Research Article

Mingjin Han*

Artificial intelligence-driven tone recognition of Guzheng: A linear prediction approach

https://doi.org/10.1515/dema-2024-0009
received August 14, 2023; accepted April 7, 2024

Abstract: The Guzheng, an ancient and widely cherished musical instrument in China, serves as a significant cultural heritage with its enchanting melodies. The advent of artificial intelligence offers a novel avenue for the automatic recognition of guzheng music. This article introduces a pitch detection and recognition approach leveraging an enhanced capsule network. By integrating relative spectrum-aware linear prediction and Mel-scale frequency cepstral coefficients into novel features and feeding them into an optimized capsule network, the method achieves precise pitch recognition from audio inputs. Evaluation on a custom dataset indicates a high accuracy in identifying distinct pitches across the guzheng's 21 strings, with an average recognition rate of 98.15%. Furthermore, to assess the algorithm's resilience to interference, comparative experiments against three other network models were conducted in various noise conditions. Our approach outperformed all others, maintaining over 96% accuracy even in noisy environments, demonstrating superior anti-interference capabilities.

Keywords: tone recognition, Guzheng, artificial intelligence, capsule network, RelAtive SpecTrAl – Perceptual Linear Predictive

MSC 2020: 05-08, 65-03

1 Introduction

The Guzheng, also known as Hanzheng and Qinzheng, is a traditional Chinese plucked instrument with over 2,500 years of history. The instrument, approximately 163 cm in length, has a slightly curved upper surface and contains multiple strings stretched over movable bridges for pitch adjustment. Its construction varies, including the common 21-string version, comprising top and bottom plates, side panels, and sound holes for amplification. Musical tone is a tone with a fixed pitch produced by the regular vibration of the pronouncing object. Tones are the most important and basic materials used in music, and the melody and harmony in music are all composed of musical tones. From the perspective of acoustic analysis, musical sound has three elements: pitch (pitch), loudness (sound intensity), and timbre, which can also be represented by fundamental frequency, amplitude, and octave. Tonal components are components of musical audio domain representations. The tonal components in the frequency spectrum can be divided into two categories: the first is the harmonic sequence whose frequency is located at an integer multiple of the fundamental frequency, and the second is the non-harmonic sequence tonal components. Rayleigh [1], Pierce [2] found that the harmonic sequences of stringed instruments are shifted by integer multiples of the fundamental frequency. Young [3], Rigaud et al. [4], Rossing and Fletcher [5] pointed out that the excursion is due to the stiffness of the strings and

* **Corresponding author: Mingjin Han**, Department of Music, Xinxiang University, Xinxiang 453003, China, e-mail: hanmingjin2020@xxu.edu.cn

that excursion can bring “warmth” to the timbre of plucked instruments. The use of technologies such as computers to analyze and understand sound and music is called “Computer Audition” or “Machine Listening” (ML). Most of the existing pitch detection algorithms focus on the correspondence between pitch and frequency (e.g., the frequency of the central C tone is 2 61.6 Hz). Klapuri and Davy [6] introduced an MT system based on the Gold–Rabiner algorithm, which calculates the frequency through the structure of the waveform. Drugman et al. [7] discussed four algorithms: the harmonic spectrum method, the Cepstrum method based on HPS, the maximum likelihood method (ML), and the weighted autocorrelation function method. The first three of them are based on frequency range calculations, and the fourth is based on time range calculations.

The advent of artificial intelligence (AI) has significantly transformed computer auditory tasks, particularly in the realm of musical tone recognition and instrument identification, including the intricate tones of the Guzheng. Among the various AI methodologies, convolutional neural networks (CNN) have emerged as a powerhouse, delivering unprecedented results across a spectrum of applications from computer audition to vision and anomaly detection [8,9]. CNNs excel at preserving the fundamental characteristics of input data, showcasing remarkable precision in image classification, retrieval, and object localization. However, their focus on local features often overshadows the holistic spatial structure of objects, a consequence of their pooling strategies. While pooling enhances network classification robustness, it simplifies the input data’s complex attributes like position, direction, and size. This simplification boosts feature detection efficiency but at the expense of a comprehensive feature understanding [10].

In order to break this inherent limitation, the network architecture can not only identify the local features of objects but also preserve the hierarchical relationship of its overall spatial features. In 2017, The capsule network model [11] was proposed, which only contains a convolutional layer and a fully connected capsule layer, which is a brand-new network structure. One of the primary benefits is their ability to capture hierarchical relationships between different features of an object, which is crucial for understanding complex structures and nuances in musical tones. Unlike CNNs, which excel at identifying spatial hierarchies in visual data but may struggle with capturing the full spatial relationships due to their pooling layers, capsule networks preserve these relationships, ensuring a more detailed and comprehensive understanding of the data. Capsule network focuses on understanding memory. It can not only memorize local features of objects but also grasp the mapping relationship between local features and overall features.

Additionally, capsule networks excel in recognizing subtle variations in Guzheng tones by encoding detailed attributes like pose and deformation, outperforming RNNs in capturing complex spatial relationships for tasks like tone recognition. Furthermore, the dynamic routing mechanism within capsule networks, which allows capsules to send information to higher-level capsules more efficiently, ensures that the network focuses on the most relevant features for the task at hand. This leads to improved performance in recognizing Guzheng tones. This mechanism also contributes to the robustness of capsule networks against interference and noise, a significant advantage over other network models that might be more sensitive to such factors.

In this article, an improved capsule network is designed for recognizing Guzheng tones. In order to enhance the robustness of the algorithm, the relative spectrum sensing linear prediction (RelAtive SpecTrAl-Perceptual Linear Predictive, RASTA-PLP) features and Mel-Scale Frequency Cepstral Coefficients (MFCC) features are fused. At the same time, the extracted features, the first-order difference coefficient, and the second-order difference coefficient are mapped to three channels to form the overall feature and then input into the improved capsule network. This article evaluates our method by crawling Guzheng music from the Internet and constructing a data set. The database includes 46 pure music Guzheng pieces crawled from various audio websites and video websites on the Internet, and digital spectrum tones are extracted from the original music. This article evaluates the accuracy and anti-interference ability of our method in recognizing 21 tones of Guzheng. The experimental results show that the improved capsule network proposed in this article has good performance.

2 Related work

Tonal components are components of musical audio domain representations. As a type of plucked instrument, the researchers analyzed the composition of the piano's tonal components. Bilbao [12] summed up the law of the frequency position of the harmonic series of piano strings and proposed to use the "dissonance coefficient" to describe the degree of frequency shift of the harmonic series. Chabassier et al. [13] found through the measurement of the vibration of piano strings. The acoustic spectrum of strings contains non-harmonic tonal components. The frequency positions of these tonal components can be expressed as a sum of shifted lower harmonic frequencies, and this phenomenon exists in plucked or percussion instruments such as guitars. Woodhouse [14] pointed out that due to the longitudinal vibration of the string, the nonlinear mixing of harmonics will be generated, which means that the low-order "harmonic pair" generated by the transverse vibration of the string as the "mother harmonic" will produce non-harmonic tonal components near higher harmonics.

The musical tone signal is composed of the fundamental tone and the overtone, and the fundamental tone determines its pitch. Therefore, the detection of the fundamental tone period is the key to the identification of piano notes [15,16]. The detection methods of the pitch period mainly include frequency domain identification and time domain identification. The short-time autocorrelation method, a traditional time-domain detection algorithm, stands out for its simplicity and widespread application. However, it is prone to errors such as pitch frequency doubling or halving.

On this basis, it is a classic improved algorithm to perform three-level center clipping operations before calculating the autocorrelation function [17,18]. Since this operation removes the part where the energy of each note is relatively concentrated in the central area and retains the energy near the peak, it can reduce the amount of calculation and speed up the operation. Researchers have proposed various smoothing filtering algorithms [19–21], the purpose of which is to filter out various interference points. Guo [22] proposed an autocorrelation method that applies the zero-insertion algorithm and the corresponding low-pass filter to three-level clipping. Yuan-yuan and Shun [23] proposed to combine the three-level center clipping autocorrelation function with the cyclic mean amplitude difference function. The mentioned algorithm demonstrates a relatively satisfactory recognition rate for music characterized by a slow tempo, yet its performance significantly drops when applied to faster-paced rhythms.

Rao et al. [24] proposed a music chord identification method based on robust scale features and measure learning SVM, which can reduce the influence of human voice on chord progression and restore the harmonic information corresponding to chords. A model is established for the harmonic information and human voice information corresponding to chords in the spectrum, and a dual-objective optimization problem is constructed to effectively reconstruct the harmonic information corresponding to the chords and remove the human voice at the same time. This approach effectively reduces noise in audio signals, but it is not very good at recognizing notes. Also, it is not very accurate at pulling out sounds from audio and cannot remove interference well.

3 Design of our method

3.1 Feature extraction

At present, the method of tone recognition is mostly a time-domain method. Through the two steps of resampling and time-domain scaling, the frequency domain information of the signal can be changed, and different tones can be distinguished. Therefore, when recognizing the tones of the Guzheng, the features used should be discriminative in recognizing different tonal factors.

We first consider how to distinguish musical tones. In the realm of tone recognition, the transition from time-domain methods to more sophisticated techniques has significantly enhanced the ability to distinguish between various tonal characteristics. Common features used to detect different tones are MFCC, linear prediction coding coefficients, linear frequency cepstral coefficients, etc. In other related fields of acoustics, perceptual linear predictive (PLP) and RASTA-PLP features are also common. Faced with the robustness problem, PLP and RASTA-PLP features are often used to enhance the robustness of the algorithm, which has been verified in musical tone recognition tasks [25]. Among these, RASTA-PLP and MFCC features stand out for their distinct advantages in recognizing and differentiating musical tones, particularly in the context of the Guzheng tone recognition.

RASTA-PLP, an evolution of the PLP approach, is specifically designed to enhance robustness in tone recognition by leveraging the human auditory system's processing mechanisms. This method selects signals processed through auditory model mechanisms, bypassing the limitations associated with linear prediction coefficient analysis in the time domain. By incorporating critical band analysis, equal loudness pre-emphasis, and intensity-loudness transformation, RASTA-PLP effectively captures the essence of musical tones. The addition of RASTA filtering, which employs a band-pass filter to attenuate stable frequency components and thus suppress channel noise, further refines this feature extraction, maintaining the signal's formant structure effectively [26].

MFCC features, renowned for their performance in tone detection tasks across different pitches, utilize a sequence of processing steps including the fast Fourier transform (FFT), Mel filtering, logarithmic operations, and the discrete cosine transform. This process results in a robust set of features capable of capturing the nuanced spectral properties of musical tones, making MFCC an invaluable tool for tone detection.

Both RASTA-PLP and MFCC features offer complementary strengths in tone recognition. RASTA-PLP's emphasis on mimicking human auditory processing and its effective noise suppression capabilities enhance the robustness and accuracy of tone recognition algorithms. Meanwhile, MFCC's detailed spectral analysis provides a reliable means for distinguishing between tones of varying pitches. The fusion of these features not only addresses the challenge of robustness in the presence of noise but also ensures a comprehensive and nuanced understanding of musical tones, thereby significantly improving the efficacy of tone recognition systems, especially in complex instruments like the Guzheng.

Taking the feature MFCC extraction process as an example, after the signal is preprocessed, it needs to undergo FFT transformation first, and the FFT transformation will calculate the spectral information of each frame and then process it through the Mel filter. The calculation process of this step is

$$P_m = \int_{f_{lm}}^{f_{um}} B_m^{(w)} |F(w)|^2 \mathrm{d}w, \tag{1}$$

where w is the index of the signal frame, $F(w)$ is the spectral information of the signal, $B(w)$ is the Mel filter, the subscript m refers to the mth Mel filter passed through, $1 \leq m \leq M$, with M being the number of Mel filters, f_{um} is the upper limit of the cutoff frequency of the filter, f_{lm} is the lower cutoff frequency, and P_m is the power obtained through the Mel filter.

In different audios, the frequency domain information of the signal will be different, which will affect the processing of the signal by the FFT and Mel filter. Both spectral information and power are changed, which in turn affects the final MFCC eigenvalues. The feature value is also discriminative for different pitch changes, so it is considered that MFCC can be used to distinguish the features of different pitches. At the same time, the correlation coefficient matrix of MFCC features is introduced. Assuming that the MFCC feature dimension is L, the correlation coefficient is calculated for the components of the jth dimension and the j'th dimension, where $1 \leq j \leq j' \leq L$, the calculation formula of the correlation coefficient matrix feature is

$$\mathrm{CR}_{jj'} = \frac{\mathrm{cov}(V_j, V_{j'})}{\sqrt{\mathrm{VAR}(V_j)} \times \sqrt{\mathrm{VAR}(V_{j'})}}, \tag{2}$$

where V_j is all components of the jth dimension, cov is the covariance operation, VAR is the variance operation, and $\mathrm{CR}_{jj'}$ is the correlation coefficient between the jth dimension and the j'th dimension components.

The features used in the algorithm fuse the relative spectrum-aware linear prediction features, MFCC features, and their correlation coefficient matrix, and the fused features can be expressed as

$$F = [F_{\mathrm{R-PLP}}, C_{\mathrm{L}}, \mathrm{CR}_{\mathrm{MFCC}}], \tag{3}$$

where $F_{\mathrm{R-PLP}}$ is the RASTA-PLP characteristic coefficient, C_{L} is the MFCC characteristic, and $\mathrm{CR}_{\mathrm{MFCC}}$ is the MFCC characteristic correlation coefficient. The splicing and fusion of these features is F, which constitutes the input feature information of the detection algorithm.

At the same time, the first-order difference coefficient and second-order difference coefficient of the feature are also calculated during feature extraction, but the feature and difference coefficient values are not spliced in series but mapped to the RGB three channels of the feature image as the input of the network information.

3.2 Network structure

The network structure used in this article is the improved Capsule Net structure. As early as 2011, the concept of the capsule took shape, and then the concept was expanded and optimized to form a capsule network, and various variants of the capsule network have emerged since the development. Unlike CNNs, which have been pivotal across various tasks, Capsule Networks address some of CNN's inherent limitations. CNNs rely on pooling and convolution operations for feature extraction, progressing from superficial to deeper layers and summarizing local features. This process, however, often leads to the loss of crucial feature information due to convolution and pooling, potentially omitting details vital to the target task. Additionally, while CNNs streamline model parameters and computational demands, this simplification might compromise performance, as CNNs struggle to learn the relative positions within input data, focusing merely on pattern presence.

Contrastingly, Capsule Networks offer a robust solution by representing features as multi-dimensional vectors within capsule structures, unlike CNNs' scalar outputs. This vector representation allows the network to capture and maintain detailed spatial hierarchies and relationships, offering a richer understanding of the input data. The magnitude of a capsule's output vector signifies the presence probability of the learned feature, with values near 1 indicating near certainty. This nuanced approach enables Capsule Networks to preserve and utilize critical information throughout the learning process, facilitating a comprehensive and precise interpretation of complex data structures, markedly enhancing the model's performance and applicability to a broader range of tasks.

The initial capsule network structure differs from the CNN network in the design of a single capsule structure, routing algorithm, and activation function. The calculation process of a single capsule structure can be simplified into the following process: First, multiply the input, and the input $\boldsymbol{v}$ of the capsule is the output of the previous capsule. Second, multiply $\boldsymbol{w}$ and the input $\boldsymbol{v}$ to obtain $\boldsymbol{u}$, and weigh the vector, that is, each weight multiply c and $\boldsymbol{u}$, respectively, where c is a scalar. Sum the weighted vectors again, denoted as $\boldsymbol{s}$. Finally, input it into the activation function; the activation function of the capsule network is the Squashing nonlinear function, which is defined as

$$\mathrm{Squashing}(s) = \frac{||s||^2}{||s||^2 + 1} \times \frac{s}{||s||}, \tag{4}$$

where the first fraction on the right side of the equal sign represents the scaling of $\boldsymbol{s}$, and the second fraction can be regarded as a unit vector in the same direction of $\boldsymbol{s}$. Through the *Squashing* function, the direction information of s can be well preserved, and the length is guaranteed to be between 0 and 1, so as to realize the compression and redistribution of the vector and reduce the amount of calculation.

During the calculation process, the weight c is updated through the routing algorithm. The sum of the weights c in the capsule layer is 1. In the initial capsule network proposed in 2017, not all of them are capsule layers, and convolutional structures also exist in shallow structures. The shallow convolutional structure extracts low-dimensional features in the input information and sends them to the Primary Caps Layer and Digit Caps Layer, which extracts deep features. There is also a convolution structure in the main capsule layer.

The output of the main capsule layer will be sent to the digital capsule layer. The digital capsule layer is a fully connected structure. The output represents the classification result, and the output is in the form of a vector. The input can be judged by the output modulus length. Whether there is a pattern in the information, does it belong to a certain category.

The initial design of the capsule network, with its reliance on shallow convolution layers, often fell short in adequately representing the intricacies of front-end extracted features. This limitation became particularly apparent when dealing with complex inputs and classification scenarios, where the algorithm's performance noticeably lagged. Enhancements to the convolutional layer structure within the capsule network have been proposed to deepen the analysis of feature information at the input stage. Such modifications aim to augment the capsule network's sensitivity to nuanced input variations, enabling a more refined distinction between different tonal tasks. This approach not only seeks to elevate the overall algorithmic performance but also to equip the capsule network with a greater adaptability to diverse inputs. Figure 1 illustrates the overall architecture of the algorithm's network. Initially, the input feature image, referred to as fusion feature F, comprised three channels. However, prior to network submission, it underwent preprocessing to be transformed into a grayscale image. The conversion process is

$$L = G \times \frac{587}{1{,}000} + R \times \frac{299}{1{,}000} + B \times \frac{114}{1{,}000}, \tag{5}$$

where R, G, and B represent the pixel values in the RGB 3 channels. The information in the three channels is reserved, the R channel represents static feature information, and the G and B channels represent dynamic feature information.

Therefore, the actual size of the input training data is (224, 224, 1). After that, the convolution layer is improved, and the size of the convolution kernel of the first convolution layer is set to (9, 9). The output is set to 64 dimensions with a stride of 4 and padding of 2. The activation function is ReLU, and the pooling function is an adaptive average pooling function. This function can automatically adjust the output to the dimensions specified in the program. In the experiment, the output setting of the pooling layer is specified as (28, 28, 64) and sent to the subsequent convolutional layer.

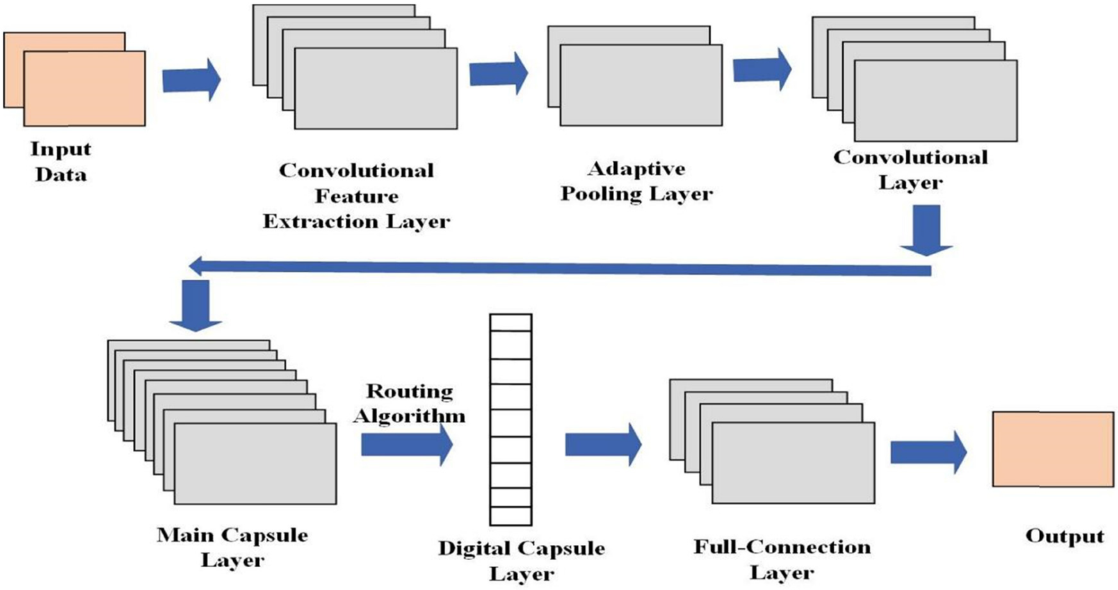

Figure 1: The overall architecture of the proposed network model.

4 Experimental evaluations

In order to verify whether the proposed method can effectively detect and recognize Guzheng tones, the method is compared with the existing work. The experimental data are derived from the self-built Guzheng music database. The database includes 46 pure music Guzheng pieces sourced from various audio websites and video websites on the Internet, and digital spectrum tones are extracted from the original music.

The feature used in the experiment is the fusion feature F. The static and dynamic features of F are mapped to RGB three channels, but the feature image is grayscaled at the input end of the network, converted into a single-channel input image, and the feature information is preserved in another way. During model training, the optimizer selects Adam, the learning rate is initially set to 0.0001, and the decay is set to 0.98.

As one of the oldest musical instruments in China, the guzheng has only five notes (do, re, mi, sol, la), but there are many modes, usually D. Guzheng has a total of 21 strings, including 5 strings for bass, bass, middle and treble, and one string for treble. In this section, we examine the accuracy of the proposed method in recognizing these 21 tones. Table 1 presents the recognition accuracy for various Guzheng tones, demonstrating the effectiveness of the proposed method across a range of musical alphabets. Notably, tones such as $\boldsymbol{a^{2}}$, $\boldsymbol{g^{1}}$, $\boldsymbol{e^{1}}$, a, G, E, and D achieved a recognition rate of 100%, indicating exceptional performance of the algorithm in these instances (it should be noted that 100% does not mean that our method can recognize the tone completely and accurately, but only means that in the data set used in this article achieve this accuracy). High accuracy levels were also observed for tones like $\boldsymbol{c^{3}}$ and $\boldsymbol{g^{2}}$, with accuracies of 99.50 and 99.70%, respectively, showcasing the method's robustness in identifying the majority of tones with great precision. However, the method exhibited lower accuracy for certain tones, specifically $\boldsymbol{d^{3}}$ and $\boldsymbol{c^{2}}$, with recognition rates of 89.30 and 88.10%, respectively, highlighting areas where the algorithm's performance could be further improved. Despite these variations, the method achieved an impressive overall average recognition rate of 98.15%.

First, the recognition accuracy of different network models under noise-free scenarios is given in Figure 2. Our approach manifests superior performance with an accuracy of over 97%, marginally eclipsing the ResNet mode. The CapsNet-base architecture, while demonstrating a respectable level of accuracy, falls behind the aforementioned models. GoogLeNet registers a modest accuracy figure, placing it at the lower spectrum of the evaluated models. At the same time, in order to verify the anti-interference effect of the algorithm proposed in this article, a set of comparative experiments are set up. In this group of experiments, the music is subjected to noise and compression processing. Noise comes from the SPIB dataset, where speaker noise, white noise, vehicle noise, pink noise, and communication channel noise are selected. To test the performance of the algorithm in unknown noise scenarios, vehicle and communication channel noises are only included in the test set. The rest of the noise types are added in both the training set and the test set. Compression compresses files into MP3 files with bit rates of 64, 128, and 192 kbit/s. In the anti-compression experiment, the situation of bit rate reduction and false high is considered, and both the training set and the test set are composed of three compressed bit rates, and the three compressed bit rates are not tested separately.

The experiment compares the improved capsule network model with the classical classification models GoogLeNet, ResNet, and the initial capsule network structure. Figure 3 displays the outcomes of the

Table 1: Recognition rate of different tones of Guzheng

Musical alphabet	Accuracy (%)	Musical alphabet	Accuracy (%)	Musical alphabet	Accuracy (%)
$\boldsymbol{d^{3}}$	89.30	$\boldsymbol{c^{3}}$	99.50	$\boldsymbol{a^{2}}$	100
$\boldsymbol{g^{2}}$	99.70	$\boldsymbol{e^{2}}$	98.90	$\boldsymbol{d^{2}}$	99.10
$\boldsymbol{c^{2}}$	88.10	$\boldsymbol{a^{1}}$	97.90	$\boldsymbol{g^{1}}$	100
$\boldsymbol{e^{1}}$	100	$\boldsymbol{d^{1}}$	98.70	$\boldsymbol{c^{1}}$	95.40
a	100	**g**	99.60	**e**	98.90
d	98.90	**c**	99.50	**A**	97.60
G	100	**E**	100	**D**	100

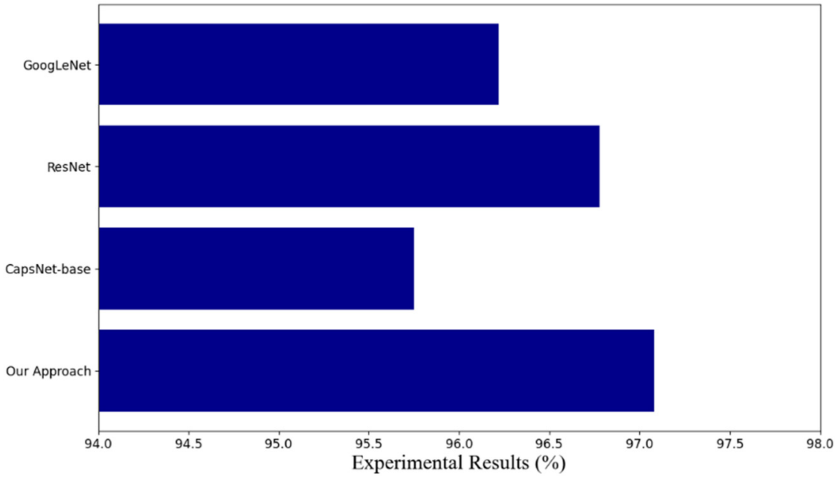

Figure 2: Experimental results of compared methods under noise-free scenario.

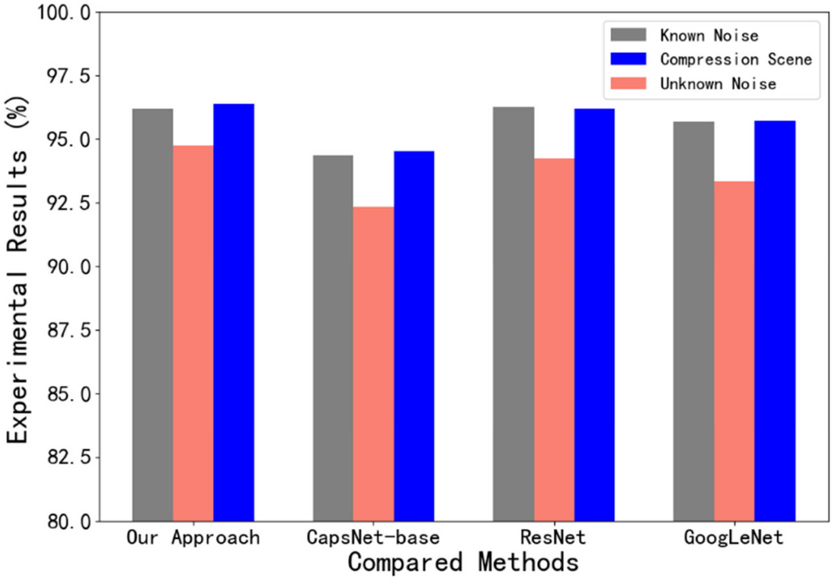

Figure 3: Experimental results of compared methods on the effect of anti-interference.

experiments, illustrating that the recognition rates across different methods are comparably close in scenarios with known noise and compression. However, in the presence of unknown noise, all four methods exhibit noticeable performance declines. Among them, CapsNet-base shows the least effective performance. In comparison, GoogLeNet marks a significant enhancement, outperforming CapsNet-base by 1.3% in known noise conditions. ResNet demonstrates further improvement. Relative to ResNet, our method achieves similar results in known noise environments but surpasses it by 0.54% in unknown noise conditions and by approximately 0.2% in compression situations. Despite a 1.96% reduction in the average recognition rate under noise compared to noise-free conditions, our method still secures an accuracy rate exceeding 96%. These findings affirm that our approach possesses robust anti-interference capabilities, enabling more precise Guzheng tone recognition amidst disturbances.

5 Conclusion

Tone recognition is an important part of musical tone recognition. Music recognition technology offers a wide range of application scenarios, and its development and application have shown a trend of "popularization" and "specialization." Guzheng is one of the famous classical musical instruments in China, and it is of great value to use advanced AI technology to automatically identify the tones of the Guzheng. Capsule networks, with their innovative design, excel at maintaining the hierarchical relationships and spatial features of objects, while their dynamic routing mechanism boosts the model's efficiency in processing and prioritizing relevant features. This approach not only markedly improves Guzheng tone recognition performance but also offers superior robustness against interference and noise, setting a new standard in music recognition technologies. Aiming at this problem, this article combines RASTA-PLP features and MFCC-related features and improves the capsule network structure for Guzheng tone detection and recognition. This article evaluates our method by crawling Guzheng music from the Internet and creating a data set. The database includes 46 pure music Guzheng pieces crawled from various audio websites and video websites on the Internet, and digital spectrum tones are extracted from the original music. This article first evaluates the accuracy of our method in recognizing 21 tones of Guzheng and shows the recognition results on different tones. The experimental results show that the improved capsule network proposed in this article has good performance. Furthermore, in order to evaluate the anti-interference ability of our method, comparative experiments were designed. By comparing experiments with other algorithms, it can be seen that the algorithm designed in this article can maintain good robustness in the face of common interference factors such as noise and compression.

Despite the innovative approach to guzheng tone recognition presented, achieving an average accuracy of 98.15%, opportunities for improvement remain. The aspiration for perfection in machine recognition of musical tones necessitates further refinement. The performance is affected by environmental factors and variability in playing techniques, such as plucking force, speed, and angle, which impact accuracy. Future efforts should refine the algorithm to better accommodate these variations through adaptive models that dynamically respond to different conditions. Expanding the dataset with a broader array of playing styles and environmental scenarios, alongside exploring machine learning techniques like transfer learning, could enhance system robustness and accuracy. Moreover, expanding the dataset to include a broader range of playing styles, environmental conditions, and Guzheng models could significantly improve the robustness and generalizability of the recognition system. ML techniques that focus on transfer learning or unsupervised learning might offer pathways to accommodate the vast diversity inherent in musical expression. The exploration of real-time processing capabilities to support live performance analysis would further advance the field of music recognition technology.

Acknowledgments: This work was supported by the Doctoral Research Start-up Project of Xinxiang University under Granted No. 1366020216.

Author contribution: The author confirms the sole responsibility for the conception of the study, presented results and manuscript preparation.

Conflict of interest: Author states no conflict of interest.

References

[1] J. W. S. B. Rayleigh, *The Theory of Sound*, vol. 2, Macmillan & Company, New York, US, 1896.
[2] L. Pierce, *Acoustics*, Springer International Publishing, Cham, 2019.
[3] R. W. Young, *Inharmonicity of plain wire piano strings*, J. Acoust. Soc. Am. **24** (1952), no. 3, 267–273.
[4] F. Rigaud, B. David, and L. Daudet, *A parametric model and estimation techniques for the inharmonicity and tuning of the piano*, J. Acoust. Soc. Am. **133** (2013), no. 5, 3107–3118.

[5] T. D. Rossing and N. H. Fletcher, *Principles of Vibration and Sound*, Springer Science & Business Media, New York, US, 2004.

[6] A. Klapuri and M. Davy, (eds.), *Signal Processing Methods for Music Transcription*, Springer Science & Business Media, New York, US, 2007.

[7] T. Drugman, G. Huybrechts, V. Klimkov, and A. Moinet, *Traditional machine learning for pitch detection*, IEEE Signal. Process. Lett. **25** (2018), no. 11, 1745–1749.

[8] Y. Liu, H. Chen, and B. Wang, *DOA estimation based on CNN for underwater acoustic array*, Appl. Acoust. **172** (2021), 107594.

[9] K. Zhang, W. Wang, Z. Lv, Y. Fan, and Y. Song, *Computer vision detection of foreign objects in coal processing using attention CNN*, Eng. Appl. Artif. Intell. **102** (2021), 104242.

[10] R. Keshari, M. Vatsa, R. Singh, and A. Noore, *Learning structure and strength of CNN filters for small sample size training*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9349–9358.

[11] T. Vu, T. D. Nguyen, D. Q. Nguyen, and D. Phung, *A capsule network-based embedding model for knowledge graph completion and search personalization*. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, June, pp. 2180–2189.

[12] S. Bilbao, *Numerical Sound Synthesis: Finite Difference Schemes and Simulation in Musical Acoustics*, John Wiley & Sons, West Sussex, UK, 2009.

[13] J. Chabassier, A. Chaigne, and P. Joly, *Time domain simulation of a piano. Part 1: model description*, ESAIM: Math. Model. Numer. Anal. **48** (2014), no. 5, 1241–1278.

[14] J. Woodhouse, *The acoustics of a plucked harp string*, J. Sound Vib. **523** (2022), 116669.

[15] A. P. Klapuri, *Multiple fundamental frequency estimation based on harmonicity and spectral smoothness*, IEEE Trans. Speech Audio Process. **11** (2003), no. 6, 804–816.

[16] N. Yang, H. Ba, W. Cai, I. Demirkol, and W. Heinzelman, *BaNa: A noise resilient fundamental frequency detection algorithm for speech and music*, IEEE/ACM Trans. Audio Speech Lang. Process. **22** (2014), no. 12, 1833–1848.

[17] Z. Cui, *Pitch extraction based on weighted autocorrelation function in speech signal processing*. In Proceedings of 2012 2nd International Conference on Computer Science and Network Technology, IEEE, 2012, December, pp. 2158–2162.

[18] J. Dubnowski, R. Schafer, and L. Rabiner, *Real-time digital hardware pitch detector*, IEEE Trans. Acoust. **24** (1976), no. 1, 2–8.

[19] S. Särkkä, *Bayesian Filtering and Smoothing* No. 3. Cambridge University Press, Cambridge, UK, 2013.

[20] G. R. Xue, C. Lin, Q. Yang, W. Xi, H. J. Zeng, Y. Yu, et al., *Scalable collaborative filtering using cluster-based smoothing*, In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005, August, pp. 114–121.

[21] H. Liu, D. Chen, and G. Sun, *Detection of fetal ECG R wave from single-lead abdominal ECG using a combination of RR time-series smoothing and template-matching approach*, IEEE Access **7** (2019), 66633–66643.

[22] J. Guo, *The stability model of piano tone tuning based on ordinary differential equations*, Appl. Math. Nonlinear Sci. **8** (2023), no. 1, 929–936.

[23] W. Yuan-yuan and Y. Shun, *Speech synthesis based on PSOLA algorithm and modified pitch parameters*, In International Conference on Computational Problem-Solving, IEEE, 2010, December, pp. 296–299.

[24] Z. Rao, X. Guan, and J. Teng, *Chord recognition based on temporal correlation support vector machine*, Appl. Sci. **6** (2016), no. 5, 157.

[25] V. Z. Këpuska and H. A. Elharati, *Robust speech recognition system using conventional and hybrid features of MFCC, LPCC, PLP, RASTA-PLP and hidden Markov model classifier in noisy conditions*, J. Comput. Commun. **3** (2015), no. 6, 1.

[26] H. Hermansky and N. Morgan, *RASTA processing of speech*, IEEE Trans. Speech Audio Process. **2** (1994), no. 4, 578–589.