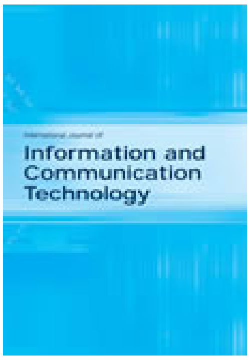


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Piano teaching-assisted beat recognition based on spatio-temporal two-branch attention

Rigega Su

Article History:

Received:	31 December 2024
Last revised:	15 January 2025
Accepted:	15 January 2025
Published online:	21 March 2025

Piano teaching-assisted beat recognition based on spatio-temporal two-branch attention

Rigega Su

Musicology and Musical Arts,
National Academy of Music 'Pancho Vladigerov',
1142 Sofia, Bulgaria
Email: 18812630659@163.com

Abstract: The challenge of beat identification in piano teaching has progressively taken the stage as intelligent education technology develops quickly. Most of the conventional beat detection systems ignore the link between the video information and the player's motions by depending just on the analysis of auditory data. Based on the spatio-temporal two-branch attention mechanism, this work presents a piano beat detection model called TempoNet to increase the accuracy and robustness of beat identification. By means of the spatio-temporal dual-branching attention mechanism, the model efficiently captures the temporal features in the audio signal and the dynamic spatio-temporal features in the video signal by deep fusion of the two-modal information. Comparatively to conventional approaches, the suggested TempoNet model shows better beat identification accuracy and robustness according to experimental results on several test datasets.

Keywords: piano teaching; beat recognition; spatio-temporal two-branch attentional mechanism; multimodal learning.

Reference to this paper should be made as follows: Su, R. (2025) 'Piano teaching-assisted beat recognition based on spatio-temporal two-branch attention', *Int. J. Information and Communication Technology*, Vol. 26, No. 5, pp.100–116.

Biographical notes: Rigega Su received her Bachelor's degree from Tianjin Conservatory of Music in 2017. In 2020, she received her Master's degree from the University of York in the UK. In 2024, she received her PhD from the National Academy of Music 'Pancho Vladigerov' in Sofia. Her research interests include digital signal processing and digital music.

1 Introduction

Intelligent teaching systems have attracted a lot of attention in the world of education as artificial intelligence technology – especially the broad application of deep learning in many spheres – develops rapidly (Chen et al., 2020; Timms, 2016). With the growing student count and need for individualised learning in music education – especially in piano instruction – the conventional teaching approach is confronting considerable difficulties. Learning piano not only depends on students' knowledge of notes and melodies but also pays more attention to the development of rhythmic sense and precision of playing (Ivanova et al., 2020). A fundamental ability in piano study, beat detection is absolutely important for students' learning effect and musical expression. The development of intelligent piano teaching systems depends much on accurate beat recognition since it not only improves students' sense of rhythm but also increases the fluency and expressiveness of their playing.

Past studies have mostly concentrated on the processing and analysis of audio signals in order of beat recognition (Ismail et al., 2018). The conventional beat recognition systems mostly rely on signal processing approaches. Usually, these techniques concentrate on extracting rhythmic aspects from audio recordings and using time-frequency domain analysis to estimate beats. When confronted with intricate music environments, these techniques are vulnerable to noise interference and environmental changes, hence their recognition accuracy decreases (Kujala and Brattico, 2009). Furthermore, the audio signal itself does not directly correlate with the performer's movement, hence these techniques are sometimes challenging to handle challenging circumstances including background noise or multi-player performance.

As deep learning technology develops constantly, more research aiming at beat recognition using neural network models is undertaken (Liu et al., 2017). Because of their better feature learning and time series data processing, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have become common alternatives for beat recognition activities. Especially models based on RNNs and long short-term memory networks (LSTMs) can efficiently capture the temporal information in audio signals, thereby enhancing the accuracy of beat prediction. Although these techniques have some acceptance results by learning the temporal characteristics of audio signals, their performance in long time periods or high noise situations usually seems to be unreliable.

Recent researchers have progressively shown, when compared with single-modal audio signal processing techniques, the improvement of multimodal learning for beat recognition tasks (Ivanko et al., 2023). Particularly in the performance of musical instruments like the piano, where the video can record not only the note changes but also significant elements like the player's hand motions, posture, and facial expressions, the introduction of video information offers fresh approaches for beat recognition. These dynamic visual cues can improve the model's capacity to fit challenging playing situations and balance the auditory information. Some researchers have effectively raised the accuracy and robustness of beat identification by mixing audio and visual for multimodal learning (Dogan and Akbulut, 2023). Still, the current multimodal techniques fall short in fully using the possible spatio-temporal correlation between audio and video information and in simpler fusion algorithms and insufficient information interaction.

In recent years, the wide application of attention mechanisms in deep learning provides an effective solution for multimodal learning. In particular, the spatio-temporal attention mechanism can automatically learn and capture the information of key moments in temporal data, so as to give higher weights to important features in audio and video. This mechanism can help the model to better focus on the key time points of rhythmic information and improve the recognition accuracy. Therefore, the spatio-temporal dual-branching attention mechanism becomes a potential innovative direction for multimodal beat recognition.

There are still certain issues even if the studies mentioned above offer several theoretical bases and technical tools for beat identification (Au and Kauffman, 2008). Most of the current approaches either limit to one analysis of audio data or the merging of multimodal information is not yet ideal. Furthermore limiting their generalisability in actual applications are most of the current models, which are tailored for certain datasets or scenarios and lack some generalising capacity.

Aiming to improve the accuracy and robustness of beat recognition by means of the fusion of multimodal information and deep learning of spatio-temporal features, this study proposes a piano teaching-assisted beat recognition model based on spatio-temporal dual-branching attention mechanism based on the above issues. Combining spatio-temporal dual-branching attention mechanism with the joint modelling of audio and video helps the model suggested in this work to not only precisely identify beats but also sustain effective recognition performance in challenging playing environments.

The main innovations include:

1 Design of spatio-temporal dual-branch network architecture. In this work, we design a spatio-temporal dual-branch network architecture which first performs dynamic weighted fusion via the attention mechanism after processing audio and visual input through two branches independently. This architecture improves the accuracy of beat detection by learning the characteristics of audio and video modalities independently and simultaneously enhances the feature extraction of important time points and key frames through the spatio-temporal attention mechanism.

2 Deep fusion of multimodal information. This work suggests a new multimodal information fusion technique merging audio and video inputs to completely use their complimentary information. Combining the note timing information in the audio signal with the hand movement and playing gesture information in the video signal helps to greatly increase the accuracy of beat identification and get over the constraints of the single-modal approach in challenging environments.

3 Introducing spatio-temporal attention mechanism to improve the model performance. This work uses the spatio-temporal attention mechanism instead of the conventional attention mechanism to pay attention to important events in both audio and video signals in the time and spatial dimensions. Especially in the event of complicated player motions and heavy background noise, this approach can enable the model to effectively capture rhythmic changes during dynamic playing and strengthen the robustness of beat identification.

These innovations make the model in this study more accurate, robust and adaptable in the piano beat recognition task, which provides technical support for the further development of the intelligent piano teaching system and provides new ideas and methods for research in related fields.

2 Relevant technologies

2.1 Beat recognition

Beat detection is a crucial chore in music information processing since it seeks to identify rhythmic structures from audio sources for analysis and deployment of intelligent music systems (Camurri et al., 2000). Particularly in piano instruction, beat recognition is essential for students since it not only helps them to precisely understand the rhythm of playing but also enhances the fluidity and harmony of playing. Usually depending on feature extraction of audio signals, traditional beat detection techniques analyse the time series information in the audio signals to detect rhythmic shifts. These techniques, meantime, can suffer from complicated rhythm alterations and noise interference.

Traditional target detection methods can usually be classified into two main categories: candidate region-based methods and regression methods. The former by generating candidate regions (e.g., selective search) and classifying each region; the latter by directly predicting bounding boxes and categories from the whole image by means of regression.

Accurate extraction of the rhythmic pulsations from the audio data and beat location determination constitute the fundamental responsibilities of beat recognition. Common signal analysis techniques are wavelet transform and short time Fourier transform (STFT), both of which are extensively applied in the task of beat recognition and have respective benefits in time-frequency analysis (Canal, 2010). By separating an audio signal into short segments in time and computing the spectrum of every segment, Short-time Fourier Transform may clearly expose the periodicity and frequency components of the signal. Its mathematical statement is:

$$X(t, f) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-j2\pi f\tau} d\tau \tag{1}$$

The time-frequency representation is denoted by $X(t, f)$; the window function is $w(\tau - t)$; the frequency is f; the time is t. The frequency properties of the audio stream at various times can be recorded with this time-frequency representation, thereafter the periodic variations of the beat can be obtained.

Particularly appropriate for signals with local abrupt changes or non-smoothness features, the wavelet transform more precisely captures the instantaneous changes in the signal than the STFT by multi-scale local analysis (Kim and Aggarwal, 2000). Wavelet transform's fundamental formula is:

$$W_{a,b}(x) = \frac{1}{\sqrt{|\alpha|}} \int_{-\infty}^{\infty} x(t) \psi\left(\frac{t-b}{\alpha}\right) dt \tag{2}$$

where $\psi(t)$ is the wavelet basis function; a and b are respectively the scale and displacement parameters; $W_{a,b}(x)$ is the transform result under scale a and displacement b. Through scale parameter a, the wavelet transform adjusts to the several frequency components of the signal. Its time-frequency localisation is better than that of the STFT and is particularly appropriate for recording audio signals with regular rhythmic fluctuations. The immediate properties of the signal may be efficiently extracted by the wavelet transform, which also provide a necessary basis for later beat localisation.

The rhythm detection technique grounded on the autocorrelation function (ACF) is another often used method in beat recognition (Kumar et al., 1997). By comparing the signal with its own time-shifted form, the autocorrelation function helps to determine the beat location by exposing the periodicity of a signal. It defines as:

$$R(\tau) = \int_{0}^{T} x(t) x(t + \tau) dt \tag{3}$$

where τ is the time delay; $R(\tau)$ is the autocorrelation function; $x(t)$ is the signal. Periodic signal peaks can be derived from the autocorrelation function, which relates to the temporal location of the beat.

But the temporal dependence between notes – especially when several notes overlap or are performed simultaneously – makes beat identification in piano performance a difficult task. Researchers have started to introduce a spatio-temporal attention mechanism to solve this issue so that the model may dynamically change the weight of attention depending on the relevance of information in many temporal and spatial dimensions by means of an attention mechanism into the model. This method can increase beat detection accuracy and reasonably simulate the spatio-temporal dependency between notes.

2.2 Spatio-temporal two-branch attention mechanisms

While using the attention mechanism to dynamically weight the attention to the key parts, which is especially appropriate for the beat recognition task with complicated spatio-temporal dependencies, the spatio-temporal dual-branch attention model combines deep feature extraction of temporal and spatial information to capture important features of the signal independently in the temporal and spatial dimensions by a dual-branch structure (Cai et al., 2024). Processing time series and spatial data in parallel allows the model to dynamically modify the weights between various time steps and spatial locations, thereby better modelling the spatio-temporal link between notes and rhythms.

First, CNN or RNN handles the input signal X_{time} in the time branch to extract the temporal properties. After feature extraction, the output temporal feature T_{out} is represented as assuming the convolutional or RNN layer in the temporal branch has L hidden layers:

$$T_{out} = CNN\left(X_{time}\right) \text{ or } T_{out} = RNN\left(X_{time}\right) \tag{4}$$

where $CNN(\cdot)$ or $RNN(\cdot)$ indicates the convolutional operation or RNN operation, respectively, whereas X_{time} is the input time-series signal; so, the output T_{out} is the features extracted via temporal branching.

Temporal branching employs a temporal attention method to let the model concentrate on the important temporal information in the input signal (Wang et al., 2024). Calculating the similarity between the features of each time step and the query vector Q_{time} allows the temporal attention mechanism to dynamically change the focus to several time steps. More especially, the following equation computes the weight of temporal attention $\alpha_{time}(t)$:

$$\alpha_{time}(t) = \frac{\exp\left(score\left(T_{out}(t), Q_{time}\right)\right)}{\sum_{t'} \exp\left(score\left(T_{out}(t'), Q_{time}\right)\right)} \tag{5}$$

Generally by utilising dot product or another similarity computation technique, $score(\cdot,\cdot)$ is a function that gauges the similarity between temporal characteristics and query vectors. By means of this method, the model can dynamically allocate attention to various time steps, hence improving attention to tempo variations and key note changes.

Particularly when numerous notes overlap or notes interact with one another, spatial branching is helpful in extracting information in the spatial dimension and deals with spatial aspects in the audio stream. Furthermore, extracted from the input spatial signal X_{space} by CNN are spatial branch features. Following spatial convolution layer processing, output spatial feature S_{out} is:

$$S_{out} = CNN\left(X_{space}\right) \tag{6}$$

Spatial branching presents a spatial attention method, much like temporal branching (Campbell et al., 2007). Through computation of the similarity between the properties of every spatial location and the query vector Q_{space}, the spatial attention mechanism modulates the degree of attention of the model in the spatial dimension. One may determine the weight of spatial attention $\alpha_{space}(s)$ by means of the following equation:

$$\alpha_{space}(s) = \frac{\exp\left(score\left(S_{out}(s), Q_{space}\right)\right)}{\sum_{s'}\exp\left(score\left(S_{out}(s'), Q_{space}\right)\right)} \tag{7}$$

where the attentional weight at spatial point s is $\alpha_{space}(s)$. Particularly in complicated sceneries with many notes overlapping, the model can efficiently find the spatial connections between notes by means of this spatial attention mechanism.

The model must combine the temporal and spatial features after finishing their extraction (Zhong et al., 2019). Combining the attributes taken from the temporal and spatial branches helps the temporal and spatial information fusion layer to enable the later beat prediction. T_{out} and S_{out} of the temporal and spatial branches are merged by weighted summation in the fusion procedure. The weighted fusion equation is:

$$Y_{final} = \alpha \cdot T_{out} + \beta \cdot S_{out} \tag{8}$$

Usually optimised by training data, Y_{final} is the final fusion feature; α and β are the weight coefficients of the temporal and spatial branches accordingly. Changing these two coefficients lets the model flexibly control the contribution of spatial and temporal elements to the final output under various application settings.

The spatio-temporal two-branch attention model presents a self-attention method to improve the interactivity and representation of spatio-temporal information by allowing the model to capture more complicated spatio-temporal connections (Tu et al., 2024). Through computation of the correlation between spatio-temporal features, the self-attention process helps to represent interdependencies between features. The self-attention formula is specifically:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

With Q the query matrix, K the key matrix, V the value matrix, and d_k the key's dimension. Calculating the similarity between the input characteristics and assigns a weight to every feature, thereby obtaining the final output by weighted summation. When considering spatio-temporal characteristics, the model may therefore effectively represent the relationships between several times and locations, so enhancing the accuracy of rhythm detection.

At last, the model produces a probability distribution via the softmax function, which denotes the beat location related to every time step. The subsequent equation computes this process:

$$\hat{Y} = softmax\left(Y_{final}\right) \tag{10}$$

where $\hat{Y}$ is the projected last beat, softmax procedure turns model output into a probability distribution used to forecast the precise position of a beat. Based on this probability distribution, the model finds which time point most likely corresponds to a beat and hence precisely locates the beat.

The pseudo-code for the spatio-temporal two-branch attention model is shown in Algorithm 1.

Algorithm 1 Pseudo-code for the spatio-temporal two-branch attention model

```
Input: Time-series input X_time, Spatial input X_space, Initial attention weights, Initial deep
neural network weights, learning rate, iterations total number
Output: Optimised time attention weights, optimised space attention weights, optimised neural
network weights
begin
    Initialise time and space attention weights ($\alpha$, $\beta$)
    Initialise deep neural network weights for time and space branches
    Initialise sliding window size for input data
    Initialise experience buffer for reinforcement learning agent (optional)
    for iteration = 1 to total_iterations do
        Receive new time-series input vector X_time
        Receive new spatial input vector X_space
        Add new input to the sliding window
        Extract time features T_out from X_time using CNN
            Extract space features S_out from X_space using CNN
        Calculate time attention weights using softmax(T_out, Q_time)
        Calculate space attention weights using softmax(S_out, Q_space)
        Apply time and space attention to features:
    T_out_weighted = T_out * attention_weights_time
    S_out_weighted = S_out * attention_weights_space
        Fuse the weighted time and space features:
    fused_features = $\alpha$ * T_out_weighted + $\beta$ * S_out_weighted
        Pass fused features through self-attention mechanism
        Compute the final output prediction Y_pred from the fused features
        Calculate the total loss function between Y_pred and the true label
        Compute the gradient of the loss function with respect to the time and space
        attention weights
        Update the time and space attention weights using gradient descent
        Compute gradients for the deep neural network weights using backpropagation
        Update the deep neural network weights using gradient descent
        If the sliding window exceeds a predefined size then
            Remove the oldest sample from the sliding window
        end if
        (Optional) use reinforcement learning agent to optimise attention parameters based
        on reward signal
        Store experience in buffer if using reinforcement learning
        If the experience buffer is full, randomly sample a small batch and compute strategy gradients
        Update RL policy parameters (optional)
    end for
    return optimised attention weights ($\alpha$, $\beta$), optimised neural network weights
end
```

In summary, the spatio-temporal dual-branching attention model extracts features through independent temporal and spatial branches and adaptively adjusts the feature weights in the spatio-temporal dimension by combining with the attention mechanism to effectively capture the spatio-temporal dependencies in rhythmic signals. By fusing the spatio-temporal features and the self-attention mechanism, the model is able to accurately perform beat recognition with strong robustness in the presence of complex rhythmic patterns and note overlap.

3 TempoNet: piano teaching aid beat recognition based on spatio-temporal bifurcated attention

3.1 Model architecture

Working together to maximise the accuracy and real-time performance of beat identification, TempoNet is composed of several modules each assigned a particular task. The general design comprises of a data preprocessing module, a temporal branching module, a spatial branching module, a spatio-temporal feature fusion module, a self-attention mechanism module and a beat recognition module. Every module is described in great length here:

1 Data preprocessing module

The main inputs in piano instruction are audio and visual. Whereas the visual feed offers the spatial layout and key motion of the piano keyboard, the audio signal has time information including the rhythm and pitch of the performance. Thus, the goal of the module on data preparation is to translate the audio and video signals into a feature representation fit for next use.

STFT initially uses time-frequency feature extraction on the audio signal X_{audio} to record the timing changes in the signal. Following STFT transformation, the audio stream X_{audio} has time-frequency elements:

$$X_{time\text{-}frequency} = STFT\left(X_{audio}\right) \tag{11}$$

CNN uses spatial feature extraction of the video signal X_{video} to produce a spatial feature representation of the piano keyboard image:

$$X_{space} = CNN\left(X_{video}\right) \tag{12}$$

These two feature representations will be forwarded correspondingly into the spatial and temporal branching modules that follow.

2 Time branching module

Processing the temporal aspects of the audio signal is the fundamental work of the temporal branching module. First retrieved in this module via a sequence of convolutional layers (Conv1D or Conv2D), the time-frequency feature $X_{time\text{-}frequency}$ of the audio is then temporal feature T_{raw}:

$$T_{raw} = Conv\left(X_{time\text{-}frequency}\right) \tag{13}$$

Then, using a temporal attention mechanism, the extracted temporal features are weighted to highlight events vital for beat identification. The temporal attention mechanism computes a weighting factor αt for the temporal features, therefore providing:

$$\alpha_t = Softmax\left(Q_{time} \cdot T_{raw}\right) \tag{14}$$

where T_{raw} is the initial temporal feature obtained by convolution and Q_{time} is the query vector of temporal features, the weighted temporal feature $T_{weighted}$ is finally produced:

$$T_{weighted} = T_{raw} \cdot \alpha_t \tag{15}$$

The temporal branching module can so efficiently concentrate on the significant events in the beat signal.

3 Spatial branching module

Processing the video information to extract spatial properties of the piano keyboard is the aim of the spatial branching module. By CNN, the video signal X_{space} extracts the spatial feature S_{raw}, thereby reflecting the spatial structure and keyboard's key state:

$$S_{raw} = CNN\left(X_{space}\right) \tag{16}$$

The spatial attention mechanism's job is to dynamically change each region's weight according on the degree of piano keyboard focus devoted to other places. Calculating the similarity of feature maps with the formula generates the spatial attention weights β_s:

$$\beta_s = Softmax\left(Q_{space} \cdot S_{raw}\right) \tag{17}$$

At last, the weighted spatial feature $S_{weighted}$ is obtained by means of attention to the spatial aspects:

$$S_{weighted} = S_{raw} \cdot \beta_s \tag{18}$$

With this module, TempoNet is able to efficiently extract spatial information from piano playing, thus providing more accurate rhythm and movement recognition.

4 Temporal and spatial feature fusion module

From the audio and video signals respectively, the temporal and spatial aspects are extracted via the temporal and spatial branching modules respectively. Two weighted features $T_{weighted}$ and $S_{weighted}$ will be coupled in the spatio-temporal feature fusion module to generate the spatio-temporal feature representation F_{fused}:

$$F_{fused} = \alpha \cdot T_{weighted} + \beta \cdot S_{weighted} \tag{19}$$

where the fusion weight coefficients are α and β and their values can be dynamically changed in line with the training process. This step's core goal is to apply weighted fusion of spatio-temporal characteristics and combining meaningful information from audio and video to increase beat detection accuracy.

5 Self-attention mechanism module

Re-weighting the spatio-temporal features via self-attention aims to improve the capacity of the model to capture the linkages and global dependencies between several features. Assuming F_{fused} as the fused spatio-temporal feature, the self-attention mechanism computes the weighted feature $F_{attended}$ by means of global feature correlation:

$$F_{attended} = Self\text{-}Attention\left(F_{fused}\right) \tag{20}$$

The self-attention method enables the model to recognise the worldwide knowledge about beats, therefore enhancing the accuracy and resilience of the model in challenging environments.

6 Beat recognition module

At last, the features $F_{attended}$ following spatio-temporal feature fusion and weighted by the self-attention mechanism feed the fully connected layer for final beat prediction in the beat detection module. Based on the weighted features using a certain formula, the fully connected layer determines the output result Y_{pred} of beat recognition:

$$Y_{pred} = FC\left(F_{attended}\right) \tag{21}$$

The result originally employed for real-time student playing beat feedback in the piano teaching system, Y_{pred} is a vector representing the state of the beat at every moment.

3.2 Assessment of indicators

In the piano teaching-assisted beat identification challenge, model performance is absolutely important. We have chosen the following four often used assessment criteria to holistically assess the TempoNet model's performance in many different facets.

1 Accuracy

Indicating the ratio of accurately anticipated beats by the model to the total expected beats, accuracy is among the most often utilised measures for evaluation of categorisation tasks. The accuracy rate of the beat recognition problem shows the general model efficacy in the beat identification challenge. The formula helps one to determine the accuracy rate:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

where *TN* is the true negative case, denoting the off-beat moment correctly predicted by the model; *FP* is the false positive case, denoting the off-beat moment incorrectly predicted by the model; and *FN* is the false negative case, denoting the beat moment incorrectly predicted by the model. Negative is indicating a beat instant the model misfits as non-beat.

2 Precision

The measure of how many of the beat moments the model detects is accuracy rate. A greater accuracy rate denotes less false positives in the model's beat prediction, thereby indicating better beat identification. The formula helps one to determine the accuracy rate:

$$Precision = \frac{TP}{TP + FP} \tag{23}$$

High accuracy indicates that the model can efficiently avoid erroneous beat identification and has a great chance of accurately predicting when forecasting for beats.

3 Recall

The model's ability to identify how many actual beats it can detect from all real beats is gauged using recall. That is to say, the model can identify more beat events the more recall it possesses. Particularly in piano teaching situations, beat recognition activities demand that one find as many accurate beat times as feasible. The formula allows one to determine the recall rate:

$$Recall = \frac{TP}{TP + FN} \tag{24}$$

High recall indicates that the model can spot more genuine beat events and lower the missed call count.

4 F1 score

Comprising the reconciled average of precision and recall, the F1 score may fully assess the completeness and accuracy of the model. The F1 score can balance the recall rate with the precision rate in the beat identification problem, therefore preventing the circumstance whereby one index is too high and the other is too low. The F1 score's formula is:

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{25}$$

Particularly in the context of imbalanced datasets, the F1 score offers a complete evaluation of the beat detection efficacy of the model and is a stronger predictor of the general performance than accuracy by itself.

4 Experimental results and analyses

4.1 Datasets

In this work, we selected a multimodal dataset including both audio and video signals in order to assess the TempoNet model in the piano beat identification challenge.

Mostly from the publicly accessible Piano-DB dataset, a multimodal dataset devoted to piano performance analysis with piano audio, video, and matching beat labels. The dataset consists of several recordings performed by several piano players; the recorded audio is of great quality; and the frame rate of the video is kept at 30 fps, which is appropriate for the tasks of piano movement detection and beat analysis. Additionally included to the study are some custom-recorded data including videos of piano playing with various tempos, therefore enhancing the variety of the dataset.

Table 1 Piano-DB dataset statistical information

Data type	*Description*
Audio files	Each sample contains a piano performance audio signal, in WAV format, with a sampling rate of 44.1 kHz, and a duration ranging from 60 to 180 seconds.
Video files	Each audio file is paired with a corresponding piano performance video, with a resolution of 1,280 × 720, a frame rate of 30 fps, and in MP4 format.
Beat labels	Each audio and video sample has corresponding beat labels, with timestamps precise to the timing of each note played. The labels are manually annotated by experts to ensure accuracy.

Regarding data preparation, STFT first transforms the audio signal to a time-frequency feature map whereby each audio clip is split into several 1-second-length time windows and the corresponding spectral features are extracted, subsequently normalised for use as input to the next network model. Each video frame was clipped to a fixed size (224 × 224) and normalised to capture changes in piano keyboard and hand movements. Pre-processing of video frames fed to the CNN helped to extract spatial characteristics. The beat labels are translated into a series of binary labels whereby non-beats are 0 and beats are 1. Model training and evaluation will base on these labels.

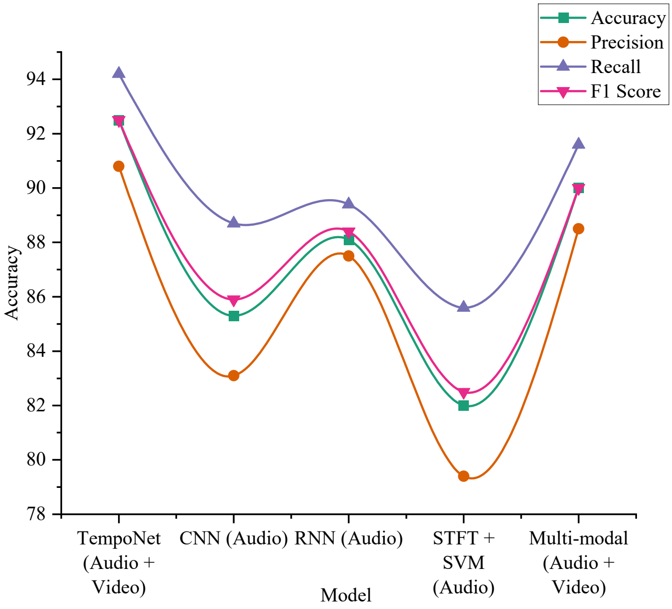
Twenty percent of the samples were used as the test set and 80% of the samples as the training set, therefore separating the dataset. To guarantee that the model can be trained in a range of circumstances, the training set includes several players, repertory and tempo variances. The performance of the model on unprocessed data is assessed using the test set to guarantee objectivity and validity of the evaluation conclusions.

4.2 Comparative experiments

In this work, we intend to assess the TempoNet model's performance in terms of accuracy, precision, recall, and F1 score by means of a comparative experiment between many other beat detection techniques and the model itself. We selected multiple representative beat recognition models, all of which were trained on the same dataset and applied consistent training and test set divisions, therefore guaranteeing the fairness of the tests. Among the models for comparison studies are conventional audio feature-based techniques as well as deep learning models including CNN and RNN.

The experimental results are shown in Figure 1.

Figure 1 Results of comparative experiments (see online version for colours)



With bimodal inputs – audio and video – TempoNet beats all other comparison models quite dramatically. First of all TempoNet performs well on important measures including precision (90.8%), recall (94.2%), and F1 score (92.5%). It also obtains an accuracy of 92.5%. By contrast, TempoNet's recognition accuracy and performance are far higher than those of conventional audio feature-based beat recognition systems, including STFT + SVM, which only reach 82.0% accuracy and 82.5% F1 sorce. Furthermore, although the performance is rather better under unimodal audio input, deep learning models including CNN (audio input) and RNN (audio input) obtain 85.3% and 88.1% respectively. Under unimodal audio input, the performance gains; however, it falls short of TempoNet's bimodal identification capacity.

Furthermore, although the Multi-modal (Audio + Video) model also uses dual-modal inputs of audio and video, its overall performance is rather less than TempoNet's because it does not integrate the spatio-temporal dual-branching attentional mechanism used by TempoNet with an accuracy of 90.0%, and with precision and recall rates of 88.5% and 91.6% respectively.

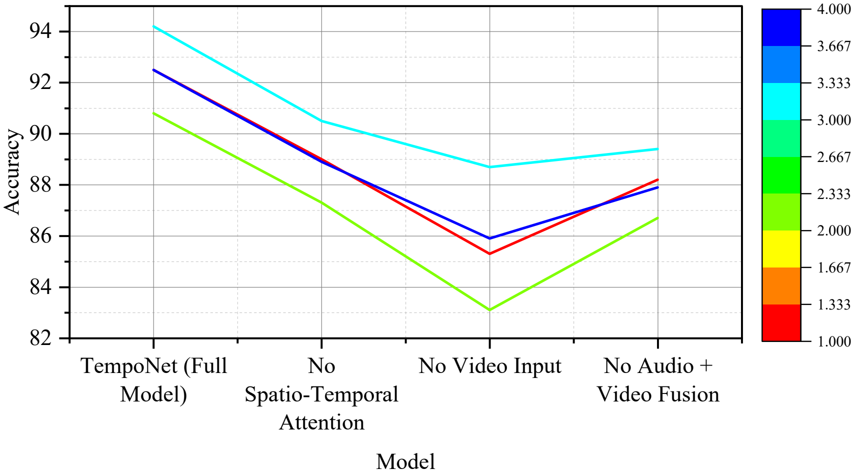
The experimental results further demonstrate the advantages of TempoNet in the beat recognition task. Through the effective fusion of multimodal data and the introduction of the spatio-temporal dual-branching attention mechanism, TempoNet shows significant improvement in accuracy, robustness and real-time performance, and has a stronger beat recognition capability compared to the traditional audio feature extraction and classification methods as well as the single-modal deep learning models.

4.3 Ablation experiments

In this work, we intended to confirm the contribution of every important technological module in the TempoNet model to beat recognition performance by designing ablation tests. We deleted the spatio-temporal two-branch attention mechanism, the video input, and the joint audio and video input from the TempoNet model, respectively, and evaluated the removed model in order to assess the influence of every module. These tests help one to better grasp how each module affects the general performance of the model.

Figure 2 provides the ablation experiments' specific findings:

Figure 2 Results of ablation experiments (see online version for colours)



The first ablation experiment eliminates the spatio-temporal two-branch attention mechanism from the TempoNet model, therefore using either a conventional structure with a single or no attention mechanism for beat recognition. The experimental results reveal that the elimination of spatiotemporal two-branch attention reduces the performance of the model; the accuracy drops from 92.5% to 89.0%; the recall and F1 score also drop. This implies that TempoNet's accuracy and resilience to capture the possible spatio-temporal correlations in the audio and visual inputs benefit much from the spatio-temporal two-branch attention mechanism.

The second ablation experiment uses just auditory information for beat detection and eliminates the video input. This experiment is to confirm the increase of the video input on beat recognition. Eliminating the video input lowers TempoNet's accuracy from 92.5% to 85.3% and causes notable declines in precision and recall as well. This implies that, particularly in the case of complicated backgrounds or several audio sources, the video input offers useful visual information that allows the model to better grasp and recognise beats.

Using only one modality – audio – the third ablation experiment eliminates the joint audio and visual inputs and streamlines the way the model is fused. With an accuracy of 88.2%, which is far lower than the 92.5% for the bimodal input, the testing results reveal that the performance of the single-modal audio input model is likewise deteriorated. This confirms even more the significant part combined audio and video inputs play in raising recognition accuracy.

From the experimental results, it can be observed that removing the spatio-temporal two-branch attention mechanism has the greatest impact on the model, with a decrease of 3.5% in the accuracy, and a significant decrease in the recall and F1 score, which indicates that this mechanism essentially improves the ability of the model to model complicated spatio-temporal relationships. Eliminating the video input reduces the accuracy by 7.2%, which emphasises the need of video information in beat recognition particularly in difficult situations. Eliminating the combined audio and visual inputs reduces the model's accuracy as well, so multimodal fusion clearly increases the model's recognition accuracy.

Through these ablation experiments, we can conclude that both the spatio-temporal two-branch attention mechanism and the joint audio and video input play a crucial role in the beat recognition performance of TempoNet. The overall performance of the model not only benefits from the complementary audio and video information, but also better handles complex beat patterns with the help of the spatio-temporal attention mechanism.

5 Conclusions

Aiming to improve the accuracy and real-time performance of tempo recognition in piano teaching by combining the multimodal information of audio and video, and introducing the spatio-temporal two-branch attention mechanism, in this paper we propose a tempo recognition model, TempoNet, based on the spatio-temporal two-branch attention mechanism for piano teaching. TempoNet greatly improves the metrics of accuracy, precision, recall, and F1 score of beat detection relative to conventional unimodal audio systems and other deep learning approaches by means of comparing experiments and validation of ablation experiments. Furthermore indicated by the experimental results are the main elements influencing model performance improvement: spatio-temporal dual-branching attention mechanism and audio-video fusion.

TempoNet performs remarkably in beat identification challenges, although there are still certain limits. First of all, especially as the efficiency of the video input depends on the device and the shooting angle, the performance of the model may deteriorate when handling some more extreme environmental conditions (e.g., excessive background noise or low video quality.). Second, the dataset for this work is mostly from a small number of piano instruction situations; so, the generalisation capacity of the model to a greater spectrum of music scenarios still needs more validation. TempoNet's computational cost is finally somewhat high, particularly in video processing, which could be a significant load for devices with low computational capacity, so restricting its adoption in useful applications.

Future studies can enhance and broaden the above constraints in the following respects:

1 Noise and low quality input processing. More strong noise reduction methods such adaptive filtering or more sophisticated deep denoising networks would help the model be more robust in challenging settings. Furthermore, methods such generative adversarial network (GAN)-based video processing module optimisation helps to enhance the performance of low-quality videos.

2 Model generalisation capability. Future studies must widen the dataset to include other kinds of musical instruments, music styles, and various performing techniques if we are to increase the applicability of the model. Furthermore, approaches like transfer learning help the model to improve generalisability and enable it to better fit the beat recognition job in various musical situations.

3 Optimisation of computational resource consumption and efficiency. Future computing efficiency can help the model to be optimised so improving its practical application value. Techniques including model compression, quantisation, and knowledge distillation can help to lower the processing overhead thereby enabling real-time beat recognition on low-resource devices. Furthermore, lightweight deep learning models can be presented to lower the computational complexity and storage needs of the models, therefore improving their viability in useful applications.

4 Personalisation and Interactivity Enhancement. TempoNet has already shown good performance in beat detection; however, future studies should include the personalisation needs in intelligent education to improve the system's interactivity and intelligent feedback capacity. For instance, including elements like the learner's emotional condition and learning development helps to deliver tailored comments on beat recognition. Simultaneously, the cognitive load and emotional changes of the learner can be better comprehended by merging bio-signal data including electroencephalogram (EEG), thereby offering more precise teaching aid.

Declarations

All authors declare that they have no conflicts of interest.

References

Au, Y.A. and Kauffman, R.J. (2008) 'The economics of mobile payments: understanding stakeholder issues for an emerging financial technology application', *Electronic Commerce Research and Applications*, Vol. 7, No. 2, pp.141–164.

Cai, Z., Tan, C., Zhang, J. et al. (2024) 'DBSTGNN-Att: dual branch spatio-temporal graph neural network with an attention mechanism for cellular network traffic prediction', *Applied Sciences*, Vol. 14, No. 5, p.2173.

Campbell Grant, E.H., Lowe, W.H. and Fagan, W.F. (2007) 'Living in the branches: population dynamics and ecological processes in dendritic networks', *Ecology Letters*, Vol. 10, No. 2, pp.165–175.

Camurri, A., Hashimoto, S., Ricchetti, M. et al. (2000) 'Eyesweb: toward gesture and affect recognition in interactive dance and music systems', *Computer Music Journal*, Vol. 24, No. 1, pp.57–69.

Canal, M.R. (2010) 'Comparison of wavelet and short time Fourier transform methods in the analysis of EMG signals', *Journal of Medical Systems*, Vol. 34, pp.91–94.

Chen, L., Chen, P. and Lin, Z. (2020) 'Artificial intelligence in education: a review', *IEEE Access*, Vol. 8, pp.75264–75278.

Dogan, G. and Akbulut, F.P. (2023) 'Multi-modal fusion learning through biosignal, audio, and visual content for detection of mental stress', *Neural Computing and Applications*, Vol. 35, No. 34, pp.24435–24454.

Ismail, S., Siddiqi, I. and Akram, U. (2018) 'Localization and classification of heart beats in phonocardiography signals – a comprehensive review', *EURASIP Journal on Advances in Signal Processing*, Vol. 2018, No. 1, pp.1–27.

Ivanko, D., Ryumin, D. and Karpov, A. (2023) 'A review of recent advances on deep learning methods for audio-visual speech recognition', *Mathematics*, Vol. 11, No. 12, p.2665.

Ivanova, I., Chernyavska, M. and Pupina, O. (2020) 'Didactic Potential of instructive etude and its explication in the process of professional development of a pianist', *Journal of History Culture and Art Research*, Vol. 9, No. 3, pp.257–266.

Kim, C.H. and Aggarwal, R. (2000) 'Wavelet transforms in power systems. Part 1: general introduction to the wavelet transforms', *Power Engineering Journal*, Vol. 14, No. 2, pp.81–87.

Kujala, T. and Brattico, E. (2009) 'Detrimental noise effects on brain's speech functions', *Biological Psychology*, Vol. 81, No. 3, pp.135–143.

Kumar, P. and Foufoula-Georgiou, E. (1997) 'Wavelet analysis for geophysical applications', *Reviews of Geophysics*, Vol. 35, No. 4, pp.385–412.

Liu, W., Wang, Z., Liu, X. et al. (2017) 'A survey of deep neural network architectures and their applications', *Neurocomputing*, Vol. 234, pp.11–26.

Timms, M.J. (2016) 'Letting artificial intelligence in education out of the box: educational cobots and smart classrooms', *International Journal of Artificial Intelligence in Education*, Vol. 26, pp.701–712.

Tu, Y., Wu, J., Lu, L. et al. (2024) 'Face forgery video detection based on expression key sequences', *Journal of King Saud University-Computer and Information Sciences*, Vol. 36, No. 7, p.102142.

Wang, Z., Tang, Y. and Zhang, Z. (2024) 'Dual-branch deep learning architecture for enhanced hourly global horizontal irradiance forecasting', *Expert Systems with Applications*, Vol. 252, p.124115.

Zhong, L., Hu, L. and Zhou, H. (2019) 'Deep learning based multi-temporal crop classification', *Remote Sensing of Environment*, Vol. 221, pp.430–443.