Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

A Review of the Research on the Combination of Artificial Intelligence and Instrument Recognition

Cenyue Yao *

College of Letters and Sciences, University of California, Santa Barbara, America

* Corresponding author: cenyue@ucsb.edu

Abstract. Artificial intelligence has great potential value in the music industry. From the perspective of music recognition based on machine learning, there are numerous possibilities in the industry of music and computer science intersection, music composition and generation; music analysis and preprocessing; music education and interaction, etc. After elaborating on music analysis and preprocessing, a potent about musical instrument recognition can be done with artificial intelligence have discovered. By clarifying previous research combining traditional and modern music features extraction methods like mel-frequency cepstral coefficients (MFCC) and attention mechanism, proposed based on music theory, with deep learning models like convolutional neural network (CNN), artificial neural network (ANN), and Keras for musical instrument recognition. By giving suggestions in this article, musical instrument recognition models' efficacy can become more precise. Therefore, in the future, researchers can have the possibility to dig into deeper and more branches of the intersection industry, resulting in more unexpected value.

Keywords: Artificial intelligence; Audio processing; Musical instrument recognition.

1. Introduction

With the advancement of technology, various branches have appeared in computer science, such as software engineering, computer networks, artificial intelligence, computer graphics, data science, etc. In particular, in recent years, artificial intelligence has attracted lots of attention and also emerged many branches, such as machine learning, deep learning, natural language processing, etc... Hence, researchers based on artificial intelligence's robust capability of processing and learning are trying to use AI models to do the processing and research about audio, images, videos, etc….

Music-related research requires the audio processing procedure. Thus, people who engage in music-associated work may take advantage of musical instrument recognition tools for composition, education, and commercialization purposes. Specifically, for instance, music composers could use recognized instruments and processed audio to be the bedrock of a music track. However, to construct a musical instrument recognition tool, researchers are better familiar with knowledge of both fields of study, music theory for the music realm, and audio feature extraction methods for the artificial intelligence domain. So, researchers with a deeper understanding of both fields of study may produce a more representative and helpful recognition tool.

Mel-Frequency Cepstral Coefficients (MFCC) is an audio feature extraction method that depends on knowledge about human auditory characteristics in music theory commonly applied in modern research [1]. There are also some not widely used extraction methods based on music theory. PITCH is directly based on pitch points in music theory; Chroma Features is consistent with the concepts of scales and chords in music theory [2]; Constant Q Transform (CQT)is designed based on the distribution of music sound, and can directly obtain the amplitude values of music signs at various note frequencies [3]. The majority of Existing research focuses on the identification of individual and ensemble musical instruments with different complexities for each method. Identifying only a single instrument from a music signal requires three steps to get the final result: 1) audio preprocessing, 2) feature extraction, and 3) classification. Recognizing multiple numbers of instruments from ensemble music signals needs three distinctive procedures: 1) sound source separation, 2) single instrument identification, 3) system integration, and optimization to recognize each musical instrument.

Current research is mostly about identifying single instruments from audio, ensemble audio for recognizing various musical instruments is the minority.

This article is dedicated to analyzing some recent research about musical instrument recognition and giving meaningful suggestions for future researchers to acquire helpful integrated methods.

2. Overview of Feature Extraction Method Research

In the late 20th century and early 21st century, due to the restriction of datasets and computational resources, deep could not be trained effectively. Fortunately, various audio processing algorithms possess different strengths for different scenarios. For example, Mel-Frequency Cepstral Coefficients (MFCC) are good at capturing spectral features in audio feature signals and are robust to noise, making them perform well under noisy conditions [1]; Linear Predictive Coding (LPC) can capture the resonance features in signals effectively, making it good at speech synthesis and recognition; and Short-Time Fourier Transform (STF) can convert time-domain signals into frequency-domain for representations, making the characteristics of signals in time and frequency clearer. There are many existing similar models, giving future researchers inspiration and opportunities for combining research. Specifically, MFCC integrates with many subsequent deep-learning models.

2.1. Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is a common feature used in audio signal processing for audio recognition, speaker identification, gender recognition, etc. The calculation of MFCC requires 5 main steps: Pre-emphasis, Signal framing and windowing, Power spectrum calculated by Discrete Fourier Transform (DFT), Mel-Frequency filter bank, and Discrete Cosine Transform (DCT) [1]. The specific steps of MFCC are shown in Fig. 1.

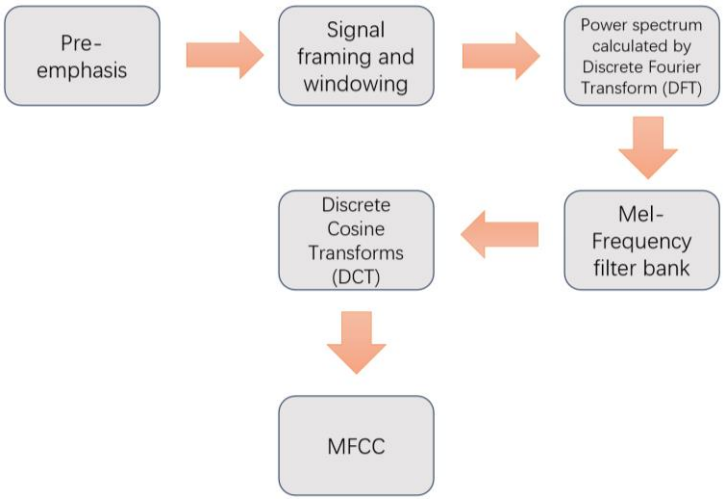

Figure 1. Flow Chart of MFCC

2.1.1. Pre-emphasis.

Pre-emphasis is an ordinary preprocessing method in the field of signal processing, dedicated to compensating for the suppressed high-frequency parts during the signal generation process.

2.1.2. Signal framing and windowing.

Signal framing includes dividing signals into different frames, through examining the sound signal in a steady period to acquire an accurate acoustic feature. Signal windowing normally uses a 20-millisecond window for short-term spectral measurements, with a 10-millisecond overlap between each frame to track the temporal characteristics of the sound signal. Applying a window to each frame helps to concentrate signal at the edges, and for common window types regarding Hanning and Hamming windows, harmonics can be enhanced, and edge effects can be reduced during Discrete Cosine Transforms (DFT).

2.1.3. Power spectrum calculated by discrete fourier transform (DFT).

The power spectrum is usually computed by DFT, representing the power distribution of the frequency components of combined signals.

2.1.4. Mel-frequency filter bank.

A mel-frequency filter bank is a sequence of filters constructed according to the high-pitch perception, used for sound analysis similar to human auditory perception and to extract a nonlinear representation of audio signal.

2.1.5. Discrete cosine transforms (DCT).

DCT represents a series of data points involving the sum of cosine functions oscillating on the different frequencies. During the calculation of MFCC, DCT would be used on the Mel-frequency filter bank to select the most accelerated index or to separate the logarithmic spectral amplitude relationships from the filter bank.

Following the 5 steps above, MFCC can be obtained.

2.2. Research Using Traditional Audio Feature Extraction Methods

In the past 4 to 5 years, there have been many investigations about deep learning with musical instrument recognition and combined traditional audio feature extraction methods to achieve the identifications.

2.2.1. Artificial neural network (ANN).

ANN is comprised of artificial neurons and can be a collection of one or more layers of neurons. Also, ANN processes input signals by imitating the connection between neurons, with a core concept to achieve complex information and learning tasks [4].

1) MFCC with ANN

This study collected a dataset of an orchestra to calculate MFCC as a music feature and applied appropriate regularization and activation functions to construct an ANN structure with multiple hidden and output layers. The training and validation set of the data were stratified regarding the ratio of 8:2 to handle the imbalance of the dataset. After 100 epochs of training, although the model's accuracy reached 0.9913 and the validation rate attained 0.9726, considering the imbalance of categories in the dataset, an additional evaluation method was required. The confusion matrix and F1 score represented the recognition accuracy of minority classes. AUC-ROC and Precision-Recall curves further demonstrated the accuracy of the model [5]. However, this model has a high accuracy reaching 97%, the performance is highly reliant on the quality and quantity of training data. If a dataset with extremely imbalanced classes is used for training, the accuracy of the model to recognize specific categories may not be ensured. Therefore, to avoid imbalanced category issues, research can further adopt data augmentation techniques by sampling more minority classes and fewer majority classes.

2.2.2. Convolutional neural network (CNN).

CNN is a feedforward neural network designed under the inspiration of visual perception. Features can be extracted directly through convolutional structures without manual intervention, having the advantages of local connectivity, weight sharing, and down sampling for dimensional reduction. The key components of CNN include convolution, padding, stride, and pooling. Also, by introducing techniques like dilated convolution and variable convolution, CNN can acquire an improvement of the perception range and can handle irrational object shapes [6].

1) MFCC with deep CNN

This study was based on the Western music recording data from the IMARAS dataset for training and modeling. Multi-layer deep neural networks and fixed learning rates were being used for training.

Through using some functions, such as ReLU combined with Max Pooling for dimensional reduction, Softmax to compute the probability of each instrument, and Dropout to prevent overfitting, the entire process showcased an escalation in accuracy and gradual improvement in each epoch. With the combination of various functions for 60 epochs, the final accuracy was well-performed at 92.8% [7]. Although the accuracy indicates the efficacy of the model to process intricate music signals is high, due to the limitation of using only the IMARAS dataset, the generalization ability of the model may be constrained to other types of multi-part music. Hence, the model can be improved by applying other classes of datasets to increase the generalization capability.

2.2.3. Keras.

Keras is a high-level Python library for deep learning and can run on TensorFlow, Theano, or CNTK. Developers can focus on the core concept of deep learning by using Keras, creating network layers without dealing with low-level tensor, shape, and numerical details. Keras has two types of frameworks, sequential API and functional API. Specifically, sequential API is based on the sequential composition of layers and is easier to apply [8].

1) MFCC with keras and independent CNN

This research collected 4 types of musical instruments (bass, drums, guitar, piano) in the Slakh dataset for screening and preprocessing, and a musical instrument recognition system was constructed using the Keras framework and functional API based on deep learning. By extracting the MFCC feature and multiple independent instrument paths from the original audio, each instrument used independently defined the CNN sub-model for recognition. The accuracy reached 92.8% after 100 epochs of training. However, distinctive instruments had different recognition effects. Percussions had the best performance for recognition, whereas guitar and piano were slightly worse. The entire model can be improved flexibly for better efficiency by replacing or adjusting specific sub-models of specific instruments but not distracting other instruments' efficacy of recognition [9]. Despite such flexibility, the differentiation in recognition performance is still affected by the recorded quality of specific datasets. If there are many noises exist in the recordings of a specific dataset, the ultimate recognitional performance may not be ensured. Thus, more distinctive datasets could be used by this model for insurance of whether such an instrument recognition model could be applied to practical situations.

2.3. Research Using Modern Audio Feature Extraction Methods

In recent years, new methods for sound feature extraction have emerged, such as attention Mechanism, Deep Feature learning, Transfer Learning, etc. So, modern audio feature extraction methods are also applied in several investigations.

2.3.1. Attention Mechanism.

The attention mechanism is dedicated to processing the key information in the input determining the significance of the input through scoring functions and assigning weights, and finally generating a context vector for ultimate output and prediction [10].

1) Attention mechanism using openMIC dataset

This study applied the OpenMIC dataset, defined the muti-instrument recognition task as a multi-instance multi-label (MIML) problem, and proposed an attention mechanism. Through the attention mechanism, the model aggregates each short-term instance prediction and compares it with other models with similar purposes. The model includes an embedding layer, an instance-level scoring layer, and an attention layer. After a series of calculations and evaluation, the model with attention mechanism's classification accuracy has improved significantly showcased by recall and F1 score [11]. The attention mechanism allows the model to give more focus on particular instruments when processing the audio. Thus, the accuracy of specific instruments can be enhanced. Nevertheless, accuracy for the model may not be ensured when the dataset applied has an extremely imbalanced type of musical instrument. Therefore, other types of deep learning models can be applied to the

research to capture more audio features efficiently, resulting in further improvements in recognition performance.

Table 1 represents each evaluation metric applied by each model respectively (N/A means not mentioned):

Table 1. Three Scheme comparing

Research	Evaluation Method	Metric	Specific Data
MFCC with ANN	N/A	Accuracy	97%
MFCC with Deep CNN	Cross-validation	Accuracy	92.8%
Attention Mechanism using OpenMIC dataset	Macro-averaged Precision, Recall, F1 Score	Precision	0.86-0.99 (varies by instrument)
		Recall	0.86-0.99 (varies by instrument)
		F1 Score	0.86-0.99 (varies by instrument)
MFCC with Keras and Separate CNN	ROC Curve, AUC ROC, Precision, Recall, F1 Score	AUC ROC	Training Set: Up to 0.96, Validation Set: Up to 0.97
		Precision	0.92 (Evaluation Set)
		Recall	0.93 (Evaluation Set)
		F1 Score	0.93 (Evaluation Set)

3. Conclusion

At present, the majority of musical instrument recognition models have great performance. Whether the model is a combination of the traditional music feature extraction method or the modern music feature extraction method with deep learning models, ultimate models often result in high accuracy with 90% and above. However, not all the models have considered common types of evaluation metrics, so the overall applicability needs to be reevaluated.

Moreover, there are spaces left for the models to improve their performance. Identification tasks can be conducted by analyzing other types of instrument recognition tools, absorbing the fort, and establishing an integrated model with various neural networks. This may result in an unexpected upshot for the performance of instrument recognition.

Except from musical instrument recognition, there might be a possibility for researchers to conduct an attractive investigation by obtaining the results from recognition and incorporating the data with the new model to generate musical notes for specific musical instruments automatically by identifying each rhythmic pattern that emerged in the audio from each musical instrument thereby outputting the final result with a series of music notes for music sheets generation.

Although there was some research about recognizing piano to conduct a music sheet, these are the minority. There are still numerous possibilities for other instruments in such specific fields being uncovered. Instrument-related works are not widely useable for the majority of fields of study, and not many investigators would put effort into analyzing musical instruments and producing music sheets for musicians to increase their efficiency in composing music. Fortunately, there is an increased amount of instrument audio emerging in new datasets, and existing datasets are upgrading rapidly. This makes it possible for researchers to more experiment within the music domain, such as integrating CNN with RNN or combining CNN with Transformer and so forth to gather the music features and classify instruments, ultimately synthesizing the results with OMR to generate reliable music sheets.

References

[1] Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. IEEE Access, 10, 122136 - 122158.

[2] Shah, A., Kattel, M., Nepal, A., & Shrestha, D. (2019). Chroma feature extraction. Chroma Feature Extraction using Fourier Transform.

[3] Schörkhuber, C., & Klapuri, A. (2010, July). Constant-Q transform toolbox for music processing. In 7th sound and music computing conference, Barcelona, Spain (pp. 3-64). SMC.

[4] Zupan, J. (1994). Introduction to artificial neural network (ANN) methods: what they are and how to use them. Acta Chimica Slovenica, 41 (3), 327.

[5] Mahanta, S. K., Khilji, A. F. U. R., & Pakray, P. (2021). Deep neural network for musical instrument recognition using mfccs. Computación y Sistemas, 25 (2), 351 - 360.

[6] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 33 (12), 6999 - 7019.

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

[8] Manaswi, N. K., & Manaswi, N. K. (2018). Understanding and working with Keras. Deep learning with applications using Python: Chatbots and face, object, and speech recognition with TensorFlow and Keras, 31 - 43.

[9] Blaszke M, Kostek B. Musical Instrument Identification Using Deep Learning Approach. *Sensors*. 2022; 22 (8): 3033.

[10] Brauwers, G., & Frasincar, F. (2021). A general survey on attention mechanisms in deep learning. IEEE Transactions on Knowledge and Data Engineering, 35 (4), 3279 - 3298.

[11] Gururani, S., Sharma, M., & Lerch, A. (2019). An attention mechanism for musical instrument recognition. arXiv preprint arXiv: 1907. 04294.