COMPARISON OF VARIOUS ALGORITHMS: RESEARCH ON PIANO AUDIO SIGNAL FEATURE IDENTIFICATION

Shuang Hao *[1]
[1]College of Film and Television, Hebei University of Science and Technology, Shijiazhuang, Hebei, China

Résumé

Cet article présente brièvement les méthodes d'extraction des caractéristiques des signaux audio de piano et notamment les algorithmes basés sur la déformation temporelle dynamique (DTW), le réseau neuronal à rétropropagation (BPNN) et le réseau neuronal à convolution (CNN), qui peuvent reconnaître les caractéristiques audio de piano. Les trois algorithmes de reconnaissance ont été comparés dans les expériences de simulation suivantes. Il a été constaté que pour certains extraits audio de piano à une ou plusieurs notes, les résultats de reconnaissance de l'algorithme CNN étaient cohérents avec les résultats standard, l'algorithme BPNN présentait quelques différences et l'algorithme de reconnaissance basé sur DTW présentait les différences les plus importantes. Avec l'augmentation du nombre de notes dans l'extrait audio de piano, la précision de reconnaissance de tous les algorithmes a diminué, mais c'est l'algorithme CNN qui a le moins diminué, et sa performance de reconnaissance était la plus importante.

Mots-clés : piano, caractéristiques audio, réseau neuronal convolutionnel, déformation temporelle dynamique

Abstract

This article briefly introduced feature extraction methods for piano audio signals and algorithms based on dynamic time warping (DTW), back-propagation neural network (BPNN), and convolutional neural network (CNN), which can recognize piano audio features. The three recognition algorithms were compared in the subsequent simulation experiments. It was found that for some single-note and multi-note piano audios, the recognition results of the CNN algorithm were consistent with the standard results, the BPNN algorithm had some differences, and the DTW-based recognition algorithm had the most differences. As the number of notes in the piano audio increased, the recognition accuracy of all the algorithms decreased, but the CNN algorithm decreased the least, and its recognition performance was highest under the same number of notes, followed by the BPNN algorithm, and the DTW-based recognition algorithm was the lowest.

Keywords: piano, audio feature, convolutional neural network, dynamic time warping

1 Introduction

The emergence time of music is difficult to examine, but throughout the development of human society, music has gradually entered people's lives and become a form of art [1]. Moreover, with the development of computer and internet technology, both music recognition and retrieval [2] require accurate identification of audio signal characteristics [3]. Digital music technology can accurately and quickly identify piano audio signals. Not only can it assist students in correcting piano intonation errors, but it can also quickly convert audio signals into piano music symbols, further assisting in the composition of piano music. Li [4] utilized model recognition technology to design a multi-note model based on HMM and confirmed its practicality. Wu [5] used a convolutional neural network (CNN) to identify piano sheet music. The experimental results verified the accuracy of the algorithm. Wang et al. [6] developed an audio identification method based on a combination of CNN and a generative adversarial network. The experiments on the AViD corpus and DAIC-WOZ dataset demonstrated that compared to other existing methods, this method effectively reduced recognition errors for depression. This article briefly introduced the feature extraction methods for piano audio signals as well as three algorithms - dynamic time warping (DTW), back-propagation neural network (BPNN), and CNN - that can recognize piano audio features. Then, these three recognition algorithms were compared through simulation experiments.

2 Extracting signal features from piano audio

When recognizing the audio signal of a piano, i.e., converting piano audio signals to sheet music with note sequences, it is first necessary to extract the features of the signal (features are indicators that reflect the characteristics of an audio, and using the indicators can distinguish different audio signals). The methods for extracting audio features include the linear prediction cepstral coefficient (LPCC) method and the Mel-frequency cepstral coefficient (MFCC) method. The LPCC method, which is based on linear predictive analysis and assumes that the audio signal is a linear autoregressive signal [7], is an audio feature extraction method that can effectively extract excitation information from the audio signal. However, in reality, audio signals are often not linear regression relationships, so the anti-noise ability of this feature is poor.

* esh968864@yeah.net

Compared with the LPCC method, because the Mel frequency scale is closer to the human ear's perception of audio signals, the audio features obtained by the MFCC method are more compatible with the auditory effect of the human ear [8].

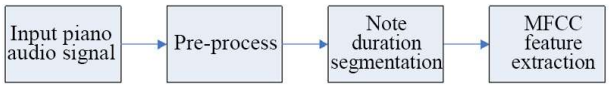

Figure 1: Piano audio feature extraction process based on MFCC

Figure 1 shows the relevant process.

① The piano audio signals to be identified are pre-processed by windowed framing [9]. The formulas for windowing and framing are:

$$\begin{cases} S_w(n) = s(n) \times w(n) \\ w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & \text{if } 0 \le n \le N-1 \\ 0 & \text{else} \end{cases} \end{cases} \quad (1)$$

where $S_w(n)$ denotes the windowed speech digital signal, $s(n)$ denotes the original speech digital signal, $w(n)$ denotes the Hamming window function, and N denotes the length of the digital signal.

② When extracting features from a piano audio signal consisting of single notes, MFCC can be used following the previous step. However, in actual piano audio signals, there are often continuous multiple notes, so the signal needs to be segmented into note values. The energy entropy ratio of each frame of audio signal is used to form a common energy entropy ratio (ratio of energy to spectral entropy in music signals) graph that changes with frame number [10]. Then, each valley point in the energy entropy ratio graph is taken as the starting point of a single note, and the corresponding point whose energy entropy ratio difference from the starting point is less than a preset threshold is taken as the end point of that single note, by moving along the frame time axis in order.

③ MFCC feature extraction is applied to each frame of the audio signal after being segmented by note duration [11].

3 Algorithms for recognizing piano audio signal features

3.1 Piano audio feature recognition method based on DTW

The previous text described the method of extracting piano audio signal features. By using the extracted MFCC features, it is possible to recognize the notes of the piano audio signal. Matching the template library of notes with the audio signal to be recognized is a method of note recognition for piano audio signals. The template library of notes stores the MFCC characteristics of different notes. The basic principle is to extract the MFCC features of the audio in the note template library and the audio to be recognized, and then perform pattern matching based on their MFCC features.

Even the same note on a piano may have different durations when played. When using a note template library to match audio to be recognized, it is highly likely that the durations will be inconsistent [12]. If the audio signal is linearly stretched or compressed as a way to match the duration of the template notes, the duration transformation of the individual segments in the signal under different circumstances will be ignored, which will result in information loss. The DTW algorithm can perform non-linear bending of audio, solving the problem of inconsistent lengths between the audio to be recognized and the template audio. Its steps are:

① Pre-emphasis, windowed framing, note duration segmentation, and MFCC feature extraction are carried out on the audio signal as described in the previous chapter.

② The DTW algorithm is used to calculate the distance between the audio to be recognized and different template audios.

③ The template audio with the smallest distance from the audio to be recognized is used as the recognition result.

3.2 Piano audio feature recognition method based on deep learning

Deep learning algorithms have gradually been applied to audio recognition. The BPNN is a classic deep learning algorithm. When used for piano audio feature recognition, the audio is first pre-processed and MFCC feature parameters are extracted. Then, the MFCC feature parameters are input into the BPNN for forward calculation in the hidden layer, and ultimately the note recognition results are output in the output layer. Compared with the DTW algorithm, the BPNN for piano audio recognition requires a large amount of data for training, but it does not need any template audio to recognize piano audio signals after training, and there is no need to adjust the note duration [13].

CNN can also be used for recognizing audio signals. The steps are as follows:

① The audio signal is processed as described previously, including pre-emphasis, windowed framing, segmentation of note duration, and extraction of MFCC features.

② The extracted MFCC features are into the CNN. The formula of convolution operation is:

$$O_i = f(O_{i-1} \otimes W_i + B_i) \quad (2)$$

where O_i and O_{i-1} are the feature maps outputted from the i-th layer and $(i-1)$-th layer, W_i is the weight in the structure of the i-th layer, B_i is the bias in the structure of the i-th layer, and $f(\cdot)$ is the activation function.

③ After obtaining the convolutional feature map, in order to reduce the subsequent computational complexity, it is compressed in the pooling layer. The specific operation is to use a pooling box to gradually slide on the feature map, and compress the feature data in the pooling box. In simple terms, multiple data in the pooling box are merged into one data, which can be the average of multiple data or the maximum value among the multiple data. The former is mean pooling, and the latter is max pooling [14].

④ The pooled convolution feature maps are then classified using the softmax function in the fully connected layer, and the corresponding music note recognition results are input into the output layer. Finally, the recognition results are output.

When the algorithm is in its training phase, the computed recognition results are compared with the corresponding label results in the training set, and this article uses cross-entropy to compute the error between them. The error is then evaluated to see if it has converged to the target range. If it reaches the preset range, the training is finished; if not, the weighted parameters in the algorithm are adjusted according to the error in the opposite direction.

4 Simulation experiments

4.1 Experimental data

The dataset used for the simulation experiments was the MAESTRO dataset. This dataset was obtained in cooperation with the organizers of international piano competitions. During the process of collecting the original piano data, a high-precision MIDI capture and playback system was used to ensure the accuracy of the data as much as possible. This dataset contained approximately 200 hours of piano audio (16-bit PCM stereo at a sampling rate of 48 kHz) and corresponding MIDI data. The MIDI data included key velocity and pedal position, which can be considered that the data set carried the note labels corresponding to the audio data.

4.2 Experimental setup

Simulation experiments were conducted on three piano audio recognition algorithms based on DTW, BPNN, and CNN, respectively. Firstly, the feature dimension of MFCC was set to 24 after orthogonal experiment. The Hamming window was used, the frame length was set to 20 ms, and the frame shift size was 10 ms.

In the DTW-based piano audio recognition algorithm, the most important part is the note template library used for matching and recognition. For the audio of notes in the template library, the dimension of extracted MFCC features was also set to 24. The Hamming window was used in windowed framing, with a frame length of 20 ms and a frame shift of 10 ms.

In the BPNN-based piano audio recognition algorithm, the relevant parameters are as follows. The number of nodes in the input, hidden, and output layers were 24, 128, and 88, respectively; the activation function was the Relu function. The epoch was set as 200.

In the CNN-based piano audio recognition algorithm, the relevant parameters are shown below. There were three convolutional layers and three mean pooling layers. During the training process, a regularization with a random dropout of 0.4 was used, and the number of epochs was set to 200.

During the testing of three piano audio recognition algorithms, the first step was to test the recognition of single-note piano audio signals by the three algorithms. Afterwards, the recognition of multi-note piano audio signals with different numbers of notes was tested, with the number of notes set at 5, 10, 15, 20, and 25, respectively. The recognition performance of the three algorithms on the piano audio signals in the aforementioned two experimental projects was tested. Precision, recall rate, and F-score were chosen as the performance indicators.

4.3 Experimental results

Due to space limitations, only waveforms of partial single- and multiple-note piano audios are shown in Figure 2. The single-note piano audio waveform in Figure 2 corresponds to the note named "c", pronounced as "do"; the multiple-note piano audio waveform corresponds to the note named "c^1def^1cd". The recognition results for the piano audio in Figure 2 using three different audio recognition algorithms are shown in Table 1. According to Table 1, the recognition results obtained by the CNN algorithm were consistent with the standard results, the BPNN algorithm differed slightly, and the DTW recognition algorithm differed the most from the standard results.

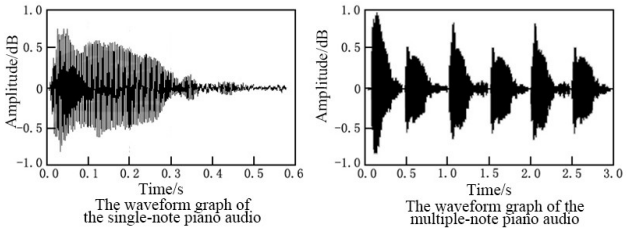

Figure 2: The waveform graphs of some single-note and multiple-note piano audios

Table 1: The recognition results of some single and multi-note piano audios

	Single-note piano audio in Figure 2	Multiple-note piano audio in Figure 2
Standard note result	c	c^1def^1cd
The recognition result of the DTW algorithm	e	$e^1ce^1f^1ce$
The recognition result of the BPNN algorithm	d	c^1ddf^1cd
The recognition result of the CNN algorithm	c	c^1def^1cd

Three recognition algorithms' recognition performance on single-note piano audio is shown in Figure 3. For the DTW recognition algorithm, the precision, recall rate, and F-value of recognizing single-note piano audio were 0.69, 0.68, and 0.68, respectively. For the BPNN algorithm, the values were 0.85, 0.84, and 0.84, respectively. For the CNN recognition algorithm, the values were 0.98, 0.97, and 0.97, respectively. It was seen from Figure 3 that the CNN recognition algorithm had the highest accuracy in recognizing single-note piano audio, followed by the BPNN algorithm, and the DTW recognition algorithm had the lowest accuracy.

The performance of three identification algorithms in identifying piano audio with multiple notes is shown in Figure 4. According to Figure 4, as the number of musical notes in the piano audio increased, the precision, recall rate, and F-values of the three identification algorithms all decreased. However, in the process of increasing the number of notes, the DTW recognition algorithm showed the greatest decrease in recognition performance, while the BPNN algorithm

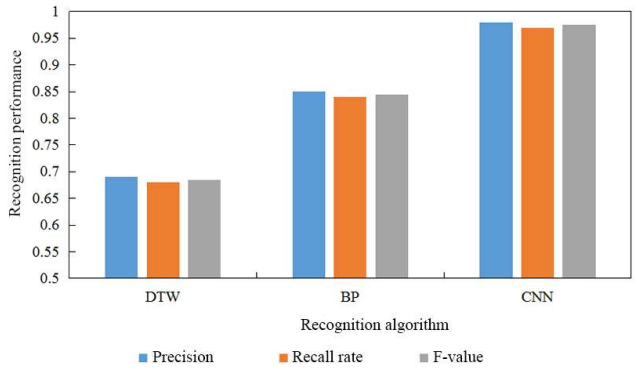

Figure 3: Performance of three recognition algorithms in identifying single-note piano audio

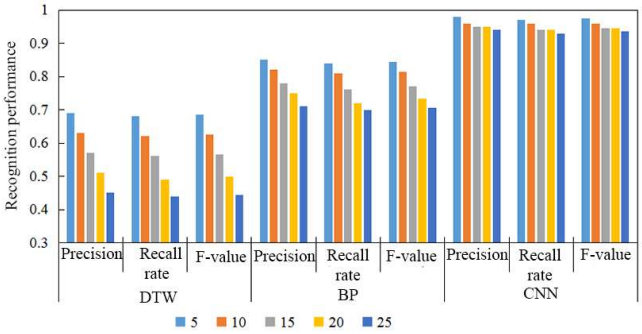

Figure 4: Recognition performance of three recognition algorithms on multiple-note piano audio

showed a relatively small decrease, and the CNN recognition algorithm showed the smallest decrease. When the number of multiple notes was the same, the recognition performance of the CNN algorithm was still the best, followed by the BPNN algorithm, and the DTW algorithm was the worst.

5 Conclusion

This paper briefly introduced the feature extraction methods of piano audio signals and three algorithms, namely DTW, BPNN, and CNN, that can recognize piano audio features. In the subsequent simulation experiments, the three recognition algorithms were compared, and the results are as follows. (1) Among the recognition results of the waveform graphs of some single-note and multi-note piano audios, the recognition results obtained by the CNN recognition algorithm were consistent with the standard results, while the BPNN algorithm differed slightly and the DTW recognition algorithm differed the most. (2) Faced with single-note piano audio signals, the CNN recognition algorithm presented the highest recognition accuracy for the single-note piano audio, followed by the BPNN algorithm, and the DTW recognition algorithm had the lowest accuracy. (3) As the number of notes in the piano audio signal increased, the recognition performance of the three recognition algorithms slightly reduced, among which the DTW recognition algorithm had the most significant reduction, followed by the BPNN algorithm, and the CNN recognition algorithm had the least reduction. Moreover, under the same number of notes, the recognition performance of the CNN algorithm was still the best, followed by the BPNN algorithm, and the DTW algorithm was the worst.

References

[1] S. A. Herff, K. N. Olsen, and R. Dean. Resilient memory for melodies: The number of intervening melodies does not influence novel melody recognition. *Q. J. Exp. Psychol*., 71:1150-1171, 2018.

[2] M. Bomgardner. MATERIALS Schlumberger pilots new lithium extraction. *Chem. Eng. News*, 99:10, 2021.

[3] Y. Wang. Research on Handwritten Note Recognition in Digital Music Classroom Based on Deep Learning. *J. Internet Technol*., 22:1443-1455, 2021.

[4] X. Li. Construction and analysis of hidden Markov model for piano notes recognition algorithm. *J. Intell. Fuzzy Syst*., 37:3293-3302, 2019.

[5] R. Wu. Research on automatic recognition algorithm of piano music based on convolution neural network. *J. Phys. Conf. Ser*., 1941:1-7, 2021.

[6] Z. Wang, L. Chen, L. Wang, and G. Diao. Recognition of audio depression based on convolutional neural network and generative antagonism network model. *IEEE Access*, 8:101181-101191, 2020.

[7] M. Brousmiche, J. Rouat, and S. Dupont. Multimodal Attentive Fusion Network for audio-visual event recognition. *Inform. Fusion*, 85:52-59, 2022.

[8] P. Hoffmann, and B. Kostek. Bass Enhancement Settings in Portable Devices Based on Music Genre Recognition. *J. Audio Eng. Soc*., 63: 980-989, 2015.

[9] X. Wang. Research on the improved method of fundamental frequency extraction for music automatic recognition of piano music. *J. Intell. Fuzzy Syst*., 35:1-7, 2018.

[10] Z. Xiao, X. Chen, and L. Zhou. Real-Time Optical Music Recognition System for Dulcimer Musical Robot. *J. Adv. Comput. Intell. Intell. Inform*., 23 :782-790, 2019.

[11] C. Shuo. The construction of internet plus piano intelligent network teaching system model. *J. Intell. Fuzzy Syst*., 37: 5819-5827, 2019.

[12] R. Wu. Research on automatic recognition algorithm of piano music based on convolution neural network. *J. Phys. Conf. Ser.,* 1941:1-7, 2021.

[12] O. Francesconi, A. Ienco, C. Nativi, and S. Roelens. Effective Recognition of Caffeine by Diaminocarbazolic Receptors. *ChemPlusChem*, 85:1369-1373, 2020.

[13] Y. Luo, and H. Yang. Teaching Applied Piano Singing While Playing Based on Xindi Applied Piano Pedagogy:Taking Fujian Vocational College of Art as an Example. *J. Contemp. Educ. Res*., 6:123-135, 2022.

[14] W. H. Lai, and C. Y. Lee. Query By Singing/Humming System Using Segment-based Melody Matching for Music Retrieval. *WSEAS Trans. Syst*., 15:157-167, 2016.

EDITORIAL BOARD - COMITÉ ÉDITORIAL

A unique software system for extended acoustic analysis and prediction.

DataKustik is the leading software development company offering state-of-the art solutions in the field of acoustics modeling. Choose from several software solutions including:

– **CadnaA** for the calculation of noise outdoors
– **CadnaB** for the calculation of sound transmission between rooms
– **CadnaR** for the assessment of sound distribution indoors

Although each software is tailored to its specific field of application, they can be connected to form a powerful package for extended calculations from the outside to the inside and vice-versa. For any type of acoustic calculations, from sound propagation outdoors to the assessment of indoor sound, choose DataKustik. For more information about DataKustik products, count on Scantek for expert and responsive sales and support.

For inquiries about DataKustic products in North America contact:

Scantek©
LISTEN – FEEL – SOLVE
A RION COMPANY

800-224-3813
info@scantekinc.com
www.scantekinc.com