Test Results

The sampling module and SA module are stacked together. In both the downsampling and upsampling modules, the size of the convolution kernel is (3,3), the stride is set to 1, and the padding mode is set to "same". Compared to using convolution kernels of size (5,5), using smaller kernels can reduce the computational complexity of the network, while using smaller and deeper kernels can achieve better performance than using larger kernels. Each downsampling block contains 3 layers of network, which are in order of size (3, 3) convolutional layer, BN layer, and Leave Relu activation layer when viewed from the direction of the input network. Each downsampling block uses a BN layer to normalize the feature information learned in this layer, avoiding overfitting. Select Leave Relu to activate the output of the downsampling layer, making the feature values of the output data smoother. Each upsampling block consists of five network layers, namely bilinear interpolation layer (BI), transposed convolutional layer of size (3,3), BN layer, dropout layer, and Relu activation layer. Abandoning the use of transposed convolution to construct upsampling blocks and instead using bilinear interpolation for upsampling, this approach reduces the number of parameters while achieving the goal of upsampling feature maps.





(C) Encoder decoder network

Fig. 10. Structure diagram of neural network model.

Model checking is performed on this basis. If the percussion waveform to be tested is an LFM percussion waveform, since the bandwidth and pulse width of the LFM percussion waveform are known, the modulation slope can also be determined. At this time, it is only necessary to measure the initial frequency according to the above process, and then perform frequency compensation to obtain the center frequency estimation of the LFM percussion waveform. The following is a simulation analysis of the frequency measurement accuracy of the LFM percussion waveform under different signal-to-noise ratios. Set up three experiments, taking the bandwidth and pulse width of LFM percussion waveform as 20MHz, 6 μ s, 40MHz, 8 μ s, 80MHz, 12 μ s, respectively, the center frequency is 380MHz, the sampling rate is 1.6GHz, the signal-to-noise ratio range is 30dB to 45dB, and the step is 1dB. The simulation results are shown in Fig. 11.

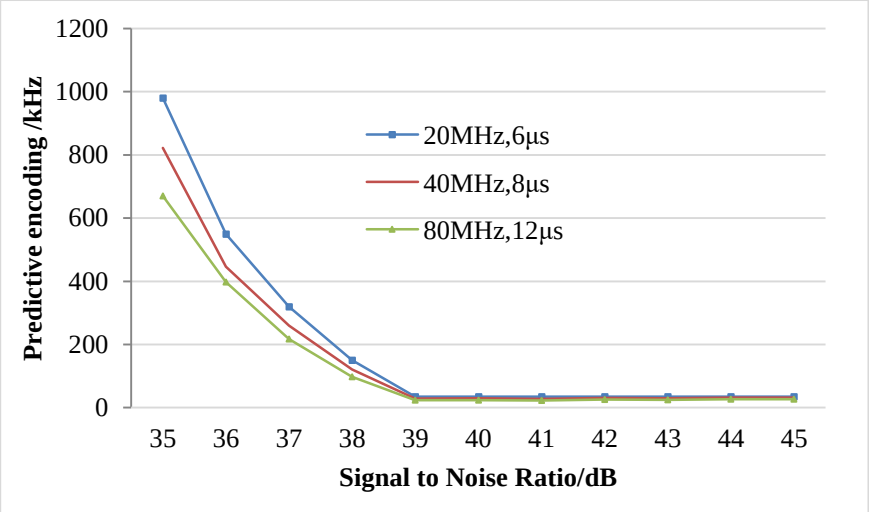


Fig. 11. The relationship between the frequency measurement error and the signal-to-noise ratio of the linear interpolation zero-crossing frequency measurement method.

As can be seen from Fig. 11, From different bandwidth and pulse width conditions, the linear interpolation zero crossing frequency measurement method has good performance in different environments, when the frequency measurement of the LFM percussion waveform is performed by the linear interpolation zero-crossing frequency measurement method, the frequency measurement error decreases with the increase of the signal-to-noise ratio. When the signal-to-noise ratio is 35dB, the frequency measurement error is close to 1MHz, which can meet the requirements of frequency measurement accuracy. When the signal-to-noise ratio is higher than 35dB, the frequency measurement error gradually decreases, and finally tends to be stable, and the frequency measurement error remains about 30kHz. Therefore, under the premise of satisfying a certain signal-to-noise ratio, the linear interpolation zero-crossing frequency measurement method has better frequency measurement accuracy and can meet the requirements of LFM percussion waveform frequency measurement.

To further verify the effectiveness of the model proposed in this paper, it is compared with the methods proposed in references [3], [7] and [10]. Reference [3] used deep learning techniques, reference [7] used multimodal sentiment classification techniques, and reference [10] used long short-term memory deep neural networks

On this basis, the waveform recognition effect of the percussion big data mining model based on the deep neural network model is tested, and the results shown in Table I are obtained.

TABLE I. THE EFFECT OF WAVEFORM RECOGNITION OF PERCUSSION BIG DATA MINING MODEL BASED ON DEEP NEURAL NETWORK MODEL

	The method described in reference [3]	The method described in reference [3]	The method described in reference [3]	The method described in this article
1	77.756	86.136	90.503	95.568
2	76.835	86.398	88.912	95.481
3	80.269	88.925	88.188	94.256
4	74.543	85.273	86.271	94.696
5	79.464	86.467	82.559	93.745
6	81.740	87.310	86.812	95.699
7	75.862	83.395	83.471	94.236
8	76.420	84.524	87.126	95.720
9	77.525	88.801	83.874	95.325
10	74.090	85.098	88.667	95.684
11	79.665	84.147	86.548	93.017
12	78.682	83.832	83.488	94.556
13	75.782	84.673	89.114	94.439
14	79.399	83.980	84.820	94.879
15	77.335	81.607	83.927	93.426

The method proposed in this article first obtains corresponding feature information through long short-term memory networks, and then inputs it into the attention calculation module to analyze the feature distribution of each data block. Compared with studies [3], [7], and [10], this article has more reliable recognition results. From the data, the method proposed in this article has better performance in waveform recognition in percussion big data mining models.

It can be seen from the above research that the percussion big data mining model based on the deep neural network model proposed in this paper has a good effect on waveform recognition.

In the process of music data mining and model construction, this method has lower implementation complexity compared to commonly used deblurring methods such as remainder theorem and time-frequency analysis. In addition, the overall design of the simulator system was completed, and the system was implemented based on a computer platform. Through analysis of test results, the accuracy of the system design and the effectiveness of frequency measurement methods were verified.

V. CONCLUSION

When performing percussion works, in some passages with complex rhythm changes or complex time signatures, the performer will also subconsciously use the head, torso and other limbs to strike the beat to remind the audience of the rhythm division logic. When there is a finger hitting passage in the music, the shape of the finger can guide the audience's understanding of the phrase. Moreover, while ensuring the timbre, the player will design the shape of the fingers when hitting or after hitting, and while ensuring the beauty of the hitting shape, it will be integrated with the inner emotion of the music. This article proposes a new deblurring algorithm based on multi-channel frequency band division. On this basis, the method of linear interpolation zero crossing frequency measurement is used to achieve fast frequency measurement of broadband frequency hopping signals. This method greatly reduces the system complexity while reducing the ADC sampling rate, and does not introduce additional deblurring errors. Finally, fast frequency measurement was achieved through linear interpolation zero crossing frequency measurement method.

The percussion big datamining and modeling methods are researched based on the deep neural network model. The simulation test shows that the percussion big data mining model based on the deep neural network model proposed in this paper has a good effect on waveform recognition.

When the signal-to-noise ratio is 35dB, the frequency measurement error is close to 1MHz, which can meet the requirements of frequency measurement accuracy. When the signal-to-noise ratio is higher than 35dB, the frequency measurement error gradually decreases and eventually stabilizes, with a frequency measurement accuracy of around 30kHz. Moreover, through comparison, it can be seen that the model in this article has better performance in sound wave recognition of percussion instruments

When simulating the frequency hopping signal echo in this article, only the target characteristics were considered, without considering the clutter and interference characteristics. At the same time, when simulating the target echo, the scattering characteristics of the target were not considered, which means that the target is considered an ideal point target. Therefore, in subsequent research work, in order to better simulate the radar environment, it is necessary to simulate clutter signals and interference signals, and analyze the echo simulation methods of extended targets.

REFERENCES

[1] Briot, J. P., & Pachet, F. (2020). Deep learning for music generation: challenges and directions. Neural Computing and Applications, 32(4), 981-993.

[2] Martín-Gutiérrez, D., Peñaloza, G. H., Belmonte-Hernández, A., & García, F. Á. (2020). A multimodal end-to-end deep learning architecture for music popularity prediction. IEEE Access, 8(2), 39361-39374.

[3] Weng, S. S., & Chen, H. C. (2020). Exploring the role of deep learning technology in the sustainable development of the music production industry. Sustainability, 12(2), 625-633.

[4] Briot, J. P. (2021). From artificial neural networks to deep learning for music generation: history, concepts and trends. Neural Computing and Applications, 33(1), 39-65.

[5] Sharma, A. K., Aggarwal, G., Bhardwaj, S., Chakrabarti, P., Chakrabarti, T., Abawajy, J. H., ... & Mahdin, H. (2021). Classification of Indian classical music with time-series matching deep learning approach. IEEE access, 9(2), 102041-102052.

[6] Pandeya, Y. R., & Lee, J. (2021). Deep learning-based late fusion of multimodal information for emotion classification of music video. Multimedia Tools and Applications, 80(2), 2887-2905.

[7] Pandeya, Y. R., Bhattarai, B., & Lee, J. (2021). Deep-learning-based multimodal emotion classification for music videos. Sensors, 21(14), 4927-4935.

[8] Rafi, Q. G., Noman, M., Prodhan, S. Z., Alam, S., & Nandi, D. (2021). Comparative analysis of three improved deep learning architectures for music genre classification. International Journal of Information Technology and Computer Science, 13(2), 1-14.

[9] Zhang, F. (2021). Research on music classification technology based on deep learning. Security and Communication Networks, 2021(1), 1-8.

[10] Hizlisoy, S., Yildirim, S., & Tufekci, Z. (2021). Music emotion recognition using convolutional long short term memory deep neural networks. Engineering Science and Technology, an International Journal, 24(3), 760-767.

[11] Bakariya, B., Singh, A., Singh, H., Raju, P., Rajpoot, R., & Mohbey, K. K. (2024). Facial emotion recognition and music recommendation system using CNN-based deep learning techniques. Evolving Systems, 15(2), 641-658.

[12] Solanki, A., & Pandey, S. (2022). Music instrument recognition using deep convolutional neural networks. International Journal of Information Technology, 14(3), 1659-1668.

[13] Dong, L. (2023). Using deep learning and genetic algorithms for melody generation and optimization in music. Soft Computing, 27(22), 17419-17433.

[14] Zinemanas, P., Rocamora, M., Miron, M., Font, F., & Serra, X. (2021). An interpretable deep learning model for automatic sound classification. Electronics, 10(7), 850-861.

[15] Rajesh, S., & Nalini, N. J. (2020). Musical instrument emotion recognition using deep recurrent neural network. Procedia Computer Science, 167(1), 16-25.

[16] Calvo-Zaragoza, J., Jr, J. H., & Pacha, A. (2020). Understanding optical music recognition. ACM Computing Surveys (CSUR), 53(4), 1-35.

[17] Yin, Z., Reuben, F., Stepney, S., & Collins, T. (2023). Deep learning's shallow gains: A comparative evaluation of algorithms for automatic music generation. Machine Learning, 112(5), 1785-1822.

[18] Kim, J., Urbano, J., Liem, C. C., & Hanjalic, A. (2020). One deep music representation to rule them all? A comparative analysis of different representation learning strategies. Neural Computing and Applications, 32(4), 1067-1093.

[19] Singh, J. (2022). An efficient deep neural network model for music classification. International Journal of Web Science, 3(3), 236-248.

[20] Zhang, Y. (2024). A Multi-sentence Music Humming Retrieval Algorithm Based on Relative Features and Deep Learning. Scalable Computing: Practice and Experience, 25(3), 1799-1806.

[21] Sheikh Fathollahi, M., & Razzazi, F. (2021). Music similarity measurement and recommendation system using convolutional neural networks. International Journal of Multimedia Information Retrieval, 10(1), 43-53.

[22] Liu, C., Feng, L., Liu, G., Wang, H., & Liu, S. (2021). Bottom-up broadcast neural network for music genre classification. Multimedia Tools and Applications, 80(3), 7313-7331.

[23] Sarkar, R., Choudhury, S., Dutta, S., Roy, A., & Saha, S. K. (2020). Recognition of emotion in music based on deep convolutional neural network. Multimedia Tools and Applications, 79(1), 765-783.

[24] Sana, S. K., Sruthi, G., Suresh, D., Rajesh, G., & Reddy, G. S. (2022). Facial emotion recognition based music system using convolutional neural networks. Materials Today: Proceedings, 62(2), 4699-4706.