CNNs in Musical Performance and Arrangement: Recognizing and Managing Bowed Instrument Techniques Across Cultures

Xinyuan Zhu
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
email: xinyuanzhu@link.cuhk.edu.cn

Clement Leung
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
email: clementleung@cuhk.edu.cn

***Abstract*—This study explores the application of AI-assisted techniques in analyzing and classifying bowed string instruments from Chinese and Western traditions, focusing on the comparison between the erhu and the violin. Using a combination of spectrogram analysis, Mel-Frequency Cepstral Coefficients (MFCCs), and Convolutional Neural Networks (CNNs), the study captures the distinct timbral and articulation differences between the two instruments. Particular attention is given to bowing techniques such as vibrato, portamento, pizzicato, which manifest differently due to structural and acoustic variations. Beyond recognition, this research contributes to AI-assisted music arrangement and composition, providing tools to analyze and synthesize playing techniques across different musical traditions. By bridging Eastern and Western bowed instrument performance styles, this approach supports both cultural heritage preservation and innovation in contemporary music production.**

***Keywords-AI-assisted music creation*; *Audio fingerprinting*; *Spectrogram matching*; *Convolutional neural networks*; *Audio signal processing*.**

I. Introduction

Music creation and arrangement rely heavily on accurate music recognition technologies, which facilitate the identification and integration of musical elements in various production contexts. Numerous techniques have emerged in the field of music recognition, significantly improving music retrieval, automated generation, and media management [1]. Central to these developments is audio fingerprinting, which segments an audio signal into small time windows and transforms them into frequency domain representations via Fourier analysis [2]–[4] . This method allows for the extraction of distinctive "fingerprints" based on unique spectral characteristics, which are used to match audio tracks across platforms like Shazam and Echoprint [5]. Complementing this, spectrogram matching uses visual representations of sound and relies on CNNs to detect patterns in these spectrograms, making it possible to classify music even in noisy environments [6] [7]. Additionally, rhythm and chord matching further enriches the recognition process by analyzing temporal features and harmonic progressions within a track, helping identify musical patterns and styles [8]. Lastly, lyrics matching extends the scope of music recognition by using Natural Language Processing (NLP) to transform sung vocals into text [9]. This text is then matched against a lyrics database to identify songs, making it particularly useful for recognizing cover versions or different renditions of the same song, where the melody might differ but the lyrics remain consistent.

These methodologies, each using unique aspects of audio processing and analysis, highlight the complexity and dynamic nature of music recognition technology. They not only improve the accuracy and efficiency of music identification but also enrich the user experience across various digital platforms, paving the way for innovative applications in the music and media industries.

Although these techniques have greatly enhanced our ability to identify and categorize music, current research predominantly focuses on the design and technological recognition of a certain musical instrument [10] [11], exploring innovative computational methods and digital fabrication that enable musical instrument identification and song matching. However, there remains a notable gap in the research specifically targeting the detailed identification of playing techniques on single instruments, especially within the domain of traditional Chinese music. This is a critical area for music learning and cataloging, as understanding and preserving the unique playing techniques of traditional music not only plays a crucial role in cultural heritage and education but also greatly enhances music creation and arrangement by providing deeper insights into the expressive potential of these instruments [12]. Furthermore, there's a need for developing technologies that can assist in the nuanced detection of specific musical styles and techniques, which are often overlooked in broader music recognition systems [13]. Thus, this paper presents a novel deep-learning model designed to recognize various playing techniques of two bowed string instruments from different musical traditions—the Western Violin and the Chinese Erhu. Section 2 details the theoretical foundations of audio processing and CNNs. Section 3 describes the system architecture and implementation. Section 4 presents experimental results for Violin and Erhu techniques. Section 5 compares the bowing techniques between the two instruments. Section 6 discusses implications and future work.

II. Theory

This section introduces the theoretical principles underlying the signal processing and deep learning operations implemented in the system. For different playing techniques, such as focusing on pitch and playing frequency, we employ various

approaches to handle pitch and playing frequency to ensure the most suitable solution for a specific technique. These techniques are described below.

1) Audio Signal Processing:

a) Pitch Shifting: Pitch shifting alters the frequency content of the audio without changing the tempo. Here, the Short-Time Fourier Transform is expressed as STFT. It can be expressed using the phase vocoder approach in the frequency domain:

$$\text{pitch\_shifted\_data} = \text{STFT}^{-1}(\text{STFT}(\text{data}) \cdot e^{j\omega\delta})$$

where δ denotes the pitch shift in radians, and ω is the angular frequency vector.

b) Audio Stretching: Time-stretching changes the duration of the audio signal. Using the phase vocoder method, the operation is defined as:

$$\text{stretched\_data} = \text{STFT}^{-1}(\text{STFT}(\text{data}) \cdot e^{j\phi})$$

where ϕ is a phase adjustment applied to maintain continuity in the time-stretched signal.

2) Feature Extraction:

a) Mel-Spectrogram: The Mel-spectrogram is computed from the Short-Time Fourier Transform (STFT) of the signal, mapped onto the Mel scale:

$$\text{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

$$\text{mel\_spec} = |\text{STFT}(\text{data})|^2$$

where f is the frequency in Hz.

b) MFCCs: Mel Frequency Cepstral Coefficients (MFCCs) are derived from the logarithm of the Mel-spectrogram, followed by a Discrete Cosine Transform (DCT):

$$\text{MFCCs} = \text{DCT}(\log(\text{mel\_spec}))$$

3) Convolutional Neural Networks:

a) Convolutional Layer: A convolutional layer in a CNN can be mathematically modeled as:

$$\text{output} = \sigma(\mathbf{W} * \mathbf{X} + \mathbf{b})$$

where $\mathbf{W}$ represents the kernel weights, $\mathbf{X}$ is the input, $\mathbf{b}$ is the bias, and σ is a nonlinear activation function, such as ReLU defined by $\sigma(x) = \max(0, x)$.

b) Pooling Layer: Pooling layers reduce the spatial dimensions of the input feature maps:

$$\text{pooled} = \max_{k,l}(\text{input}[i + k, j + l])$$

for max pooling over the window defined by indices k and l.

This detailed theoretical foundation ensures the robustness and comprehensiveness of the specifications, guiding the practical implementation of the proposed audio processing and CNN methodologies.

III. Architecture and Details

This section elaborates on the system architecture and implementation specifics designed for the project, including algorithm selection, software development practices, and Python programming techniques utilized. Each component's functionality and its integration within the system are clarified for thorough understanding.

A. System Architecture

The system is structured around several key functionalities which include:

- **Data Preprocessing:** Initial steps such as normalization, noise reduction, and data augmentation are applied to the audio data to enhance model performance and robustness. In addition, we cut training pieces to be 3 seconds long each, labling them with 1 if they have certain feature and 0 if they don't have. Each playing technique is associated with a separate dataset comprising approximately 500 audio segments, with 70% allocated for training, 15% for validation, and 15% for testing.
- **Feature Extraction:** Utilizing the `librosa` library, features such as MFCCs, spectral contrast, and tonnetz are extracted, which are crucial for the audio signal analysis.
- **Convolutional Neural Network (CNN):** The model employs CNN architectures, implemented using TensorFlow and Keras, to process and classify audio data effectively.

B. Implementation Details

1) Algorithm Selection: The project employs CNNs due to their effectiveness in audio and image processing tasks. CNNs are chosen for their ability to identify hierarchical patterns in data, which is essential for the analysis of complex audio signals.

2) Software Development: Python is selected as the main programming language, supported by its extensive libraries and frameworks that facilitate the implementation of data science and machine learning algorithms efficiently.

3) Frameworks and Libraries: Key frameworks and libraries used in the project include:

- **TensorFlow and Keras:** For designing, training, and validating deep learning models. These frameworks offer comprehensive tools that aid in the rapid development and deployment of ML models.
- **Librosa:** A library for music and audio analysis, providing the necessary functionalities to implement music information retrieval systems.

4) Model Development: The model takes features from Mel-spectrograms and MFCCs as input, represented as 2D time–frequency matrices. It consists of three convolutional layers with 3×3 kernels, each followed by max-pooling. The model was trained for 100 epochs with a batch size of 32. A final dense layer with softmax activation handles classification. Training uses categorical cross-entropy loss and the Adam optimizer.

C. Instrument Introduction

This study explores two bowed string instruments from different musical traditions: the Violin, a cornerstone of Western classical musicand the Erhu, a representative Chinese traditional instrument. While both instruments share similarities in their bowed playing technique, they exhibit distinct structural, tonal, and expressive differences.

1) Violin: The Violin is a Western bowed string instrument that has been a central part of orchestral, chamber, and solo music for centuries as shown in Figure 1(a). It typically has four strings tuned in perfect fifths (G-D-A-E) and is played with a horsehair bow. Unlike the Erhu, the Violin has a fingerboard, allowing for precise pitch control and a broader range of fingering techniques. It is known for its brilliant and resonant tone, capable of a wide range of expressive dynamics, from delicate pianissimo to powerful fortissimo.

2) Erhu: The Erhu, as shown in Figure 1(b), is a traditional bowed instrument with a distinctive timbre that is often described as mimicking the human voice. Unlike the Violin, the Erhu has only two strings, tuned a perfect fifth apart, and lacks a fingerboard, which allows for continuous gliding motions, producing characteristic portamento effects. The bow is positioned between the two strings, requiring a unique bowing technique where the player alternates between inner and outer strings. The Erhu's sound is softer and more nasal compared to the Violin, and it is widely used in Chinese folk, traditional, and contemporary music.

By comparing these two bowed string instruments, this study aims to highlight the unique playing techniques, timbral qualities, and expressive characteristics that differentiate Chinese traditional and Western classical music traditions.

(a) Violin (b) Erhu

Figure 1: The Violin and Erhu

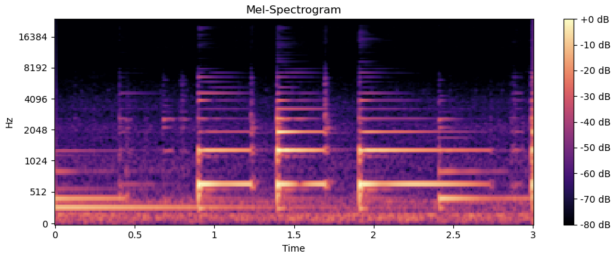
IV. Results

This section focuses on the detection and analysis of four essential playing techniques for the violin: portamento, pizzicato, vibrato, and chords, as well as three techniques for the erhu: pizzicato, portamento, and horse neighing. Detailed explanations and visual representations of each technique are provided in the following subsections.

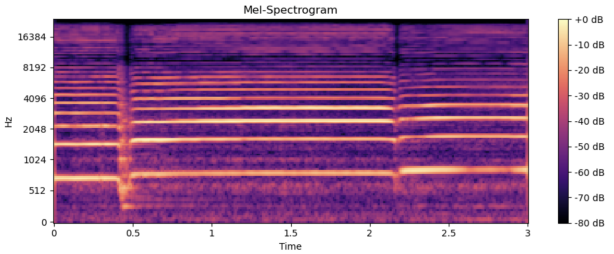
For the full code, training pieces, and testing cases, please refer to the following GitHub and Google Drive repositories: GitHub Repository, Google Drive Repository.

A. Violin pizzicato

The violin's pizzicato showcases its versatility and dynamic articulation. By plucking the strings instead of bowing, it produces crisp, percussive tones that range from delicate to forceful. This technique adds rhythmic clarity and timbral variety, enriching both classical and contemporary compositions. Beyond its technical role, pizzicato enhances expressive depth, allowing performers to craft playful, agile, or dramatic effects, highlighting the violin's adaptability across musical genres.



(a) Pizzicato spectrogram



(b) Spectrogram without pizzicato

Figure 2: Comparison of violin pizzicato audio spectrograms

Figure 2(a) showcases pizzicato, the core distinguishing feature is the discrete, rapidly decaying frequency components, evident in the short, isolated horizontal bands. Each note exhibits a sharp attack followed by a quick fade. The gaps between notes indicate the lack of sustained bow pressure, reinforcing the percussive and transient nature of pizzicato articulation. In contrast, Figure 2(b) shows a violin performance without pizzicato, characterized by continuous and sustained horizontal bands that indicate prolonged bowing. The accuracy under 20 test cases is 100% with a threshold of 0.38.

B. Violin Vibrato

Violin vibrato is an essential and nearly omnipresent technique, appearing in almost every performance. It is created by oscillating the fingertip on the string, producing continuous pitch variations. This enriches the tone, adding warmth, depth, and expressiveness. Vibrato is so frequently used that a note without it often feels unusual in classical violin playing.

This detailed spectrogram Figure 3 vividly illustrates a violin performance incorporating vibrato, which is clearly discernible through the distinct, wavering patterns of the

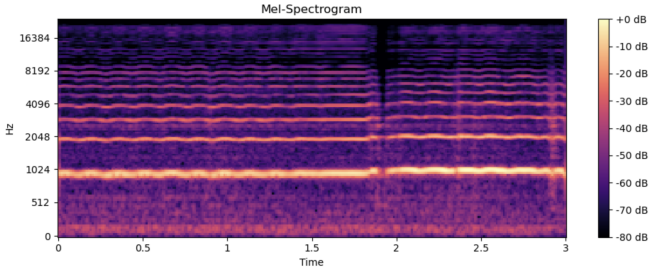


Figure 3: Vibrato spectrogram

frequency bands. These oscillating bands represent the subtle pitch variations characteristic of vibrato. The analysis demonstrates exceptional precision, achieving an accuracy of 100% at a threshold value of 0.18, underscoring the reliability and effectiveness of the method used to detect and quantify this acoustic feature.

C. Violin Portamento

Violin portamento is a smooth sliding technique that connects two notes seamlessly. It is produced by gliding the finger along the string while maintaining contact, creating a continuous pitch transition. This effect adds expressive fluidity. Portamento is frequently used in both classical and contemporary music to enhance emotional depth, making melodic passages sound more lyrical and connected. Its subtle or exaggerated application depends on stylistic interpretation, shaping the expressiveness of a performance.

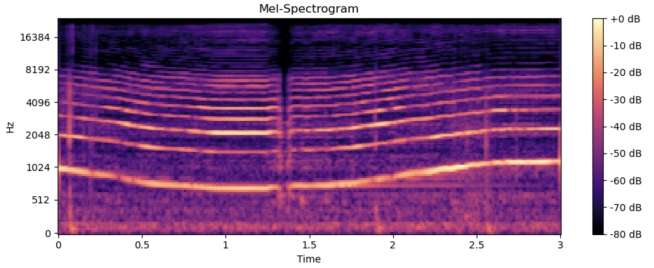


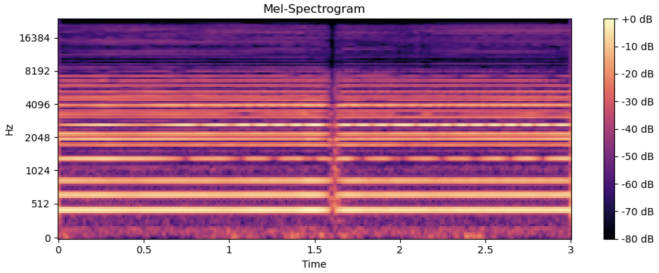
Figure 4: Portamento spectrogram

Figure 4 represents violin with portamento, characterized by smooth, diagonal transitions between frequencies. These sloping lines indicate a continuous pitch glide, as the player's finger slides between notes without discrete separation. The accuracy is 100 % at the threshold of 0.07.

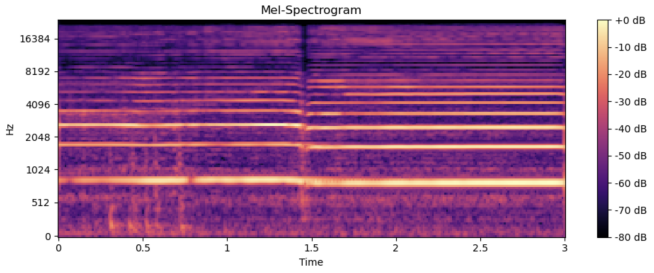
D. Violin Chords

Violin chords involve playing multiple strings simultaneously, creating a rich harmonic texture. They can be performed as double stops, where two notes are played together, or as triple/quadruple stops, where three or four strings are struck in succession. This technique adds depth, power, and resonance, commonly found in orchestral, solo, and folk music for dramatic or harmonic emphasis.

Figure 5(a) represents a series of chords, as clearly evidenced by the densely packed horizontal bands that span



(a) Chords spectrogram



(b) Spectrogram without chords

Figure 5: Comparison of chord audio spectrograms

across multiple frequency ranges. These bands indicate the simultaneous vibration of multiple strings, each contributing to the overall harmonic structure. This dense clustering of frequencies is a hallmark of polyphonic music, where multiple pitches are sounded together to form harmonies. In contrast, Figure 5(b), which represents a violin playing without chords, displays a markedly different pattern. Here, the horizontal bands are fewer in number and more evenly spaced, reflecting the individual notes being played in a monophonic manner. Each band corresponds to a single pitch, with the spacing between them indicating the intervals of the melody. Figure 6 below shows the model accuracy for detecting chords on violin.

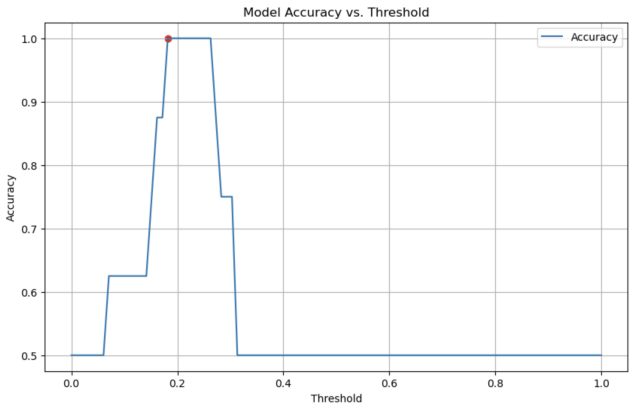
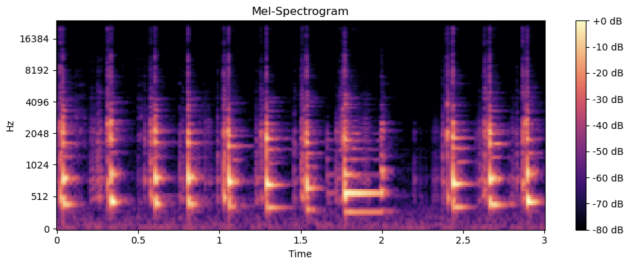


Figure 6: Model effectiveness in detecting chords on violin.

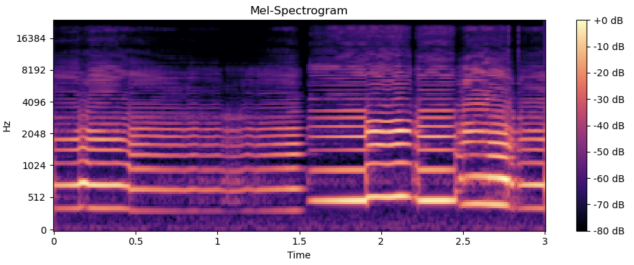
E. Erhu Pizzicato

The pizzicato technique on the erhu exemplifies the instrument's versatility and expressive range. This technique

involves plucking the strings with the fingers, rather than using the bow, producing a sharp, percussive sound that contrasts with the typical bowed tones of the instrument. The pizzicato's unique timbre and rhythmic precision make it an effective tool for creating lively, staccato passages that add dynamic contrast to musical phrases. This technique's ability to produce clear, articulated notes with a distinct percussive quality allows performers to inject a new layer of expression into their music, enhancing the emotional depth and rhythmic complexity of a piece.



(a) Pizzicato spectrogram



(b) Spectrogram without pizzicato

Figure 7: Comparison of erhu pizzicato audio spectrograms

In Figure 7(a), the characteristics of a pizzicato played on the erhu are visually evident. The defining feature is the presence of sharp, distinct vertical lines that appear at regular intervals. These lines represent the short, percussive nature of the pizzicato notes, which decay quickly and lack sustained resonance. In contrast, Figure 7(b) lacks these pizzicato features. Instead, it displays smoother, more continuous horizontal bands, indicative of sustained, bowed notes typical of traditional erhu playing. Here, we employ 200 test pieces to evaluate the model. The highest accuracy achieved is 98.02%, with the best threshold set at 0.578. Here, we demonstrate the ROC curve as shown in Figure 8. For this model, the precision is 0.9756, the recall is 0.9877, the F1-score is 0.9816, and the AUC score is 0.9850.

F. Erhu Portamento

The portamento technique on the erhu showcases the instrument's ability to convey emotional depth and seamless movement between notes. This technique involves sliding the pitch between two notes, creating a smooth, continuous transition rather than a distinct jump. The erhu's rich, fluid sound is often used to mimic the human voice, allowing the performer to evoke a sense of intimacy and vulnerability. Portamento is frequently employed in both traditional and

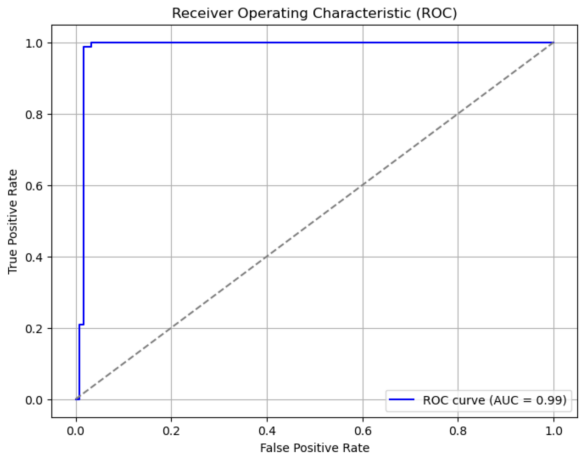


Figure 8: ROC curve for detecting pizzicato on erhu.

contemporary erhu music to enhance the lyrical quality of melodies, providing a sense of narrative flow and emotional continuity.

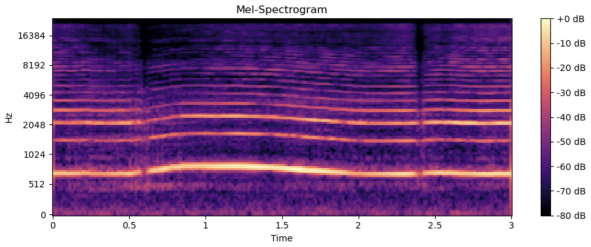


Figure 9: Portamento spectrogram

In Figure 9, the defining characteristics of portamento on the erhu can be observed. The smooth, continuous transitions between frequencies are evident, marked by sloped, connected lines that represent the sliding motion of the player's fingers along the string. This creates a fluid, expressive sound with gradual pitch changes.

As the portamento effect on the Erhu is not always distinct, and normal bowing can sometimes produce portamento-like characteristics, the accuracy is 62.8% at a threshold of 0.25.

G. Erhu Horse Neighing

The horse neighing technique on the erhu is a distinctive and evocative expression of the instrument's ability to mimic natural sounds. This technique involves a combination of fast bowing, specific finger pressure, and sliding motions that produce a sound resembling the neighing of a horse. The erhu's two strings and the player's control over bowing speed and intensity allow for the creation of sharp, high-pitched sounds that imitate the rhythm and tone of a horse's whinny.

In Figure 10, the characteristics of the "horse neighing" sound on the erhu are evident. The texture is dense and irregular, with rapid, fluctuating frequency patterns that create a chaotic and dynamic visual representation. These features correspond to the high-pitched, vibrating sound that mimics

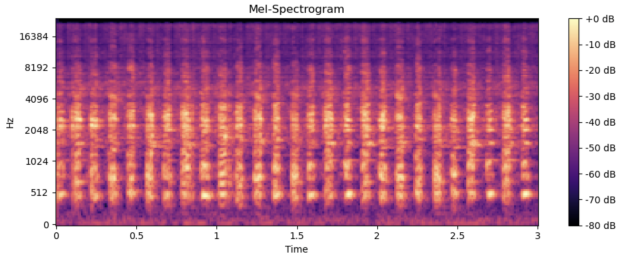


Figure 10: Horse neighing spectrogram

the neighing of a horse, achieved by rapid bowing and finger movements.

Here, we employ 180 test pieces to evaluate the model. The accuracy and the optimal identification threshold are presented below. The highest accuracy comes to 100%, with the best threshold set at 0.03.

V. Comparison of Bowing Techniques

The erhu and violin, as representatives of Chinese and Western bowed string instruments, exhibit distinct differences in playing techniques, articulation, and spectral characteristics. These differences arise from their structural design, musical traditions, and expressive focus, which are clearly reflected in their Mel-spectrograms.

The violin, with its four-string setup and fingerboard, allows for precise articulation, harmonic richness, and diverse bowing techniques. The spectrograms of violin performances show dense harmonic overtones, sustained resonance, and well-defined pitch transitions. In contrast, the erhu's two-string, fretless design enables continuous pitch glides, broader vibrato, and unique timbral effects, as reflected in its spectrogram. The absence of a fingerboard results in smoother portamento, appearing as gradual frequency slopes. Erhu vibrato is broader and more fluid, leading to a more expressive but less structured modulation compared to the violin.

While both instruments share core bowing techniques, such as vibrato, portamento, and pizzicato, their execution differs significantly. The violin excels in precision, harmonic layering, and articulation control, while the erhu prioritizes expressive fluidity, dynamic phrasing, and microtonal variation. By combining deep learning with spectral analysis, this study effectively distinguishes Chinese and Western bowed instrument performance styles, demonstrating the potential of AI in capturing nuanced musical expression.

VI. Conclusion and Future Work

This study applied AI-assisted techniques to analyze and classify bowing techniques of two culturally distinct bowed string instruments—the erhu and the violin. The experiments focused on detecting and analyzing four essential violin techniques—portamento, pizzicato, vibrato, and chords—as well as three primary erhu techniques—portamento, pizzicato, and horse neighing. The spectrogram analysis revealed distinct spectral patterns for each technique, demonstrating how structural differences between the instruments affect their articulation and sound production. These classification results could support AI-assisted music arrangement by automatically annotating performance techniques, enabling intelligent audio mixing and digital orchestration.

Beyond technical classification, this study provides valuable insights into the contrasts between Chinese and Western bowed instruments, bridging traditional musicology with computational analysis. The findings contribute to both cultural heritage preservation and modern AI-assisted music composition, offering potential applications in automated music arrangement, interactive music synthesis, and digital instrument modeling. In addition, while this study focuses on violin and erhu, the approach can be extended to other bowed instruments, subject to retraining on appropriately labeled datasets.

Future work will focus on expanding the dataset to include additional bowing techniques and other western and chinese instruments, refining model accuracy through advanced neural network architectures, and exploring real-time performance analysis. In addition, future work will compare CNNs with RNNs and transformer-based architectures to explore their effectiveness on time-series classification. By integrating AI with musicology, this research paves the way for a deeper understanding of global musical traditions and their computational representations.

References

[1] C. H. Chen, Ed., "Handbook of pattern recognition and computer vision," World Scientific, 2015.

[2] D. P. W. Ellis, B. Whitman, T. Jehan, and P. Lamere, "The Echo Nest musical fingerprint," in Proc. 2010 Int. Symp. Music Information Retrieval, 2010.

[3] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system with an efficient search strategy," Journal of New Music Research, vol. 32, no. 2, pp. 211-221, 2003.

[4] A. Wang, "An industrial strength audio search algorithm," in Int. Conf. Music Information Retrieval (ISMIR), 2003.

[5] D. P. Ellis, B. Whitman, and A. Porter, "Echoprint: An open music identification service," 2011.

[6] M. Young, "The Technical Writer's Handbook," Mill Valley, CA: University Science, 1989.

[7] D. Williams, A. Pooransingh, and J. Saitoo, "Efficient music identification using ORB descriptors of the spectrogram image," EURASIP Journal on Audio, Speech, and Music Processing, pp. 1-17, 2017.

[8] A. Shenoy and Y. Wang, "Key, chord, and rhythm tracking of popular music recordings," Computer Music Journal, vol. 29, no. 3, pp. 75-86, 2005.

[9] Z. Guo, Q. Wang, G. Liu, J. Guo, and Y. Lu, "A music retrieval system using melody and lyric," in Proc. 2012 IEEE Int. Conf. Multimedia and Expo Workshops, pp. 343-348, 2012.

[10] R. C. Rujia, A. Ghobakhlou, and A. Narayanan, "Musical instrument recognition in polyphonic audio through convolutional neural networks and spectrograms," 2024.

[11] G. A. V. M. Giri and M. L. Radhitya, "Musical instrument classification using audio features and convolutional neural network," Journal of Applied Informatics and Computing, vol. 8, no. 1, pp. 226-234, 2024.

[12] R. Michon, O. S. Julius, M. Wright, C. Chafe, J. Granzow, and G. Wang, "Mobile music, sensors, physical modeling, and digital fabrication: Articulating the augmented mobile instrument," Appl. Sci., vol. 7, no. 12, pp. 1311, 2017.

[13] A. Acquilino and G. Scavone, "Current state and future directions of technologies for music instrument pedagogy," Frontiers in Psychology, vol. 13, 2022, https://doi.org/10.3389/fpsyg.2022.835609.