Hindawi
Journal of Sensors
Volume 2022, Article ID 8175834, 9 pages
https://doi.org/10.1155/2022/8175834

Research Article

Digital Music Feature Recognition Based on Wireless Sensing Technology

Xiaoning Wang,[1] **Wei Guo**,[2] **and Weiwei Tong**[3]

[1]*School of Music, Fujian Normal University, Fuzhou, Fujian 350108, China*
[2]*School of the Arts, Xiamen University, Xiamen, Fujian 361005, China*
[3]*School of Education, Mahasarakham University, Mahasarakham 44150, Thailand*

Correspondence should be addressed to Wei Guo; wguo@xmu.edu.cn

Received 11 November 2021; Accepted 12 January 2022; Published 18 March 2022

Academic Editor: Gengxin Sun

With the rapid development of information technology, digital music is subsequently increasing in large quantities, and how a good integration of vocal input and recognition technology can be transformed into digital music can greatly improve the efficiency of music production while ensuring the quality and effect of music. This paper focuses on the implementation and application of human voice input and recognition technology in digital music creation, enabling users to generate digital music forms by simply humming a melodic fragment of a piece of music into a microphone. The paper begins with an introduction to digital music and speech recognition technology and goes on to describe the respective characteristics of various audio formats, which are selected as data sources for digital music creation based on the advantages of the files in terms of retrieval. Following that, the method of extracting musical information from music is described, and the main melody is successfully extracted from the multitrack file to extract the corresponding musical performance information. The feature extraction of humming input melody is further described in detail. The traditional speech recognition method of using short-time energy and short-time overzero rate features for speech endpoint detection is analyzed. Combining the characteristics of humming music, the method of cutting notes by two-stage cutting mode, i.e., combining energy saliency index, overzero rate, and pitch change, is adopted to cut notes, which leads to a substantial improvement in performance. The algorithm uses the melody extraction algorithm to obtain the melody line, merges the short-time segments of the melody line to reduce the error rate of emotion recognition, uses the melody line to segment the music signal to generate segmented segments, then abstracts the features of the segmented segments through a CNN-based structural model, and inputs the output of the model to the regressor in cascade with the melody contour features of the corresponding segmented segments to finally obtain the emotion *V*/*A* value of the segmented segments.

1. Introduction

In today's digital and networked era, multimedia data has become a major part of the data transmitted on the Internet information superhighway. Multimedia technology is characterized by interactive and integrated processing of audio, text, and graphic information [1]. In multimedia systems, multimedia content such as audio, image, and video currently occupies 70% of the network, and the number is growing rapidly. Voice and music are the most familiar and accustomed ways to deliver information, and sound media is the most important media other than visual media, occupying 20% of the total information volume [2]. Large-capacity, high-speed storage systems provide the basic guarantee for massive storage of sound, and the use of sound media in various industries is becoming more and more widespread [3]. Also, the implementation and application of human voice input and recognition technology in digital music creation become increasingly important. With the improvement of computer performance and Internet bandwidth, as well as the development of multimedia information compression technology and video/audio streaming

technology, the realization and application of vocal input and recognition technology in digital music creation provide a good basis and guarantee [4]. However, in the process of digital music creation, the old traditional music production methods were followed, which could not reflect the advantages and strengths of digital music and could not improve the efficiency of music production as well as the quality and popularity of music [5]. Therefore, people are no longer satisfied with music creation through the general traditional mode, and human voice input and recognition technology provides a more efficient method for digital music creation. Music as an important media resource, music creation has a very important significance for music database and digital library construction.

The sheer volume of multimedia resources on the web has prompted digital music creation to become a mainstream mode of music production [6]. People need efficient ways to compose digital music, which in turn meets the demand for massive music resources on the web. In addition, digital music composition has broad research prospects and great application value in karaoke retrieval and assisted video retrieval [7]. Given the characteristics of music itself, digital music creation is completely different from traditional music creation. The current digital music creation is based on text, which includes the name of the music, the lyricist, the singer, and the instrument played, and this information is integrated in the computer [8]. The recognition of music based on similar singers' voices or similar styles and rhythms or even similar background music sounds has become a digital music creation method that is gaining attention. This problem is cross-cutting, involves a wide range of content, and is comprehensive, involving computer science, information science, acoustics, musicology, psychology, and so on. Speech is the most dominant form that people use to communicate [9]. Therefore, speech recognition has an extremely important position in digital music creation. Sometimes we can naturally identify the singer when we hear a piece of music because their voices are different, and in general, the technology of speech recognition includes research in many fields such as acoustics, linguistics, and information processing. The scope of application is very wide. It is widely used and has been researched for a long time and has achieved very good scientific results [10]. Nowadays, audio retrieval is a kind of application related to audio information, and music as a very special kind of audio, its retrieval has been in the retrieval for lyrics, and the process of retrieval is also through a certain music or simply humming a certain lyric to find music in the music library. Up to now, the use of audio for retrieval is still very rare [11]. Therefore, the use of related technology to achieve similar music retrieval can not only change the current manual retrieval method but also singers can use the system to retrieve songs similar to their own according to their own voice and style, which not only can save a lot of time and achieve better results but also users can automatically select their favorite songs from a large number of song music libraries to meet personalized music recommendations and services.

With the development of speech signal processing technology, the system used to process audio signals now relies more and more on the effective content of the processed signal, which first and foremost is to preprocess the audio signal, extract its effective signal part (meaningful part), remove the useless part, provide an effective preprocessing method for removing the unwanted part of the audio, and can better improve the data processing. The efficiency and performance of data processing can be improved. At the same time, due to the rapid development of multimedia technology and network technology, audio resources such as songs are becoming more and more abundant, and the access to them is becoming more and more diversified and simple. How to retrieve the songs you need accurately and conveniently in the vast library of song resources has become an urgent problem. At present, there are two major types of song retrieval methods: text-based annotation methods and content-based methods. At present, all practical song retrieval systems use text-based methods, such as Baidu Music Search, JiuKuMusic.com, and Search.com. This method requires first manually annotating songs in the song database with characteristics such as song name, singer, and song classification, and then using keyword matching methods to find them. This method has some defects that are difficult to make up in practice. First, many features of songs are difficult to be accurately described by text (e.g., singer's voice characteristics, song style, rhythm, and background music tone), so it is difficult to search for these features; second, the search accuracy of text matching depends largely on the accuracy of text annotation, while the evaluation of many features (e.g., song classification and mood) is highly subjective and the accuracy is difficult to be guaranteed. Third, the text-based approach cannot be realized for the similar song retrieval demand of "finding songs by songs."

2. Related Work

For quite a long time in the past, music composition required a high level of musical theoretical knowledge and practical skills, so it has always been the case that only those who had specialized education in music were able to do so [12]. By now, many music lovers are familiar with digital music creation. Music creation has developed mainly with the development of computers, from the initial creation of music by professionals, to the creation of music by people with their own hobby of music, which is full of personalized colors, to the music that affects all aspects of people's lives now. The history of music creation also began with the emergence of electronic instruments in the 1930s and later in the 1980s with the birth of MIDI technology, which is now more widely seen in the establishment of various music studios [13]. Nowadays, more people use digital music technology to create music, so it is convenient for more composers to get rid of the old way of creating music manually, and instead, composers can use the Internet to find more factors and ways to create music, and it is convenient and fast to create music. Sonar is just one of the powerful computing software in the computer, which has more information about music creation and also provides a broader platform for musicians to show themselves; they only need to copy, paste, and other simple operations by

clicking the mouse to create music, they do not need to imagine the music performance, tone, etc., and then modify it again and again; the new way of music creation has greatly changed the way of composers in the past. The new way of composing music has greatly changed the way composers used to compose [14].

Composers do not have to worry about the difficulty of playing their works, the complexity of the scoring process, and other technical concerns [15]. Many nonmusic majors are now using their computer skills to compose music according to their own understanding of music and hobbies. And they have achieved very good results. Thus, the use of digital technology for music composition is characterized by diversity, which is manifested in many aspects, including the genre, content, and style of music [16]. Moreover, with the rapid development of computers today, digital diversity is also reflected in the important influence that computing technology brings to the field of music composition, for example, the style of composition and aesthetic orientation. Nowadays, people use computers to digitally process the audio of music in order to get the rhythm they want. People use music creation software to create music according to their own preferences so that every music lover can easily and conveniently record, edit, and other digital processing of audio in the process of creation. The digital process requires a thorough knowledge of digital audio processing technology. Among the many software programs that use computing software to create and debug MIDI music, the most practical one is Cakewalk, which is not only a tool for music lovers to create music but also for nonmusic majors to become composers by using the software to create high-quality music [17]. It requires a systematic study of music knowledge and continuous exploration of music itself. The advantage of computer software is to satisfy the dream of ordinary people to create music [18].

The process of editing audio digitally is mainly done on traditional audio, but this is very difficult for audio programs. The whole editing process is very troublesome, and the editing and processing methods are very limited and imprecise, mainly because it needs to be done with external equipment [19]. Because there are many ways to edit and process audio in this way, there are many ways to process audio in any way one can think of. In addition, this method of processing is characterized by the speed of the audio processing and the promptness of the feedback, and the success of the creation can be played and auditioned immediately. At the same time, the quality and accuracy of the audio in the editing are very high [20]. The range of adjustment for each editing-related function of the software is large. The last feature is that in the editing and processing process, no work is required from the creator, just a simple pair of computers and music editing software can do all the work, so you can get professional-grade results at a civilian price [21].

3. Digital Music Based on Recognition Technology

3.1. Algorithm for the Implementation of Human Voice Input Recognition System. The first basis for judgment is to calculate the features at the audio frame level and at the segment level; using certain regulations, the actual calculated feature values are compared with the set thresholds to identify segments of a piece of audio into three parts: silence, pure music, and speech-music mix. In the song, the sound can be divided into three categories: silence, pure music except silence, and speech-music mixture.

It is very difficult to classify the recognition of pure music and speech-music mixed segments in the music signal (the same song) because of the high confusion susceptibility of pure music and speech-music mixture. So only using two features, short-time energy and overzero rate, cannot achieve the classification effect well. In this paper, we propose a new algorithm based on human voice input and recognition technology, as shown in Figure 1.

In music signal preprocessing and feature extraction, the commonly used feature parameters are frame average energy, overzero rate or average overzero rate, resonance peak, fundamental frequency, linear prediction coefficient, and other parameters. Short-time energy is the main energy accumulated in a signal about the sampling point within a short-time audio frame, and its short-time energy calculation formula is as follows:

$$E_n = \sum_{i=1}^{n} [a(q)w(n-q)]^2. \tag{1}$$

In the formula, n is the nth short-time frame, $a(q)$ indicates the nth short-time frame within the mth sample point signal value, N is the window length, and $w(n)$ indicates the length of the N window function. The above equation can also be rewritten as

$$E_n = \sum_{i=1}^{n} [a(q)w(n-q)]^2 = \sum_{i=1}^{n} x^2(q) * h(q). \tag{2}$$

The formula $h(n)$ is defined as follows:

$$h(q) = e^2(j) + q^2. \tag{3}$$

The short-time energy can be regarded as the output of the square of the speech signal after a linear filter with impulse response $h(n)$ by the formula. Therefore, the nature of the short-time energy is to some extent related to the choice of the window function, that is, what type of window function is used and how long the window function should be chosen; if the window length is very long, the smoothing effect of the grant window will be obvious, and the corresponding curve of the short-time energy also changes slowly with time so that the characteristics of the change about the language is not well reflected; if the window length is too short, it will appear that the short-time energy changes. If the window length is too short, there will be a dramatic change in the short-time energy with time, so it becomes very difficult to get the smoothing energy function, so, in general, the window length is chosen within 10 ms-30 ms.

Short-time overzero rate refers to the number of times the value of the sampled signal changes between positive

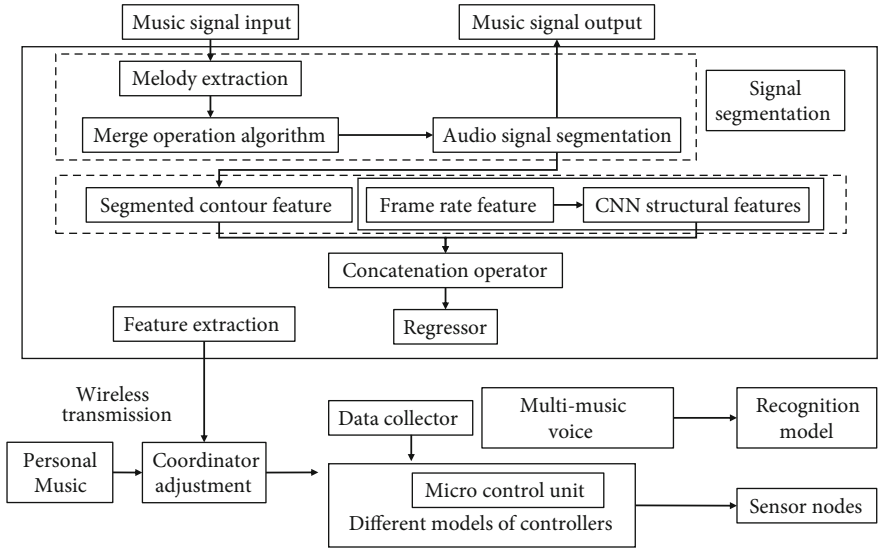

Figure 1: Principle of recognition model for pure music and speech-music hybrid clips.

and negative in a short time frame, i.e., the number of times it crosses the zero value (horizontal axis). It is the response of the average frequency of the audio signal over a short period of time and is calculated by the formula

$$Z_q = \sum_{i=1}^{n} w(q-m), \tag{4}$$

where sgn [] is the symbolic function defined as follows:

$$\text{sgn}\,[q] = \begin{cases} 1, & x(q) > 0, \\ -1, & x(q) < 0. \end{cases} \tag{5}$$

As mentioned above, the short-time overzero rate is sensitive to noise, and if the noise crosses the axes randomly and repeatedly in the computer application, many overzero artifacts can occur, which can have an important impact on the results. Therefore, in order to improve the robustness, the original signal is bandpass filtered during the operation and certain permissions can be set for the overzero rate, as shown in Figure 2. By calculating the average short-time energy and the standard deviation of the excess zero rate of the audio fragment to be recognized, we can distinguish whether the music fragment is a pure music fragment or a speech-music hybrid fragment.

3.2. Wireless Sensor-Assisted Identification. Here, the signal oscillation of the noise is guaranteed to be unaffected by the result of the overzero rate as long as it is kept within the overthreshold. Audio fragments are proposed on the basis of audio frame features. For all audio frames that make up audio, calculating the mean, variance, standard deviation, and other statistics of their audio frame features is the basic method to obtain audio fragments. In terms of audio recognition rules, the purpose of audio recognition is to roughly classify the extracted audio clips into three parts: silence, pure music, and speech-music mix. Since there is a clear distinction between these three audio categories, the recognition can be performed by the method based on the average short-time energy and the standard deviation of the overzero rate of the audio clips. Through experimental analysis, the average short time energy and overzero rate are the main two features of the standard deviation, as shown in Figure 3. The values of these two features of the pure music signal are smaller than those of speech, and by comparing the basic standard deviation of the speech waveform and the overzero rate, it can be found that there are obvious differences in their standard deviations of the overzero rate, where the music segment is located on the left side of the dotted line and the speech segment is on the right side.

The amplitude of the audio signal is small and inaudible to the human ear is the mute. The energy spectrum is characterized by low energy over a long period of time and is particularly distinctive in that the overzero rate of the mute is very different from the rest of the spectrum. Although there are also very short intervals of lower energy between each word, so it cannot be used as a silent zone. The feature of silence ratio is used here with the following rule.

(1) A silent clip is defined as when the share of silent frames in a clip exceeds the threshold value ST

(2) The definition of a silent frame is when the energy of an audio clip is well below a certain threshold. The frequency of the current sound and the loudness of the sound have a relationship to the threshold setting, the louder the sound, the higher the threshold

For this reason, the method used for extraction is the threshold ET determination method: an audio frame is considered to be silent when its temporal energy is below the threshold RT when the average ratio of the temporal energy within a 3-second window for sliding is shorter than the threshold RT. An audio clip is considered to be a pure music clip if the two characteristic values of the average short-term energy and the standard deviation of the overzero rate meet

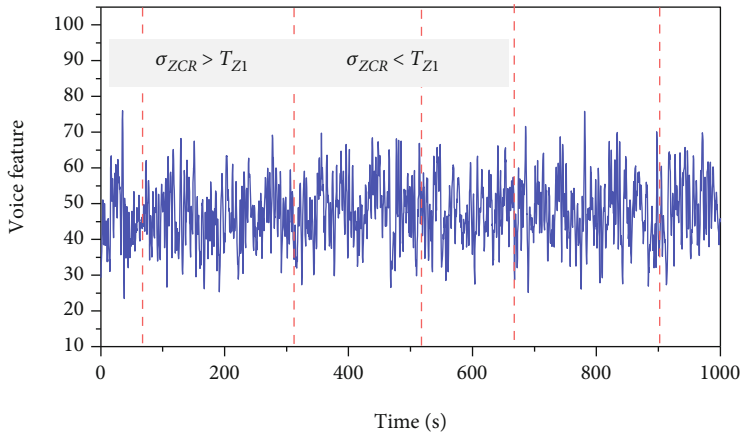

Figure 2: Threshold overzero rate diagram.

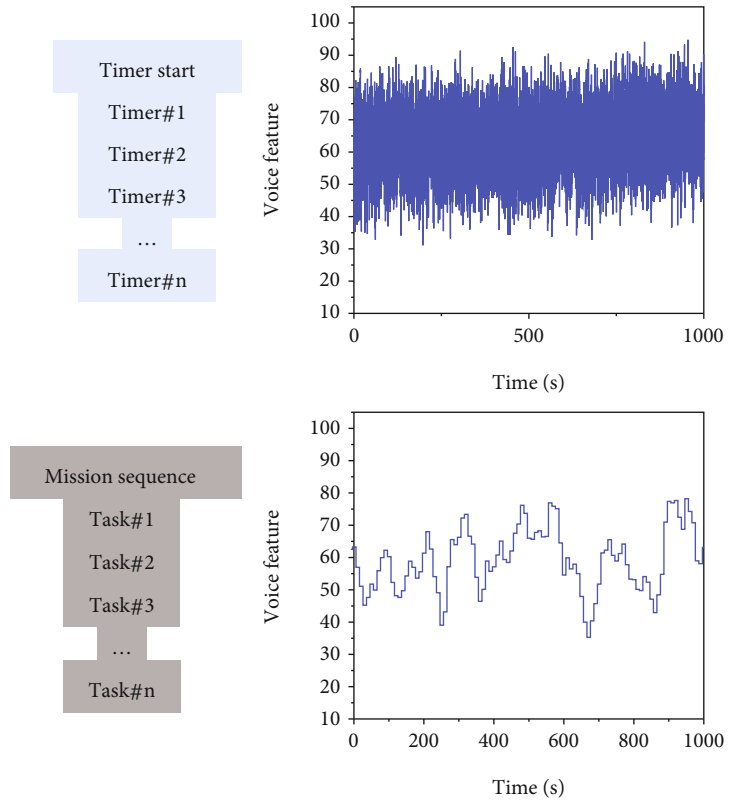

Figure 3: Feature decision table.

certain conditions; otherwise, it is a mixed speech-music clip.

3.3. Human Voice Input Recognition System. In nature, the wide variety of sounds that humans can perceive is ultimately generated by oscillations. Therefore, the first thing you come across when performing audio signal processing, and the most intuitive description of an audio signal, is the time domain waveform of the audio signal. The sound is converted into an electrical signal after passing through the transducer, and the audio signal acquisition is realized, which is the first job to be done in all audio processing systems. The electrical signal can be visually observed with an oscilloscope as the external sound changes. Since computers can only process digital signals, to draw the waveform of an audio signal in a computer, the analog audio signal must first be digitized, and then, the waveform of the audio signal is drawn based on the sampling values of each sampling point. In the waveform diagram, the changes in the energy of the audio signal can be observed very clearly, and even the time period of each note can be identified. The specific steps of digital music creation are as follows.

Sampling theorem: a time-continuous signal $m(t)$ with a frequency band limited to $(0, f)$ Hz, if $T \leq 1/2f$ seconds is the criterion for equally spaced sampling, then $m(t)$ will be completely determined by the resulting sampling value. 300 to

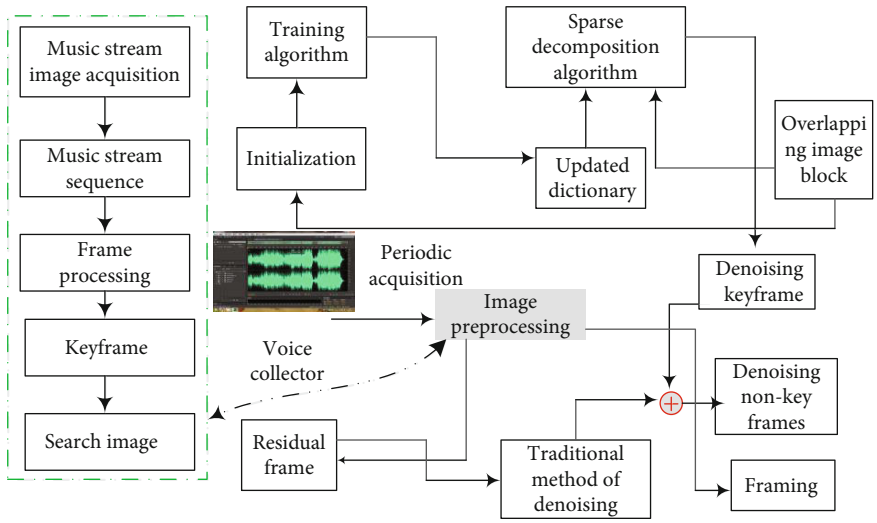

Figure 4: Architecture of the human voice input recognition system.

3400 Hz is the normal value of the human speech signal frequency band. Therefore, when the voice signal is digitally transformed, there are certain regulations for the sampling frequency, which is generally not more than 8000 Hz. 0.016~16 kHz is the most basic range according to the analysis of normal human hearing; as a young person hears more clearly, he can hear the sound of 20 kHz, so in general, equipment is often used much higher than 8000 Hz sampling frequency, so that is more enough to get a higher quality sound. The frequency range of music is related to a specific instrument. The frequency range of piano is relatively wide, from 27.5 Hz to 5000 Hz, so 10 kHz is enough to contain all the information, and the frequencies used are 22.05 kHz, 44.1 kHz, etc. What is quantization? It is mainly a process of representing the analog sampling value, which is represented by a preselected level. For the level of the analog signal according to the need for sampling, the sampling value $X(T)$ is infinite; if the size of this sample value is to be expressed in N binary digital signal, then N binary signal is expressed in M (2 of the nth power) level value. So the sampling value is divided into M discrete levels, and this process is the quantization level.

Based on what was described in the previous sections, a system was studied and developed to accept vocal humming input, retrieve it through a database, and get the user's humming name. The flow of the whole prototype system is divided into three modules. The following figure shows the functional block diagram of the system, and Figure 4 shows the block diagram of the system implementation. The higher the sampling frequency, the more accurately the discrete signal sequence will reflect the input continuous signal, which is easy to understand because the higher the sampling frequency, the less information will be lost.

After calculating the pitch saliency to obtain melodic pitch candidates, the algorithm proposes to use the continuity of pitch saliency, i.e., combining the continuity of auditory stream cues and pitch saliency to create pitch contours to reduce the problem of discontinuity of the same sound source pitch sequence due to the difficulty of distinguishing similar pitches by auditory stream cues, on top of creating pitch contours based on auditory stream cues that maintain continuity in time and pitch. Considering that the accompaniment is generally used for the modification of main notes or for the repetition of musical fragments, the repetition property of the accompaniment is proposed in the selection of melodic pitch contours. Since the repetition property is expressed in the set of pitch contours as pitch contours of equal length and pitch at different times, the dynamic time regularization (DTW) algorithm will be used to calculate the similarity between pitch contours and reduce the nonlinear deviation introduced by the difference of note length. Finally, based on the long-time relationship between adjacent pitch contours, octave errors are proposed to be detected based on the average of the pitch-weighted mean values of adjacent pitch contours in time for each frame, and melodic pitch lines are formed by smoothing melodic pitch contours using the Viterbi smoothing algorithm. Since there are strict inequalities in the pitch contours in terms of time length, the length difference range of the pitch contours satisfying the period relationship is set to. In order to remove the nonlinear deviation caused by unequal pitch contour lengths when calculating the period of pitch contours, the DTW algorithm is used to calculate the similarity between pitch contours. Considering that the difference in pitch saliency between pitch contours satisfying the period relationship is not very different, the DTW algorithm is used to calculate the difference in pitch saliency between the two pitch contours Cm and Cn, and the satisfying pitch contour is removed from the period for which the pitch contour mC is calculated, since the melodic pitch contour belonging to the dominant also has a certain long periodicity. Secondly, the system uses MFCC coefficients and short-time energy and overzero rate as feature parameters and audio retrieval technology as the recognition framework and uses GMM algorithm to train model parameters of songs, carries out the calculation of model similarity between sample songs and song feature library, realizes a song personality calculation and recommendation system, and verifies the system performance through experiments.

Programming sound in computers: the main part of programming sound in computers is the recording, playback, and operation of wa files through the sound card. In the main system of the computer, Windows, the API is used to support multimedia operations, which can be divided into two main types: low-level interfaces and high-level interfaces. The low-level interface consists of a lot of functions starting with wave, while the high-level interface is applied in two ways: they are sending messages and sending strings. When programming sound using the low-level pretext, the low-level API functions and the data structures used for sound programming and thus the handles are used.

4. Simulation Experiments and Result Analysis

Thirty humming audio clips of 10 to 15 seconds in length were used for the experiment of retrieval, and the retrieval results were output as the top three closest songs. The audio acquisition device was an external microphone. The experimental results are shown in Figure 5. From the experimental results, it can be seen that in the ideal case, the retrieval system can obtain an accuracy rate of nearly 60% for humming retrieval. The ideal situation is a situation where the user hums notes with small pauses between notes, the hummed notes are accurate, and the sampling environment is less noisy. The accuracy of system retrieval is highly dependent on the accuracy of the hummed pitch, the consistency of the hummed rhythm, and the accuracy of the MIDI information in the database. When the complexity of the MIDI file is high, such as more chord tones, and the MIDI file producer adds more subsidiary information, the retrieval rate decreases significantly. The phenomenon is related to the method of automatic extraction of MIDI file features, which is still to be improved by future research work. The hummer should have obvious pauses when humming, and the retrieval result is not satisfactory if the humming is too continuous. For the retrieval of continuous humming, more in-depth research on the note segmentation algorithm is needed.

Firstly, the system uses average short-time energy and standard deviation of overzero rate as feature parameters to accurately distinguish pure music and speech-music mixed fragments in the same song according to audio recognition rules, to achieve the function of removing pure music parts in songs. The dataset for the simulation experiments was taken from the introduced dataset DEAM15, containing a total of 489 tracks in MP3 format. Of these, 431 tracks of 45 seconds in duration were used as the development set, while the remaining 58 tracks were used as the test set. The sentiment annotation of the dataset is based on the Thayer sentiment model, with each annotation having a V/A value in the range $[-1, 1]$ and an annotation interval of 0.5 seconds. The simulations are run on Ubuntu 14.04 in the PyTorch framework, with an Intel Core i7-5930k 3.4 GHz CPU, 32 GB RAM, and TITANX 12G graphics card, and a total of 5 different random divisions of the development set are used for model training. Among them, 411 firsts were divided as the training set and 20 firsts as the validation set, and the validation set had to be randomly selected according to the genre distribution of the test dataset to ensure that the datasets matched. The evaluation index is evaluated by RMSE, which is the standard deviation of the difference between the predicted and true values of the dataset. The algorithm uses the openSMILE toolbox to extract the frame-level features of the segmented segments. The features are composed of 65 low-level acoustic descriptors, including MFCC, spectral features, and features related to the human voice. To use melodic contour features for emotion recognition, melodic contour features based on pitch, duration, vibrato, and contour type are extracted, totaling 10 features. To prevent overfitting of the model, a regularization method with random deactivation of 0.5 is used. The sequence information of short time segments is relatively small, which is prone to produce wrong emotion recognition results, so this chapter proposes the method of merging short time segments. In order to verify the effectiveness of this merging method, simulation experiments are conducted before and after the merging of short time segments, and the experimental results shown in Figure 6.

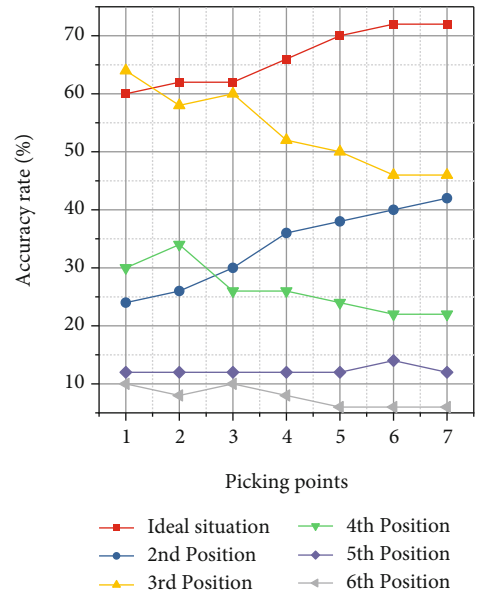

Figure 5: Audio picking experiment results.

It can be concluded from this that comparing the values of pleasantness and activation before and after merging, the merging method reduces their values, indicating that the merging method can reduce the false recognition of short-time segments. To identify the emotions of segmented segments, the features of the segmented segments need to be extracted first. To verify the effectiveness of the features extracted based on the CNN structural model and the melodic contour features, the two methods will be removed separately for testing and compared with the complete algorithm in this chapter, and the final results are shown in Figure 7. The dynamic music emotion recognition algorithm based on melody extraction and convolutional neural network is proposed for music emotions that are not uniformly distributed with time points and in order to abstract the features within adjacent emotion change points. The experimental results show that the algorithm in this chapter

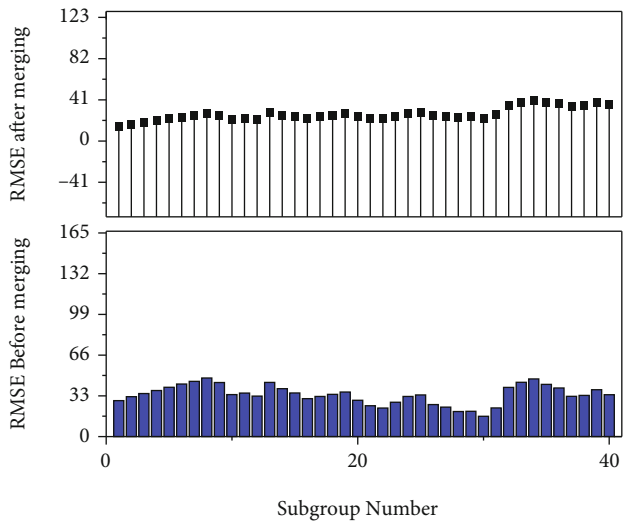

Figure 6: RMSE before and after merging short time periods.

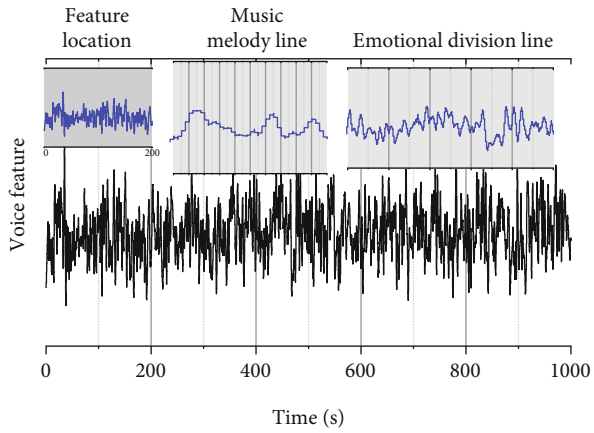

Figure 7: Music melody line and emotional division line.

achieves results close to the best recognition algorithm and greatly reduces the number of parameters of the model. The audio fragment feature used in this method is the zero-percentage standard deviation (ZCR_STD), which is defined as the standard deviation of the zero percentage of each frame in an audio fragment.

From this, it can be concluded that removing the CNN structural model significantly reduces the recognition accuracy of *V*/*A* compared to the algorithm in this chapter, which illustrates the importance of the CNN structural model for the algorithm recognition, and this result also reflects the effectiveness of the segmentation method in this chapter. For the melodic contour feature, adding this feature can further improve the recognition accuracy of the algorithm, which illustrates the effectiveness of the feature. The regressor, as the last stage of the algorithm, has many methods to choose from. To improve the recognition accuracy of the algorithm, this chapter compares regressors such as multivariate linear regression (MLR), SVR, and neutral network (NN). Among them, SVR has a 3rd polynomial kernel and NN is a single hidden layer network with 14 units.

Design and implement a prototype system that uses an audio recognition algorithm based on average short-time energy and standard deviation of overzero rate and a song personality calculation algorithm based on MFCC and GMM. The system can achieve the function of accurately removing the mute and pure music parts of a song by extracting feature parameters such as short-time energy and overzero rate of the song, extracting the speech features of the song using MFCC technique and generating the template of the song using GMM algorithm, and then performing similarity calculation of the song file using the song template library for similar song retrieval, which can accurately retrieve from the music library the songs that are similar to the sample songs that are similar (have the same characteristics or style) to the sample songs from the music library, which can achieve the requirement of personalized music recommendation. The system is developed in C++ language using VC++ compilation environment, and all functional modules are encapsulated by dynamic link libraries. The modular design of the system is realized to enhance the scalability of the system. All functions are processed by multithreaded processing technology to improve the calculation speed of the system, and at the same time, the fault tolerance and the ability to handle abnormal errors of the system are fully considered to realize the design of the reliability of the system and the ability to handle data resources.

5. Conclusion

This paper researches the implementation of human voice input and recognition technology in digital music creation, studies and analyzes the key technologies such as preprocessing technology, feature parameter extraction technology, and Gaussian mixture model algorithm of music retrieval system, and proposes the concept of "song personality" to summarize features such as song style, rhythm, and background music. We propose an audio recognition algorithm based on average short-time energy and standard deviation of overzero rate, which can distinguish pure music and mixed speech-music fragments in the same song more accurately and achieve high accuracy in processing songs of different styles, different singers, and different languages. Meanwhile, according to the need of similar song creation, a method of song personality calculation and creation based on MFCC and GMM is proposed and designed to realize the digital music creation and retrieval function to better realize the requirement of personalized digital music creation. A high accuracy recognition algorithm for pure music and speech-music hybrid audio clips based on average short time energy and standard deviation of overzero rate is proposed. The method of accurately distinguishing pure music and speech-music mixed fragments in the same song is investigated, which solves the problem of high confusion susceptibility of pure music and speech-music mixed fragment recognition and provides an effective preprocessing method for removing unwanted parts of the song. The experimental results show that by processing songs with different styles, different singers, and different languages, the average detection rate is 92.08% for pure music fragments and 96.33% for

speech-music hybrid fragments after smoothing, and the average recognition correct rate is 92.30% for pure music and 96.36% for speech-music hybrid.

By processing each note, the intensity, length, and relative pitch characteristics of the whole humming melody are extracted for the implementation of vocal input and recognition technology in digital music composition. In the melody retrieval part, a combination of exact matching algorithm and fuzzy matching algorithm is used according to the special characteristics of the humming melody to finally design the system for the implementation and application of vocal input and recognition technology in digital music composition.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. H. Barnes, G. Choby, A. J. Smith et al., "Creation of a new educational podcast: headmirror's ENT in a nutshell," *Otolaryngology head and neck surgery*, vol. 163, no. 4, pp. 623–625, 2020.

[2] S. Belikovetsky, Y. A. Solewicz, M. Yampolskiy, J. Toh, and Y. Elovici, "Digital audio signature for 3D printing integrity," *ieee transactions on information forensics and security*, vol. 14, no. 5, pp. 1127–1141, 2019.

[3] M. Bender, E. Gal-Or, and T. Geylani, "Attracting artists to music streaming platforms," *European journal of operational research*, vol. 290, no. 3, pp. 1083–1097, 2021.

[4] T. Bhangale and R. Patole, "Tampering detection in digital audio recording based on statistical reverberation features," in *Advances in Intelligent Systems and Computing*, pp. 583–591, Springer, Singapore, 2019.

[5] A. Goyal, S. K. Shukla, and R. K. Sarin, "Identification of source mobile hand sets using audio latency feature," *Forensic science international*, vol. 298, pp. 332–335, 2019.

[6] A. Goyal, S. K. Shukla, and R. K. Sarin, "A comparative study of audio latency feature of Motorola and Samsung mobile phones in forensic identification," *Indian journal of science and technology*, vol. 14, no. 4, pp. 319–324, 2021.

[7] T. Hodgson, "Quantifying music: imagined metrics in digital startup culture," *Culture theory and critique*, vol. 61, no. 4, pp. 424–439, 2020.

[8] G. Hua, H. Liao, Q. Wang, H. Zhang, and D. Ye, "Detection of electric network frequency in audio recordings-from theory to practical detectors," *ieee transactions on information forensics and security*, vol. 16, pp. 236–248, 2021.

[9] B. R. Ismanto, T. M. Kusuma, and D. Anggraini, "Indonesian Music classification on folk and dangdut genre based on rolloff spectral feature using support vector machine (SVM) algorithm," *Indonesian journal of computing and cybernetics systems*, vol. 15, no. 1, pp. 11–20, 2021.

[10] S. Li, Q. Luo, L. Qiu, and S. Bandyopadhyay, "Optimal pricing model of digital music: subscription, ownership or mixed?," *Production and Operations Management*, vol. 29, no. 3, pp. 688–704, 2020.

[11] J. Wu, B. Chen, W. Luo, and Y. Fang, "Audio steganography based on iterative adversarial attacks against convolutional neural networks," *ieee transactions on information forensics and security*, vol. 15, pp. 2282–2294, 2020.

[12] Z. Xie, W. Lu, X. Liu, Y. Xue, and Y. Yeung, "Copy-move detection of digital audio based on multi-feature decision," *in Workshop on Information Security Applications*, vol. 43, pp. 37–46, 2018.

[13] Z. Liu, Y. Huang, and J. Huang, "Patchwork-based audio watermarking robust against de-synchronization and recapturing attacks," *ieee transactions on information forensics and security*, vol. 14, no. 5, pp. 1171–1180, 2019.

[14] A. McPherson and K. Tahiroglu, "Idiomatic patterns and aesthetic influence in computer music languages," *Organised sound*, vol. 25, no. 1, pp. 53–63, 2020.

[15] L. M. Meier and V. R. Manzerolle, "Rising tides? Data capture, platform accumulation, and new monopolies in the digital music economy," *new media & society*, vol. 21, no. 3, pp. 543–561, 2019.

[16] M. Ritts and K. Bakker, "Conservation acoustics: animal sounds, audible natures, cheap nature," *Geoforum*, vol. 124, pp. 144–155, 2021.

[17] A. Rodman and S. Trivedi, "Podcasting: a roadmap to the future of medical education," *Seminars in Nephrology*, vol. 40, no. 3, pp. 279–283, 2020.

[18] G. Rodríguez-Arauz, N. Ramírez-Esparza, A. García-Sierra, E. G. Ikizer, and M. J. Fernández-Gómez, "You go before me, please: behavioral politeness and interdependent self as markers of Simpatía in Latinas," *Cultural Diversity & Ethnic Minority Psychology*, vol. 25, no. 3, pp. 379–387, 2019.

[19] A. Roper, "From print to digital: first steps in collecting digital music publications in UK legal deposit libraries," *Alexandria the journal of national and international library and information issues*, vol. 30, no. 1, pp. 32–53, 2020.

[20] J. Shen, M. Tao, Q. Qu, D. Tao, and Y. Rui, "Toward efficient indexing structure for scalable content-based music retrieval," *multimedia systems*, vol. 25, no. 6, pp. 639–653, 2019.

[21] T. Shen, J. Jia, Y. Li et al., "PEIA: personality and emotion integrated attentive model for music recommendation on social media platforms," *in National Conference on Artificial Intelligence*, vol. 34, no. 1, pp. 206–213, 2020.