


EMPIRICAL RESEARCH

Open Access



Effective acoustic parameters for automatic classification of performed and synthesized Guzheng music

Huiwen Xue¹, Chenxin Sun¹, Mingcheng Tang¹, Chenrui Hu¹, Zhengqing Yuan², Min Huang¹ and Zhongzhe Xiao^{1*} 

Abstract

This study focuses on exploring the acoustic differences between synthesized Guzheng pieces and real Guzheng performances, with the aim of improving the quality of synthesized Guzheng music. A dataset with consideration of generalizability with multiple sources and genres is constructed as the basis of analysis. Classification accuracy up to 93.30% with a single feature put forward the fact that although the synthesized Guzheng pieces in subjective perception evaluation are recognized by human listeners, there is a very significant difference to the performed Guzheng music. With features compensating to each other, a combination of only three features can achieve a nearly perfect classification accuracy of 99.73%, with the essential two features related to spectral flux and an auxiliary feature related to MFCC. The conclusion of this work points out a potential future improvement direction in Guzheng synthesized algorithms with spectral flux properties.

Keywords Guzheng, Acoustic feature analysis, Instrument recognition, Synthesizer

1 Introduction

Music is an art of creativity, imagination, and aesthetic experience and plays a vital role in social life. MIR (Music Information Retrieval) [1] is a discipline that studies how to use computer technology to analyze, extract, and organize musical information automatically. MIR's research areas cover various aspects, including music signal processing [2], music feature extraction [3], music classification [4] and annotation [5], music retrieval [6], music recommendation [7], and music generation [4]. Automatic musical instrument recognition plays a pivotal role as a critical subtask in the field of music information

retrieval (MIR). Its primary objective is to identify various types of instruments within raw music, specifically at different time intervals [8]. In recent years, researchers have used different deep-learning methods for instrument recognition. Zhong et al. [9] addressed hierarchical multi-label music tagging with joint training methods of a single DNN. Ashraf et al. [10] designed a hybrid CNN and RNN variant model for music classification. The proposed hybrid architecture combining CNN and Bi-GRU, utilizing Mel-spectrogram, demonstrated a high accuracy of 89.30%, and the hybridization of CNN and LSTM, employing MFCC, achieved an accuracy of 76.40%. Gonzalez et al. [11] used the geometric distance to study the timbral similarity between audios of different sounds and instruments and proposed a machine learning algorithm to evaluate timbral similarities. Lekshmi and Rajeev [12] discovered that the architectural choice of CNN with score-level fusion on Mel-spectro/model-gram has merit in recognizing the predominant instruments. In MIR, music genre classification is an essential multimedia

*Correspondence:

Zhongzhe Xiao
xiaozhongzhe@suda.edu.cn

¹ School of Optoelectronic Science and Engineering, Soochow University, Suzhou, China

² School of Artificial Intelligence, Anhui Polytechnic University, Wuhu, China



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

research domain for classifying music databases. PMG-Net [13] is a deep neural network-based method that automatically classifies Persian music genres. The feature selection strategy designed by Jaishankar et al. [14] based on African Buffalo Optimization (ABO) can achieve a high average accuracy of 82% when used with a neural network classifier.

Although deep learning has achieved remarkable success in many areas, its limitations are also obvious [15]. Deep learning typically needs a large number of examples to learn a simple concept [16], its models are less interpretable [17], and the ability to generalize models to small samples of data or data outside the domain [18] may be limited. There is still much room for improvement in the synthesis algorithm of traditional Chinese instruments. To address these challenges, we constructed a dataset with both performed and synthesized Guzheng samples for analysis to conduct parameter recognition.

The Guzheng is a traditional Chinese folk instrument with 21 nylon or steel strings [19], one of China's unique and critical ethnic instruments, and it can be traced back to China's Qin Dynasty (221-206 BC). In recent years, with the continuous development of electronic music synthesizer technology, synthesizers can emulate the timbre characteristics of various instruments [20], making them widely used as substitutes for traditional instrument performances. Currently, the mainstream synthesizers in the market use pulse code modulation technique, where instrument performances are sampled to obtain waveforms [21]. Subsequently, the waveforms and their corresponding synthesis coefficients are encoded and stored in the electronic synthesizer's read-only memory (ROM). Nonetheless, the Guzheng, a distinctive plucked instrument, necessitates specific playing techniques musicians employ to convey their musical interpretation, resulting in a more intricate tonal structure inherent in Guzheng performances. The existing synthetic algorithms fail to capture these playing techniques' nuanced characteristics fully.

In this work, audio samples of performed Guzheng and synthesized Guzheng music were constructed first. Second, a proper feature set was selected for feature extraction. Following that, Sequential Forward Selection (SFS) was chosen for feature selection. Finally, a random forest model was employed for a classification task. This study contributes to the preservation and evolution of traditional Chinese music in the era of AI music.

2 Related work

This section presents a brief introduction regarding the studies of the Guzheng. The focus lies on developing the Guzheng research field and the significance of the work in music.

2.1 Recent breakthroughs in Guzheng research

The essential elements of music are loudness, pitch, duration, and melody [19]. Guzheng music has a unique sound and performance technique with distinctive characteristics. In 2020, Qi et al. [22] from Tsinghua University established connections between Guzheng notes and stimuli from other sensory modalities, offering insights into the perception of traditional Chinese music. Li et al. [23] of Fudan University introduced an end-to-end Guzheng playing technique detection system using Fully Convolutional Networks suitable for variable-length audio. Zhang [24] proposed a Guzheng music conversion model based on Star generative adversarial networks (GAN) with effective conversion between similar music styles. In 2022, Chen [19] trained a Long Short-Term Memory network to generate realistic Guzheng music, a breakthrough in Guzheng music synthesis. This model generates Guzheng music with the characteristics of the technique labeled, including telekinesis, finger shake, yellow represents arpeggio, and pink represents big and small pinch.

2.2 Previous studies on music instrument classification

There was a proliferation of survey methods for classifying musical instruments, each with unique approaches and insights. In 2015, Masood et al. [25] ventured into this field with an innovative neural network-driven method, particularly emphasizing the classification of instruments such as the piano, flute, violin, drums, and guitar. In 2016, Bhalke et al. [26] brought a fresh perspective by introducing a novel feature extraction method for instrument classification, hinged on the MFCC features grounded in the Fractional Fourier Transform. This momentum was carried forward when Nagawade et al. [27] delved deeper into the instrument identification space, using Mel-frequency cepstral coefficients (MFCC) and showcasing how these features could notably enhance classification precision.

As research progressed, Avci et al. [28] conducted a classification study on distinguishing violin, and viola sounds for identical notes in 2018. To achieve this, they harnessed 16 statistical features from an extensive collection of 512 recordings per instrument and employed an array of classifiers such as LDA, k-NN, SVM, and RF. In 2020, Racharla et al. [29] focused on predominant musical instrument classification based on spectral features, utilizing the SVM classifier on the IRMAS dataset.

In 2021, Shah et al. [30] built machine-learning models and convolutional neural networks for genre-based music classification. The evolution of research in this domain culminated in 2022 with Solanki's [31] groundbreaking work on recognizing polyphonic instruments. Their approach utilized a state-of-the-art deep convolutional

neural network framework, further enriching this fascinating field’s tapestry of methods and techniques.

These research advances in instrument classification have enriched the methodology and techniques, providing important references and insights into the study of Guzheng music.

2.3 Comparative acoustic analysis in musical instruments

The acoustic analysis of musical instruments is rich in diverse methods, especially those aimed at understanding and modeling the timbral properties of musical instruments. The acoustic properties of musical instruments have been extensively analyzed in past studies. For instance, a comparative evaluation of interpolation methods for the directivity of sound sources such as speakers, singers, or musical instruments was conducted in an anechoic environment [32]. The music signal processing field also employs many techniques and representations initially developed for speech signal processing that are now used in music analysis [33]. In addition, evaluating musical instruments often involves determining the objectively measurable characteristics of a particular instrument that are closely related to the perceived timbre and playability as judged by the musician and then proposing ways to optimize those characteristics [34]. These studies shed light on the fact that methods and techniques for the acoustic analysis of musical instruments vary and evolve. They provide valuable background and frames of reference that enable us to understand better and evaluate synthesized Guzheng music’s acoustic characteristics.

3 Method

This work’s automatic classification scheme of performed and synthesized Guzheng music includes dataset construction, feature extraction, feature selection, and final classification. The flowchart of this scheme is shown in Fig. 1.

3.1 Multiple source dataset construction

As the fundamental of a machine learning-based classification task, the first step is constructing a proper dataset. A multiple-source Guzheng dataset containing performed and synthesized samples is constructed for this purpose. The actual performed Guzheng samples

are collected mainly in two ways, by amateurish players and professional players, respectively. Amateurish players come from our team, and the samples are recorded with an Audio Technica AT2020USB+ microphone. The professionally played samples are extracted from publicly accessible Guzheng music pieces online from 14 different performers. The recording environments of the Internet professional Guzheng samples differ from each other. Aiming to distinguish the tone color of the performed or synthesized Guzheng sound, the music pieces are more comprehensive than the range of traditional Guzheng music works. Instead, five distinct genres were included in the dataset construction: traditional music, children’s songs, popular songs, folk songs, and film and television music. The data collection criteria as multiple sources and genres ensured the generalizability of the classification networks trained from this dataset by weakening the influence of recording conditions, playing level, and the specified music pieces.

The synthesized part of the dataset is constructed with two different synthesis software on two different devices. The first one is the Guzheng synthesizer of GarageBand, which covers a series of different musical instruments on an iPhone, and the other is iGuzheng, specially designed for Guzheng, on an iPad. The synthesized Guzheng samples are collected by playing the virtual Guzhengs by an amateurish player who knows several different instruments. The music pieces selection in the synthesized samples collection remains consistent with the performed samples in the previously mentioned five genres.

The audio pieces, performed or synthesized, were segmented into data samples with durations ranging from 5 to 10 s, according to the borders of musical phrases. The final constructed dataset contains a total audio duration of 6960 s, which includes actual performed samples by amateurish players of 1680 s, from professional players of 2880 s, synthesized samples from GarageBand of 1920 s, and from iGuzheng of 2400 s. The number of segmented samples of each genre is listed in Table 1.

Since different Guzheng playing techniques may affect auditory perception, we categorized Guzheng recordings with synthesizers by some playing technique, which are frequently used in classic Guzheng compositions. The number of segmented samples of each technique is listed in Table 2.

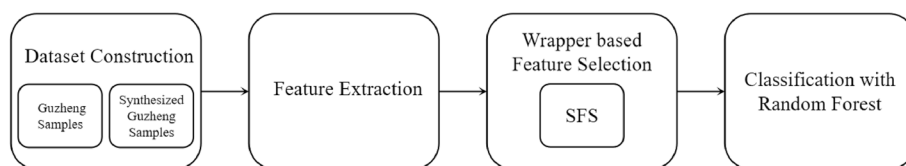


Fig. 1 Flowchart of automatic classification of performed and synthesized Guzheng music

Table 1 Number of pieces of different types of music

	Traditional	Children's	Popular	Folk	Film/TV	Total
Performed	227	76	190	197	74	764
Synthesized	167	86	174	181	80	688

Table 2 Number of pieces of different Guzheng techniques

	Glissando	Tremolo	Portamento/ Bends	Total
Performed	59	117	78	254
Synthesized	62	77	75	214

Glissando (guā zòu) is a right-hand playing technique, which means running the thumb or index finger rapidly across several strings to create a running sound effect, either away from or toward the player. Tremolo (yáo zhì) is another right-hand technique that involves “rocking” back and forth on the strings with the index finger or thumb of the right hand, plucking a string in rapid succession. Portamento (huá yīn) is a left-handed technique in which the strings are pressed with the left hand to change the pitch of the ringing tone to the desired pitch. Bend (àn yīn) is pressing a string with the left hand to raise the pitch of a string by a minor second or more before playing. Since the bends and the glissando often occur in the same segment and both change pitch, we categorize them together as one category. We also recorded each sound of the string according to the pluck methods of sounding the strings, each lasting 1–2 s. Pluck is simply plucking the strings with fingers, including gōu, tuō, mǒ, and dǎ.

3.2 Feature extraction

One of the goals of this work is to clarify what kinds of acoustic features can distinguish the actual performed and synthesized Guzheng sound, thus providing a possible improvement direction in future synthesis algorithms. Since the Guzheng dataset constructed in this work contains only 1452 samples in both classes (performed as positive samples and synthesized as negative samples), too many features in objective analyzing will cause problems such as overfitting. Thus, a small-scaled feature set, The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [35], is selected in this work. This dataset can be extracted with an audio feature extraction toolkit, OpenSmile [36], with the “sv02” configuration. This feature set has been previously employed in music identification-related works [35]. There are 88 acoustic parameters provided by this set, including a standard

parameter set with 62 parameters and an extended parameter set with 26 parameters.

The parameters provided in this set are mainly sorted as frequency-related parameters, including pitch, jitter, and formant (not applicable for Guzheng music); energy/amplitude-related parameters, including shimmer, loudness, harmonics-to-noise ratio (HNR); and spectral parameters, including alpha ratio, Hammarberg index, spectral slope, spectral flux, and MFCCs. The features calculated from statistics of the above parameters can reflect the intrinsic characteristics of audio signals. They thus can be used to distinguish the subtle difference between performed and synthesized Guzheng sound.

3.3 Wrapper-based feature selection

As the feature set applied in this work is a standard usage feature set for audio signals, initially proposed for affective computing, only some of the features in this set are suitable for our task of performing and synthesizing Guzheng music classification. Thus, a feature selection section is essential to clarify which parameters lead to the subtle tone color difference.

In order to make the classification more accurate, the Sequential Forward Selection (SFS) based method, which is a wrapper approach method, is chosen [37]. SFS is designed to iteratively build a feature subset by progressively adding the most informative features to improve classification performance. The SFS process begins with an empty feature set, and it adds one feature at each iteration by evaluating the performance of a specified classifier with the best performance. SFS terminates when the desired number of selected features or desired classification performance is reached.

A slight modification is introduced into the SFS process in our feature selection. First, individual features are evaluated by classification accuracy using only a single feature, and some “good” features are kept as a feature subset pool for further evaluation. As the whole feature set contains only 88 features, and the number of features in the kept subset pool can be even less, instead of combining each of the other features with the best one in the second iteration, we evaluate all the possible combinations of two features in the subset pool, to find out the best pair of features with the minor influence of redundancy information between two similar

features. From the third iteration, each remaining feature is added to the best combination of the previous iteration, as normal SFS does.

3.4 Classification with random forest

One of the critical links for a classification task is the selection of a classifier. In this work, the investigation between the performed and synthesized Guzheng music is carried out on a relatively small-scaled dataset, and we aim to point out the powerful distinguishable features between the two considered types, which can provide good generalizability and regardless of influence factors such as recording conditions. Based on these situations, a random forest is chosen as the classifier for this task.

As a method of ensemble learning, a random forest consists of multiple decision trees, and each decision tree is trained independently with different random samples and features. The random sampling strategy in constructing individual trees in the random forest helps reduce the model's overfitting tendency in the case of the small dataset and improves the generalizability. Although the random forest also provides a feature selection scheme in building the trees, we need to find the most efficient features explicitly. Thus, the feature selection property of the random forest is neglected in this work, while the features are screened with the wrapper method, as mentioned in the Section 3.3.

In the process of building the decision trees, the best parameters of the decision tree nodes are decided by measuring the heterogeneity index of the output variables [38], such as information entropy, as shown in Eq. (1):

$$\text{Ent}(D) = - \sum_{k=1}^{|Y|} p_k \log_2 p_k \quad (1)$$

where p_k denotes the probability that the sample belongs to the k^{th} category, $k = 1$ for performed samples, $k = 2$ for synthesized samples, and D refers to the whole dataset.

4 Experiment and results

In this section, an objective evaluation, as an automatic classification of performed and synthesized Guzheng music based on a random forest model, is taken out to find out the most compelling features in this task and to prove the performance of these features. For comparison purposes, a subjective perception evaluation by human listeners is also conducted before the objective evaluation.

4.1 Subjective evaluation

As a task concerning the difference between the tone colors of similar but differently generated audio signals, the automatic classification will be valuable if human listeners do not distinguish the two kinds of audio signals. Before the objective automatic classification, we evaluate the performed and synthesized Guzheng music by subjective perception evaluation.

We compared participants with a musical background or learning experience versus those without a musical background. We also selected college students with amateur Guzheng experience because they may be more sensitive to the characteristics of Guzheng music, including timbre, intonation, and string effects, to increase the credibility of the assessment results. To avoid selection bias, we randomly selected 25 college students with majors in science, engineering, business, and liberal arts. All participants possessed normal perceptual abilities, verified during their college entrance physicals. The participants can be divided into three groups. The first group consisted of 6 people who had yet to learn music. The second group consisted of 11 people with some experience in music learning, including singing, piano, guitar, and violin, but they had yet to be exposed to any Guzheng learning. The third group consisted of 8 people with amateur-level proficiency playing the Guzheng.

In the subjective evaluations, firstly, each participant was asked to listen to 20 randomly selected audio samples, which consisted of 10 performed samples and ten synthesized samples, played in random order. The participants needed to point out the type of each heard sample (performed or synthesized).

The distribution of the subjective perceived accuracies is shown in Fig. 2.

The average accuracy of all 25 participants is 62.00%, only a little higher than the chance level of 50% for the binary classification task, with a wide accuracy range from 40% to 85%. This result indicates that the current Guzheng synthesizers perform relatively well in imitating actual Guzheng tone color, and human listeners cannot recognize the difference between them. The average accuracies for the three groups of participants are 59.16%, 55.00%, and 73.75%, respectively. Listeners with a general, amateur learning experience of music present only advance to those with a music background, presenting even lower ability in perceptive of synthesized Guzheng. The average accuracy of the third group is 73.75%, which is significantly higher than other listeners. This result indicates that the training experience of Guzheng may improve the sensitivity to the Guzheng tone color.

Then, each participant was asked to listen to 30 randomly selected audio samples, with 10 recordings of each

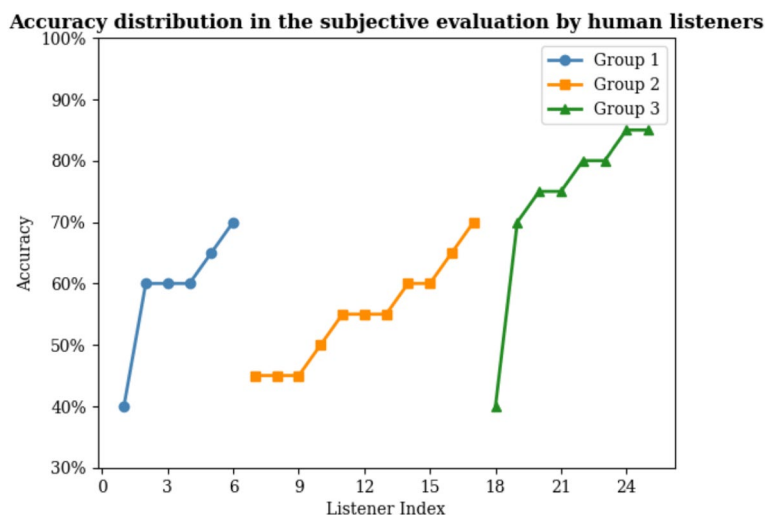


Fig. 2 Accuracy distribution in the subjective evaluation by human listeners

Table 3 Accuracies of different playing techniques

Accuracy	Glissando	Tremolo	Portamento/Bend
Group 1	65.00%	56.67%	66.67%
Group 2	54.54%	56.36%	45.45%
Group 3	71.25%	72.50%	70.00%
All	62.40%	61.60%	58.40%

playing technique. The participants needed to point out the type of each heard sample (performed or synthesized) as before. The accuracies of classification for recordings containing different playing techniques are shown in Table 3.

The overall accuracies of all 25 participants regarding the three playing techniques are 62.40%, 61.60%, and 58.40%, ranging from 58.40 to 62.40%, only a little higher than the chance level of 50% for the binary classification task. This result indicates that synthesized Guzheng music is very realistic in terms of playing techniques, so it is difficult for people to distinguish between the performed and synthesized Guzheng recordings. The accuracy of the first two groups regarding Glissando and Tremolo could have been higher, suggesting that the synthesized Guzheng music realistically mimics basic Guzheng playing techniques. It is worth noting that the accuracy of the first two groups regarding Portamento/Bend is only 66.67% and 45.45%, indicating that the synthesized Guzheng music already imitates the delicate playing technique of Portamento/Bend very well. Moreover, the accuracy of the second group with a basic knowledge of music was lower than that of the first group in all three classification tasks. However, the accuracy of the third group with amateur experience in Guzheng

learning was approximately 70% in all three playing techniques, where the accuracy of Portamento/Bend is slightly lower than any of the other two techniques. As a result, the technique of playing synthesized Guzheng music has yet to be able to trick people with Guzheng experience. The technique of synthesized Guzheng music is still some way from being perfect.

In order to get a more specific view of the degree of specialization of the participants who had experience with the Guzheng, another test was conducted. The third group was asked to listen to 20 recordings each about the four methods of sound production and then to determine which method of plucking the strings produced the sounds heard. Consequently, accuracy was less than 30% for each of them. The explanations for this result are as follows. First, although they had some amateur Guzheng experience, they had a lower training level than professional Guzheng players. They did not receive sufficiently in-depth training to distinguish subtle differences in technique. Second, amateurs of Guzheng may need more aural experience to recognize different methods of string plucking, and they may be more focused on the performance of the Guzheng piece as a whole. Third, the audio of the plucking is recorded from a cell phone, and the audio quality may need to be improved to present the nuances of the plucking method.

In summary, the subjective evaluation of the constructed dataset shows that ordinary human listeners cannot accurately distinguish the performed and synthesized Guzheng music. Synthesized Guzheng music already mimics the playing technique well, but it is still far from perfection. Automatic approaches present better performance on this task.

4.2 Objective experiment

The purposes of the objective experiment are twofold. First, we aim to verify whether acoustic features-based machine learning approaches can yield better performance in distinguishing performed and synthesized Guzheng music than human listeners do. Second, we aim to find the most efficient features that characterize the difference between the two kinds of music, thus navigating Guzheng synthesizer algorithms to a further natural level. The classifier in this evaluation is chosen as a random forest.

First, the individual features provided in the GeMAPS feature set are evaluated and separated upon the classification accuracies with every feature. For this binary classification task, features with accuracies higher than 50% can contain information of difference to some degree and form a potential feature pool for further analysis. Ten features, with the lowest single feature accuracy at 53.69%, are kept in this feature pool and listed in Table 4.

From Table 4, we can see two significant features, both related to spectral flux, present single feature classification accuracies up to higher than 93%, which are already higher than the best human perceived accuracy in our subjective evaluation. This result means that upon current Guzheng synthesis algorithms, there still exists a pronounced discrepancy in spectral, which might need to be more perceptible by the human auditory system. Other features related to MFCCs, Alpha Ratio, or jitter, also carry different information between the performed and synthesized Guzheng music and present a similar accuracy level to human perception.

The histograms of several features selected from the potential pool are shown in Fig. 3. Figure 3a shows the best single feature, “spectralFlux_sma3_amean,” with a classification accuracy of 93.30%. The histograms of the

two types (performed and synthesized) are separate. Figure 3b and c are features with moderate classification accuracies of 73.60% and 66.83%, and both present a large part of overlap between the two types. Figure 3d corresponds to a feature with a classification rate as low as 53.69%. The accuracy close to the chance level accords with the almost complete overlap between the two types.

Considering the possible information redundancy and complementary among the features, mixing features according to the performance sequence may not necessarily lead to the best accuracy. To reduce the computation burden, we combined all the other features with the best two features (with accuracies higher than 93%). The complementary properties of the features cause further improvement than the best two features. Fourteen combinations of feature pairs reach classification accuracy higher than 95%, which are listed in Table 5.

The best two pairs of features reach classification accuracies over 99%, and both pairs contain the same feature, “spectralFluxV_sma3nz_stddevNorm,” which is also a spectral flux-related feature as the best two individual features. This phenomenon again proved that the spectral flux represents the most significant defect in the state-of-the-art Guzheng synthesize algorithm. The almost perfect classification in this approach shows there is still improving space in Guzheng synthesized algorithms. It is worth noticing that the combination of the best two individual features, indexed as 1 and 2, only presents an accuracy of 95.35%, slightly better than the single feature cases. This result is because the two features are highly redundant to each other.

Following the SFS manner, feature combinations of 3 features are evaluated, with the features in the potential pool combined with the best two pairs, as shown in Table 6. Further improvement in comparison to the two feature cases is obtained. All the combinations of three features with better performance than the best two features case (99.45%) are listed in Table 6. Although features 1 and 2 presented similar performances as a single feature, in three feature combinations, feature 2 is always better than feature 1. The third feature in the best combination is surprisingly feature 9, “mfcc3V_sma3nz_stddevNorm,” which only has 54.10% accuracy on itself. However, it is an MFCC-related feature that can carry auditory perceptual information. Adding other features to the combination of features 2, 5, and 9 into combinations of 4 features does not yield further improvement in classification accuracy. Thus, the analysis terminated with the three features combination. The accuracy with all

Table 4 Accuracies of single parameters selected in the potential pool (%)

	Parameters	Accuracy
(1)	spectralFlux_sma3_amean	93.30
(2)	spectralFluxV_sma3nz_amean	93.09
(3)	slopeV500-1500_sma3nz_amean	73.60
(4)	mfcc1V_sma3nz_stddevNorm	66.83
(5)	spectralFluxV_sma3nz_stddevNorm	63.54
(6)	F2amplitudeLogRelF0_sma3nz_stddevNorm	62.93
(7)	alphaRatioUV_sma3nz_amean	60.67
(8)	F3frequency_sma3nz_stddevNorm	56.57
(9)	mfcc3V_sma3nz_stddevNorm	54.10
(10)	jitterLocal_sma3nz_stddevNorm	53.69

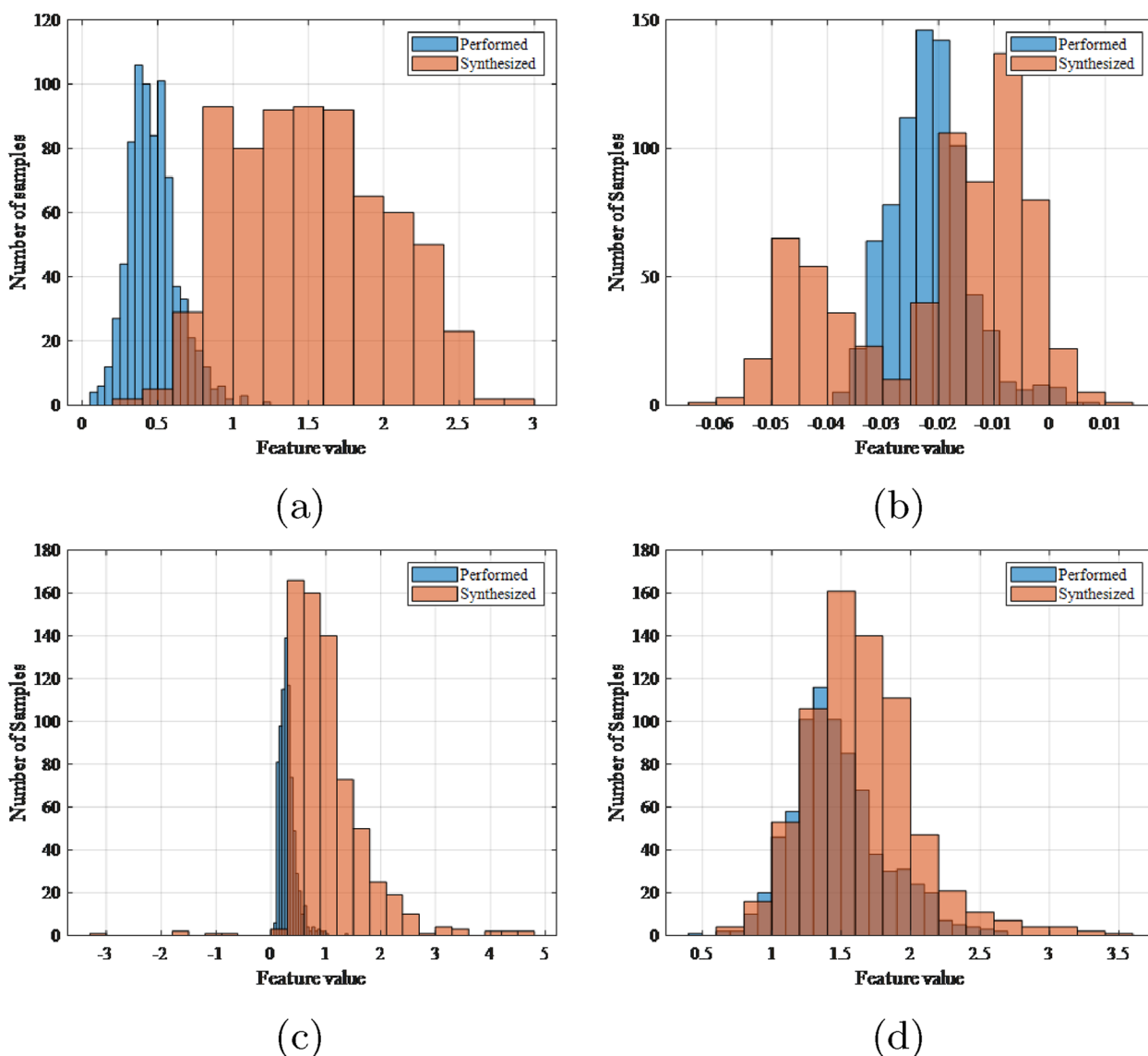


Fig. 3 Histograms of selected features, with different levels of accuracies: **a** spectralFlux_sma3_amean. **b** slopeV500-1500_sma3nz_amean. **c** mfcc1V_sma3nz_stddevNorm. **d** jitterLocal_sma3nz_stddevNorms

ten features in the potential pool is only 99.52, which is lower than the best accuracy with only three features. The redundancy and information conflict between the features may cause this result.

5 Discussions

In the discussion, we first explain the experimental methodology of this paper and then explore in detail two features that have been shown to be effective in classification tasks, followed by our validation of the experimental methodology of this paper using LSTM models. Finally, we discuss the broader contributions of this study and reflect on its scalability to other research areas.

5.1 Human and technical evaluations of synthesized Guzheng music

To distinguish synthesized Guzheng music from actual performed Guzheng music, subjective and objective evaluations are carried out in this work.

Upon the result of subjective evaluation, the human perception accuracy ranges from 40.00 to 80.00%, with an average of 57.75%, confirming that the current applied Guzheng synthesis algorithms are good enough to “deceive” human ears of most ordinary people. However, two listeners are amateur Guzheng players who gave more accurate judgment, up to 80.00%, indicating that people more familiar with the Guzheng instrument

Table 5 Accuracy of combinations of two parameters

Parameter combinations	Accuracy
(2) (5)	99.45
(1) (5)	99.38
(2) (4)	97.67
(1) (4)	97.26
(2) (8)	96.44
(1) (8)	96.31
(2) (6)	96.16
(2) (10)	96.03
(1) (6)	96.03
(2) (3)	95.83
(1) (3)	95.49
(1) (10)	95.42
(1) (2)	95.35
(2) (9)	95.08

Table 6 Accuracy of combinations of three parameters

Parameter combinations	Accuracy
(2) (5) (9)	99.73
(2) (5) (7)	99.66
(2) (5) (8)	99.66
(2) (5) (10)	99.59
(2) (5) (6)	99.59
(2) (5) (3)	99.59
(2) (5) (4)	99.59
(1) (5) (2)	99.59
(1) (5) (3)	99.52
(1) (5) (4)	99.45

can perceive more subtle cues lost in the synthesis. Thus, there is still improvement space for Guzheng synthesized algorithms.

In order to make it clear how to improve the Guzheng synthesized algorithms, acoustic feature-based analysis, with the evaluation criterion of classification accuracy, is conducted as objective evaluations. Opposite to the results in subjective evaluation, the accuracy in automatic classification of performed and synthesized Guzheng music can reach as high as 99.73% with only three features, and the best single feature can also provide high accuracies up to 93.30%, which is far better than the best human perceptual accuracy. The most crucial features presenting the defect in the Guzheng synthesizer are spectral flux-based features, which measure the spectral change between two successive frames. Features 2 and 5 (as indexed in Table 5)

are proved to compensate for each other, representing the spectral flux in means of mean value and standard deviations.

This study used an acoustic feature-based approach to assess the differences between Guzheng synthesized music and real Guzheng performances. This approach was chosen for a variety of reasons. First, acoustic features provide a clear and intuitive way to identify and explain specific differences between the two. This is particularly important in music, especially for a complex traditional instrument such as the Guzheng, as we wanted to provide clear, actionable feedback to the synthesis algorithm.

5.2 Potential causes for timbral discrepancies in synthesized Guzheng

This part constitutes one of the revision suggestions for the paper on Guzheng and synthesizer timbre recognition. Several potential reasons for the observed differences in spectral flux and Mel-frequency cepstral coefficients (MFCC) between synthesized and real samples can be outlined as follows.

5.2.1 Synthesizer parameter settings

The parameter settings of the synthesizer may significantly influence the spectral characteristics of the synthesized sound. For instance, specific filtering or modulation techniques employed by the synthesizer [39] can impact the numerical values of spectral flux and MFCC.

Filter types and characteristics: The synthesizer may utilize diverse filter types and characteristics to emulate the sound spectrum. Parameters like passband, stopband, and center frequency in these filters affect the spectral shape of the synthesized sound [40], thus influencing the computation of spectral flux and MFCC.

Modulation methods: The synthesizer might employ various modulation methods, such as amplitude or frequency modulation. These modulation techniques introduce extra spectral components, altering the spectral features of the synthesized sound and, consequently, impacting the computation of spectral flux and MFCC.

5.2.2 Theoretical foundations of synthesizer

The synthesis process might be grounded in a particular theoretical model [41], while the physical properties of a real instrument govern the Guzheng sound. Disparities in this theoretical foundation could manifest as differences in observed spectral flux and MFCC.

This result is because synthesizers generate audio through mathematical models, and their theoretical foundation

depends on the algorithms and models used. These models are often idealized and differ from real instruments' physical vibrations and resonances. The physical structure and vibration patterns of actual instruments determine real sound sources. Therefore, the theoretical disparities between synthesizers and real sound sources manifest as differences in acoustic characteristics, subsequently influencing observed variations in spectral flux and MFCC, among other acoustic features.

5.2.3 Sampling rate and bit depth

Discrepancies in sampling rate and bit depth between the synthesizer and Guzheng sound sources may contribute to variations in frequency domain features [42].

These differences can affect the number of samples and the precision of representation per sample, influencing frequency domain features. Higher sampling rates and bit depths capture audio signal details and dynamic range more accurately. Therefore, distinctions in these parameters between the synthesizer and Guzheng sound source may contribute to observed variations in frequency domain features, such as spectral flux and MFCC. The two features, spectral flux, and MFCC, could play a role in facilitating the improvement of algorithms for synthesizing Guzheng music.

5.3 Enhanced evaluation using LSTM

Since our data samples are audio compositions split into pieces with a duration of 5 to 10 s, we employed a Long Short-Term Memory (LSTM) neural network model to enhance the evaluation of our proposed method. LSTM excels at capturing and modeling dynamic features in time-series data, allowing us to understand feature information in audio data better. The experimental steps were as follows.

Firstly, we loaded the dataset from a CSV file containing audio feature representations and respective instrument labels. Then, we applied label coding to convert textual instrument labels into numerical values and then divided 70% of the dataset into a training set and 30% into a test set.

Our LSTM architecture comprises two layers, each housing 100 hidden neurons. After the LSTM layers, a dedicated classification layer was introduced, complete with dropout regularization, fully connected layers, and a sigmoid activation function. In the model training phase, we used the Adam optimizer to precisely optimize the model parameters, choosing a learning rate of 0.001 and adopting cross-entropy loss as the loss function.

The model underwent extensive training over 100 epochs. Finally, the LSTM-based model achieved an outstanding accuracy rate of 99.54%, validating our method's

effectiveness and demonstrating the high quality of the dataset we created.

It is worth noting that although LSTM, a deep learning model, achieves about the same high accuracy as random forests in our task, it takes much longer to complete the training, and it needs to be more interpretable. This paper's random forest model is more efficient than LSTM. It does not require a lot of computational resources and time and requires minimal data volume so that it can be applied to a broader range of tasks.

5.4 Contribution and extensibility to other research

There are several reasons for choosing the Guzheng as the primary research object. Firstly, Guzheng is one of China's oldest traditional musical instruments and has great cultural significance. The timbre of Guzheng is unique and distinctly different from that of other instruments, thus presenting an intriguing challenge for timbre classification studies. In addition, in recent years, the integration of Guzheng music into modern music creation has been increasing. Therefore, illuminating the differences between synthesized Guzheng music and real performance of Guzheng music is of practical significance to music makers. Guzheng is musically expressive and contains a variety of plucking techniques and musical structures, which poses challenges for timbre classification and synthesis techniques. There are few detailed studies of specific timbre attributes of the guzheng in the existing literature; therefore, our study attempts to fill this gap to some extent.

This study reveals acoustic differences between synthesized and real Guzheng performances, providing valuable insights with far-reaching implications for broader categorization efforts. The coexistence of subjective human assessment juxtaposed with objective machine learning highlights the ability of computational methods to reveal complex patterns that are usually undetectable to humans. Furthermore, the methods and insights gained from the Guzheng can be generalized to other traditional musical instruments, especially those with similar cultures, thus highlighting the versatility of the methods employed. This study reveals key differences between synthesized and real performances, assisting in improving synthesizer algorithms that could contribute to the advancement of the field of digital music synthesis.

The research methods used in this paper are highly extensible. The Random Forest model we use has the distinct advantage of being highly robust and accurate and has inherent feature selection capabilities that help focus on the most relevant acoustic features in different musical instruments. However, for some instruments, further discretion may still be required to introduce or

exclude specific features to capture their unique acoustic characteristics. For example, the guitar or violin may have a unique harmonic structure requiring additional features to characterize.

The construction and preprocessing of datasets may require different approaches for different instruments, as the differences between synthesized and real music for other instruments may be more difficult to capture than for the Guzheng, which may require more data or more sophisticated data enhancement techniques. For different instruments, the parameters of the random forest may need to be tuned for optimal performance. In addition, some instruments may have more complex acoustic properties, which may require more elaborate feature extraction strategies.

In conclusion, although our approach is designed for the Guzheng, its core ideas and techniques provide a solid foundation that can be adapted and optimized to the needs of different instruments, providing valuable insights for instrumental, cross-cultural research in music classification and synthesis techniques.

6 Conclusion

This paper analyzes the difference between the acoustic properties of actual performed and synthesized Guzheng music, aiming at the potential improvement direction in Guzheng synthesizing algorithms. A dataset with both performed and synthesized Guzheng samples is constructed for analysis. Multiple sources and genres are emphasized in constructing the dataset to ensure generalizability. Subjective evaluation of this dataset proved two facts. First, the state-of-the-art algorithms can be seen with adequate performance facing ordinary people; second, they need further improvement concerning people who are more familiar and sensitive to Guzheng music. Acoustic feature analysis according to accuracies on random forest proved that the current synthesized Guzheng music can be nearly perfectly separated from actual performed music with as few as three features up to the accuracy of 99.73%. The most efficient features, which are also the key points in improving the synthesized algorithms, are focused on spectral flux properties.

Upon the findings in this paper, we clarified the most essential “defect” in current Guzheng synthesized algorithms, and the next step of our work is to add this key finding into the next generation of Guzheng synthesizers and hope to be able to create more natural Guzheng music anywhere without must having such a non-portable instrument at hand.

Abbreviation

MIR Music Information Retrieval

Acknowledgements

Not applicable.

Authors' contributions

HX designed and programmed the proposed algorithm and wrote this paper. CS and MT participated in the algorithm design and paper writing processes. ZY participated in the algorithm programming and paper writing processes. CH participated in the data collection process. ZX and MH supervised the research. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No.61906128).

Availability of data and materials

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

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 16 June 2023 Accepted: 16 November 2023

Published online: 01 December 2023

References

1. J.S. Downie, Music information retrieval. *Ann. Rev. Inf. Sci. Technol.* **37**(1), 295–340 (2003)
2. A. Ribbrock, F. Kurth, in *2002 IEEE Workshop on Multimedia Signal Processing*. A full-text retrieval approach to content-based audio identification, pp. 194–197. IEEE, St. Thomas, VI (2002). <https://doi.org/10.1109/MMSP.2002.1203280>
3. O. Lartillot, P. Toivainen, T. Eerola, in *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation eV, Albert-Ludwigs-Universität Freiburg, March 7–9, 2007*. A matlab toolbox for music information retrieval, pp. 261–268. Springer, Berlin (2008). https://doi.org/10.1007/978-3-540-78246-9_31.
4. M.A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, M. Slaney, Content-based music information retrieval: Current directions and future challenges. *Proc. IEEE*. **96**(4), 668–696 (2008)
5. M. Lesaffre, Music information retrieval: conceptuel framework, annotation and user behaviour. Ph.D. thesis, Ghent University (2006)
6. M. Kaminskas, F. Ricci, Contextual music information retrieval and recommendation: State of the art and challenges. *Comput. Sci. Rev.* **6**(2–3), 89–119 (2012)
7. R. Typke, F. Wiering, R.C. Veltkamp, J.D. Reiss, G.A. Wiggins, et al., in *Proc. 6th international conference on music information retrieval*. A survey of music information retrieval systems. London, UK: Queen Mary, University of London; 2005. pp. 153–160.
8. S. K. Dash, S. S. Solanki, S. Chakraborty, A Comprehensive Review on Audio based Musical Instrument Recognition: Human-Machine Interaction towards Industry 4.0. *J. Sci. Ind. Res.* **82**(1), 26–37 (2023)
9. Z. Zhong, M. Hirano, K. Shimada, K. Tateishi, S. Takahashi, Y. Mitsufuji, in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. An attention-based approach to hierarchical multi-label music instrument classification, pp. 1–5. IEEE, Rhodes Island (2023). <https://doi.org/10.1109/ICASSP49357.2023.10095162>
10. M. Ashraf, F. Abid, I.U. Din, J. Rasheed, M. Yesiltepe, S.F. Yeo, M.T. Ersoy, A hybrid cnn and rnn variant model for music classification. *Appl. Sci.* **13**(3), 1476 (2023)

11. Y. Gonzalez, R.C. Prati, Similarity of musical timbres using fft-acoustic descriptor analysis and machine learning. *Eng.* **4**(1), 555–568 (2023)
12. C.R. Lekshmi, R. Rajeev, Multiple Predominant Instruments Recognition in Polyphonic Music Using Spectro/Modgd-gram Fusion. *Circuits Syst. Signal Process* **42**, 3464–3484 (2023). <https://doi.org/10.1007/s00034-022-02278-y>
13. N. Farajzadeh, N. Sadeghzadeh, M. Hashemzadeh, Pmg-net: Persian music genre classification using deep neural networks. *Entertain. Comput.* **44**, 100518 (2023)
14. B. Jaishankar, R. Anitha, F.D. Shadrach, M. Sivarathinabala, V. Balamurugan, Music genre classification using african buffalo optimization. *Comput. Syst. Sci. Eng.* **44**(2), 1823–1836 (2023)
15. N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z.B. Celik, A. Swami, in *2016 IEEE European symposium on security and privacy (EuroS &P)*. The limitations of deep learning in adversarial settings, pp. 372–387. IEEE, Saarbrucken (2016)
16. B. Zohuri, M. Moghaddam, Deep learning limitations and flaws. *Mod. Approaches Mater. Sci.* **2**, 241–250 (2020)
17. G. Marcus, Deep learning: a critical appraisal. (2018). arXiv preprint arXiv:1801.00631
18. F. Chollet, The limitations of deep learning. *Deep learning with Python* (2017)
19. S. Chen, Y. Zhong, R. Du, Automatic composition of guzheng (chinese zither) music using long short-term memory network (lstm) and reinforcement learning (rl). *Sci. Rep.* **12**(1), 15829 (2022)
20. M. Vail, *The synthesizer: a comprehensive guide to understanding, programming, playing, and recording the ultimate electronic music instrument*. Oxford University Press (2014)
21. W. Wagner, Radiometric calibration of small-footprint full-waveform airborne laser scanner measurements: basic physical concepts. *ISPRS J Photogramm. Remote Sens.* **65**(6), 505–513 (2010)
22. Y. Qi, F. Huang, Z. Li, X. Wan, Crossmodal correspondences in the sounds of chinese instruments. *Perception* **49**(1), 81–97 (2020)
23. D. Li, Y. Wu, Q. Li, J. Zhao, Y. Yu, F. Xia, W. Li, Playing technique detection by fusing note onset information in guzheng performance. (2022). arXiv preprint arXiv:2209.08774
24. L. Zhang, Analysis of guzheng music style transformation based on generative adversarial networks. *Mob. Inf. Syst.* **2022**, 1–9 (2022)
25. S. Masood, S. Gupta, S. Khan, in *2015 Annual IEEE India Conference (INDICON)*. Novel approach for musical instrument identification using neural network (2015), pp. 1–5. <https://doi.org/10.1109/INDICON.2015.7443497>
26. D. Bhalke, C.R. Rao, D.S. Bormane, Automatic musical instrument classification using fractional fourier transform based-mfcc features and counter propagation neural network. *J. Intell. Inf. Syst.* **46**, 425–446 (2016)
27. M.S. Nagawade, V.R. Ratnaparkhe, in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. Musical instrument identification using mfcc (2017), pp. 2198–2202. <https://doi.org/10.1109/RTEICT.2017.8256990>
28. K. Avci, M. Arican, K. Polat, in *2018 26th Signal Processing and Communications Applications Conference (SIU)*. Machine learning based classification of violin and viola instrument sounds for the same notes (2018), pp. 1–4. <https://doi.org/10.1109/SIU.2018.8404422>
29. K. Racharla, V. Kumar, C.B. Jayant, A. Khairkar, P. Harish, in *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*. Predominant musical instrument classification based on spectral features, pp. 617–622. IEEE, Noida (2020). <https://doi.org/10.1109/SPIN48934.2020.9071125>
30. V. Shah, A. Tandle, N. Sharma, V. Sheth, in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. Genre based music classification using machine learning and convolutional neural networks, pp. 1–8. IEEE, Kharagpur (2021). <https://doi.org/10.1109/ICCCNT51525.2021.9579597>
31. A. Solanki, S. Pandey, Music instrument recognition using deep convolutional neural networks. *Int. J. Inf. Technol.* **14**(3), 1659–1668 (2022)
32. D. Ackermann, F. Brinkmann, F. Zotter, M. Kob, S. Weinzierl, Comparative evaluation of interpolation methods for the directivity of musical instruments. *EURASIP J. Audio Speech Music Process.* **2021**, 1–14 (2021)
33. M. Muller, D.P.W. Ellis, A. Klapuri, G. Richard, Signal processing for music analysis. *IEEE J. Sel. Top. Signal Process.* **5**(6), 1088–1110 (2011). <https://doi.org/10.1109/JSTSP.2011.2112333>
34. D.M. Campbell, Evaluating musical instruments. *Phys. Today* **67**(4), 35–40 (2014). <https://doi.org/10.1063/PT.3.2347>
35. B. Ye, X. Yuan, G. Peng, W. Zeng, in *2022 6th Asian Conference on Artificial Intelligence Technology (ACAIT)*. A novel speech emotion model based on cnn and lstm networks (2022), pp. 1–4. <https://doi.org/10.1109/ACAIT56212.2022.10137926>
36. F. Eyben, M. Wöllmer, B. Schuller, in *Proceedings of the 18th ACM International Conference on Multimedia, Opensmile: The munich versatile and fast open-source audio feature extractor, MM '10* (Association for Computing Machinery, New York, 2010), p. 1459–1462. <https://doi.org/10.1145/1873951.1874246>
37. V. Giedrimas, S. Omanovič, in *2015 IEEE 3rd Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*. The impact of mobile architectures on component-based software engineering (2015), pp. 1–6. <https://doi.org/10.1109/AIEEE.2015.7367317>
38. R. Genuer, J. M. Poggi, C. Tuleau-Malot, N. Villa-Vialaneix, Random forests for big data. *Big Data Research* **9**, 28–46 (2017)
39. M.R. Hasanabadi, M. Behdad, D. Gharavian, in *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mfccgan: A novel mfcc-based speech synthesizer using adversarial learning, pp. 1–5. IEEE, Rhodes Island (2023). <https://doi.org/10.1109/ICASSP49357.2023.10095873>
40. X. Wang, S. Wang, Y. Guo, in *2021 IEEE 3rd Eurasia Conference on IOT, Communication and Engineering (ECICE)*. Research on speech feature extraction and synthesis algorithm based on eemd, pp. 362–365. IEEE, Yunlin (2021). <https://doi.org/10.1109/ECICE52819.2021.9645625>
41. U.G. Echeverria, F.E.G. Castro, J.M.D.B. López, in *2010 20th International Conference on Electronics Communications and Computers (CONIELECOMP)*. Comparison between a hardware and a software synthesizer, pp. 311–314. IEEE, Cholula, Puebla (2010). <https://doi.org/10.1109/CONIELECOMP.2010.5440747>
42. S.A. Tripathy, A.A. Sakkeer, U. Utkarsh, D. Saini, S.J. Narayanan, S. Tiwari, K. Pattabiraman, R.T. Shankarappa, in *2023 2nd Edition of IEEE Delhi Section Flagship Conference (DELCON)*. Sound ai engine for detection and classification of overlapping sounds in home environment, pp. 1–6. IEEE, Rajpura (2023). <https://doi.org/10.1109/DELCON57910.2023.10127311>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.