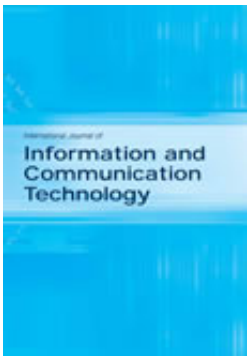


International Journal of Information and Communication Technology

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

Acoustic noise recognition of ancient buildings based on space time joint processing and deep learning

Ziqing Tang, Zhengguang Li

DOI: 10.1504/IJICT.2025.10071983

Article History:

Received:	16 May 2025
Last revised:	29 May 2025
Accepted:	29 May 2025
Published online:	10 July 2025

Acoustic noise recognition of ancient buildings based on space time joint processing and deep learning

Ziqing Tang*

Architecture College,
Taiyuan University of Technology,
Taiyuan 030024, China
Email: ziqing2024s@126.com
*Corresponding author

Zhengguang Li

School of Civil Engineering and Architecture,
Zhejiang University of Science and Technology,
Hangzhou 310023, China
Email: izhgu@zust.edu.cn

Abstract: Acoustic noise recognition of ancient buildings is crucial for the protection and study of ancient buildings, but the traditional methods have problems such as insufficient feature extraction and weak generalisation ability in complex scenes, which are ineffective for noise recognition. Therefore, this paper proposes an acoustic noise recognition method for ancient buildings based on space time joint processing and deep learning, which utilises space time joint processing to pre-process acoustic signals and extract effective features, and then classifies and recognises them through a deep learning model. Experiments show that the method shows excellent performance in terms of recognition accuracy and robustness, providing new ideas and effective means for the recognition of acoustic noise of ancient buildings, which helps to better protect and study ancient buildings.

Keywords: noise recognition; spatiotemporal joint processing; deep learning; acoustic signals; ancient architecture.

Reference to this paper should be made as follows: Tang, Z. and Li, Z. (2025) 'Acoustic noise recognition of ancient buildings based on space time joint processing and deep learning', *Int. J. Information and Communication Technology*, Vol. 26, No. 25, pp.70–86.

Biographical notes: Ziqing Tang received his doctorate degree from Taiyuan University of Technology in 2020. He is currently a lecturer in the School of Architecture, Taiyuan University of Technology. His research interests include the acoustics of ancient buildings and deep learning.

Zhengguang Li received his doctorate degree from Zhejiang University in 2012. He is currently a lecturer in the School of Civil and Architectural Engineering, Zhejiang University of Science and Technology. His research interests include building environmental acoustics, noise control technology, noise effects, and building energy efficiency.

1 Introduction

As a precious heritage of human civilisation, ancient buildings embody rich historical and cultural significance. In the realm of ancient building preservation and research, the maintenance of the acoustic environment is of paramount importance. However, with the development of modern society, ancient buildings face the threat of various acoustic noises, such as those generated by tourism activities, neighbouring construction and changes in the natural environment (Gibbs, 2010). These noises not only affect the tourists' visiting experience, but also may potentially harm the structure and artefacts of the ancient buildings themselves. Therefore, it is important to accurately identify and analyse the acoustic noise of ancient buildings in order to develop effective protection measures. Traditional acoustic noise identification methods mainly rely on a single time-domain or frequency-domain analysis, and these methods have achieved some results in simple environments. However, in complex acoustic scenarios of ancient buildings, traditional methods have many limitations (Zhu et al., 2023). On the one hand, there are various sources of acoustic noise in ancient buildings, including anthropogenic noise, natural noise, and mechanical noise inside the building, etc. Various types of noise are intertwined with each other in the time and frequency domains, which makes it difficult for the traditional methods to accurately differentiate and recognise different noise sources. On the other hand, the acoustic environment of ancient buildings has unique spatial and temporal characteristics, for example, the spatial structure of the building, building materials, and air flow and other factors have an impact on the propagation of noise, and the traditional method fails to take these spatial and temporal factors into account, resulting in a lack of recognition accuracy and robustness (Jiang et al., 2019).

In recent years, deep learning techniques have made significant progress in the field of pattern recognition and classification, providing new ideas for solving complex acoustic noise recognition problems. Acoustic recognition takes a segment of acoustic samples as input, and analyses the features of the input signal in order to classify and identify the identity of the speaker. There have been significant advancements in the field of voiceprint recognition and speech processing. Yang et al. (2016) proposed a method for temperament accuracy assessment and intelligent processing based on voiceprint recognition, which enhanced the accuracy of traditional voiceprint recognition in the presence of environmental noise and vocal variations. Li et al. (2020) introduced an adaptive threshold algorithm based on OTSU and deep learning to address internal and external speaker similarity value and threshold calculation issues in open-set speech recognition, thereby improving recognition accuracy. During the same period, Yao et al. (2021) developed a streamer voice recognition method for live broadcasts based on RawNet-SA and gated recurrent units (GRUs). By integrating self-attention mechanisms and GRUs, this method effectively boosted the efficiency of streamer voice feature extraction and aggregation, demonstrating great potential in real-time voice recognition applications. Khdier et al. (2021) designed a speech recognition system for noisy environments based on deep learning algorithms. They implemented RW-CNN methods using convolutional neural networks (CNN) based on MFCC features and raw waveforms, achieving a remarkable recognition accuracy rate of 96%. This research highlighted the effectiveness of deep learning in handling complex noise conditions. Kunjir (2022) proposed a dual-path attention mechanism and weighted clustering domain loss to enhance speaker and emotion recognition accuracy. Furthermore, a critical

enhanced loss function was introduced to tackle training efficiency challenges, providing new insights into optimising the training process of recognition models. Barakat et al. (2024) assessed the accuracy of speech recognition systems designed using Euclidean distance functions and genetic algorithms. Their study found that incorporating genetic algorithms significantly elevated recognition rates and decision-making speed, indicating the advantages of hybrid algorithms in improving recognition performance. In terms of speech separation and noise reduction, Zhang et al. (2024) proposed a two stages network method for voice devices, achieving speech separation and noise-echo reduction for specific speakers in noisy and echoic conditions. Its innovation lies in the cascaded structure of attention mechanisms and temporal convolutional networks (TCNs). In speaker verification, Jin et al. (2023) presented a single channel speech separation based network combined with MFCCT features to enhance accuracy in multi speaker scenarios. The innovation is integrating the speech separation model with the speaker verification task, evaluating separation quality through downstream task accuracy. As for real-time speech processing, Zhou et al. (2024) utilised distributed acoustic sensing technology with the MFCC method to achieve real-time speech reproduction and recognition, proposing an accurate and fast processing approach. These studies contribute to the development of speech recognition and processing technologies, offering references for the research on acoustic noise recognition of ancient buildings.

Deep learning models are able to automatically learn complex features and patterns in the data with strong nonlinear fitting capabilities. Meanwhile, the space time joint processing technique takes into account the correlation of acoustic signals in the time and space dimensions, and is able to capture the characteristic information of noise more comprehensively (Li et al., 2020). Among various deep learning models, support vector machine (SVM) is chosen for its effectiveness in classification tasks with high-dimensional data and strong generalisation ability, while CNN is utilised for its powerful feature extraction and learning capabilities, which are crucial for handling the complex acoustic noise in ancient buildings. Based on this, this paper proposes an acoustic noise recognition method for ancient buildings based on space time joint processing and deep learning. First, the acoustic signal is pre-processed using space time joint processing to extract the space time features of acoustic noise, fully considering the complexity of the acoustic environment of ancient buildings. Then, the extracted features are input into the deep learning model for training and classification, and the powerful learning capability of deep learning is utilised to achieve accurate recognition of different noise types. Through experimental verification, the method has excellent performance in terms of recognition accuracy and robustness, providing a new technical means for the recognition of acoustic noise in ancient buildings.

The main innovations and contributions of this work include:

1 This paper introduces the space time joint processing technology to break through the limitation of the traditional acoustic noise identification method that only relies on a single time domain or frequency domain analysis. By simultaneously mining the correlation characteristics of acoustic signals in the time and space dimensions, the temporal and spatial characteristics of various types of noise sources in the acoustic environment of ancient buildings are accurately captured, which lays the foundation for the application of spatial-temporal joint processing technology in the field, and effectively improves the feature extraction accuracy.

2 In the face of the complexity and diversity of the acoustic noise of ancient buildings, this paper optimises the deep learning model. Design a reasonable network structure and parameter configuration, so that the model accurately learns and recognises acoustic noise features to improve classification accuracy. At the same time, data enhancement and regularisation techniques are used to enhance the generalisation ability of the model, ensure its stable recognition under different environmental conditions, and improve the adaptability of the model.

3 This paper innovatively integrates space time joint processing and deep learning techniques. Space time joint processing provides space time feature extraction capability, deep learning automatically learns feature complex patterns, and the combination of the two enables the model to comprehensively and accurately analyse the acoustic noise of ancient buildings, realising efficient and accurate recognition. Compared with the traditional single means, the fusion method significantly improves the accuracy, robustness and adaptability, expanding the scope of application of the technology.

2 Relevant technologies

2.1 Space time adaptive processing technology

The joint space-time processing technique, which originated in radar signal processing, laid the foundation for space time adaptive processing (STAP), and this concept was introduced by Klemm (1999). The primary goal of STAP is to leverage both temporal and spatial dimensions of signal information to enhance target detection and recognition performance. With the development of acoustic signal processing, this technique has been gradually introduced into the field of acoustics, especially in the noise processing in complex acoustic environments, showing unique advantages. In the recognition of acoustic noise in ancient buildings, the space-time joint processing technology collects acoustic signals through multiple acoustic sensors in different spatial locations at the same time, forming multi-channel data, which contain signal changes in the time domain and also reflect the differences in acoustic characteristics in different spatial locations. By analysing these multi-channel data, the technique can effectively capture the temporal variations and spatial distribution of noise sources, which is crucial for improving the accuracy of feature extraction. The simultaneous consideration of time and space dimensions allows for a more comprehensive representation of the noise characteristics, enabling better differentiation between various noise types in the complex environment of ancient buildings. The joint air-time adaptive processing technique is a joint processing of the time series (time) output from multiple beams (air). In active sonar, the array spatial response characteristics and variability of reverberant, interfering and desired signals are deeply analysed, and a single pulse is used to realise the joint spatial and temporal processing. The specific method is to introduce the complex analysed signal first, and then carry out the segmentation processing, and finally establish the minimum variance distortionless response (MVDR) spatial and temporal joint processing model (Dai et al., 2016). The MVDR spatial and temporal joint processing model is able to accurately extract specific types of noise signals from the complex acoustic environment, such as extracting the interior of ancient buildings from the background noises of the

tourists' conversations and the ambient wind noises, and so on. The MVDR model can accurately extract specific types of noise signals from complex acoustic environments, for example, the structural vibration noise inside ancient buildings from the background noise such as tourists' conversations, environmental wind noise, etc., so as to provide more accurate acoustic information for the protection and maintenance of ancient buildings (Shaw et al., 2021).

Let the number of receiving array elements be N, the vectorial azimuth angle between the target and the receiving array be Ψ, the waveform of the transmitted pulse signal be denoted as $s(t)$, and the echo signal received at the nth array element be:

$$x(n,t) = s\left(t - \tau - \Delta\tau_n, f\right) \tag{1}$$

where $\Delta\tau_n$ is the time difference of each array element relative to the reference array element, f is the Doppler shift, and τ is the echo arrival time.

Aligning the scanning angle Ψ, each array element first receives the data $x(n, t)$ with a time delay $\Delta\tau_n$, introduces an auxiliary signal all the way orthogonal to the output time series of each array element, and constructs the resolved signal:

$$y(n,t) = x(n,t) + jH[x(n,t)] \tag{2}$$

Similarly, the resolved signal of the transmit signal of the l^{th} Doppler channel is constructed, the resolved signal of the transmit signal after Doppler frequency shift compensation:

$$z(l,k) = D_l[s(k)] + H\left\{D_l[s(k)]\right\} \tag{3}$$

In both equations, H denotes the Hilbert transform and D_l denotes the frequency shift transform of the l^{th} Doppler channel. In order to obtain a stable and full-rank covariance matrix estimation, the signal time sampled data is divided into M segments, each segment contains K sampling points, and M takes a value of a magnitude that should guarantee a stable estimation of the correlation between each array element and each data segment. The transmissionisation of the l^{th} Doppler channel of the constructed beam is given as:

$$B_l(\Psi') = \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=m}^{m+K} w_l(n,m)y(n,k)z^*(l,k) \tag{4}$$

where w_l is a complex weight vector in the spatial and temporal domains. The output power is:

$$P_l(\Psi') = \left|B_l(\Psi')\right|^2 = \left|\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=m}^{m+K} w_l(n,m)y(n,k)z^*(l,k)\right|^2 \tag{5}$$

where, $y(n, k)z^*(l, k)$ denotes the complex space-time correlation of the received signal with the matched signal of the l^{th} Doppler channel, which is a complex constant for the desired signal component. Let the signal and noise (non-expected signal) correlation coefficients be zero, i.e., uncorrelated, from which constraints can be introduced:

$$\sum_{n=1}^{N}\sum_{m=1}^{M} w_l(n,m) = M \tag{6}$$

Under this constraint, the $P_l(\Psi)$ output is minimised by adjusting the weight vector, i.e., the output power of the noise is minimised under the condition of guaranteeing that the desired signal passes all the way through. Thus for a given beam Ψ' and Doppler channel, the joint airtime white adaptation processing can be expressed as the following constrained optimisation problem:

$$S = \begin{cases} \min\limits_{w_l(n,m)} \left\{ P_l(\Psi') \right\} \\ s.t. \sum\limits_{n=1}^{N} \sum\limits_{m=1}^{M} w_l(n,m) \end{cases} \tag{7}$$

Further considering the output of multiple Doppler channels, the Doppler channel with the largest output can be finalised by the following equation:

$$\max_l S = \begin{cases} \min\limits_{w_l(n,m)} \left\{ P_l(\Psi') \right\} \\ s.t. \sum\limits_{n=1}^{N} \sum\limits_{m=1}^{M} w_l(n,m) \end{cases} \tag{8}$$

2.2 Theoretical foundations of deep learning

Neural network technology imitates the structure and function of neurons in the human brain, and is used for efficient information processing and pattern recognition. The first model was constructed by McCulloch and Pitts in 1943, and has been continuously developed and improved since then, which mainly consists of convolutional layers, activation functions, fully connected layers, normalisation layers and other key components. In the field of acoustic noise recognition of ancient buildings, neural networks are capable of deep learning and accurate classification of acoustic features after joint processing in space and time, thus realising efficient recognition of acoustic noise (Tang and Sun, 2023). The optimisation of the deep learning model is based on extensive experimentation and validation. The number of hidden layers and nodes is determined by evaluating the model's performance on a validation set, aiming to balance complexity and avoid overfitting. Parameters such as the learning rate and regularisation terms are tuned using grid search and cross-validation to enhance convergence and generalisation. This systematic approach ensures the model's effectiveness in recognising acoustic noise in ancient buildings.

Convolutional layer is a basic structure in the neural network, through the sliding convolution kernel and the input data correlation operation, the local depth features of the input data extraction. The convolutional layer used in this paper is a one-dimensional convolutional layer, for an input vector $x \in \mathbb{R}^{L\times 1}$, the output $y \in \mathbb{R}^{L\times 1}$ of a one-dimensional convolutional layer, which has a convolutional kernel $w \in \mathbb{R}^{m\times 1}$, the computational process can be expressed as follows.

$$y(i) = \sum_{k=0}^{m-1} w(k)x(k-i) \ , 0 \le i < L' \tag{9}$$

where m is the size of the convolution kernel, which is satisfied without input padding $L = L–m + 1$. The convolution kernel $w \in \mathbb{R}^{m\times 1}$, is a learnable parameter that is optimised with the gradient descent algorithm trained on the dataset. When the channel feature dimension size of the input and output is not 1, for a one-dimensional convolutional layer with an input vector $x \in \mathbb{R}^{L\times c}$ and an output $y \in \mathbb{R}^{L\times c}$ with a convolutional kernel $w \in \mathbb{R}^{m\times 1}$, the computational process can be expressed as:

$$y\left(c_j^{'}\right) = \sum_{k=0}^{C-1} w\left(c_j^{'}, k\right) \star x(k), 0 \le c_j^{'} < C^{'} \tag{10}$$

where denotes the mutual correlation operation process. When the convolution kernel size is taken as $m = 1$, the convolution operation at this point is called point-by-point convolution; when the input feature map is grouped for convolution operation and the number of groups is the same as the number of input channels C, the convolution operation at this point is called deep convolution.

The activation function is a crucial component in neural networks, enabling a nonlinear mapping between inputs and outputs, which endows the network with the ability to model complex relationships. It also restricts the output values to a specific range (Demir et al., 2020). The Sigmoid activation function, widely used in neural networks, confines output values between 0 and 1. As the input x approaches positive infinity, $g(x)$ tends to 1, while it approaches 0 as x heads toward negative infinity. This function enhances the network's nonlinear fitting capability by mapping neuron outputs to the (0, 1) range. However, due to its derivative being close to 0 for input values far from 0, excessive use of Sigmoid can lead to gradient cumulants during backpropagation, resulting in gradient dispersion.

$$g(x) = \frac{1}{1+e^{-x}}, g'(x) = g(x)(1-g(x)) \tag{11}$$

The ReLU activation function, one of the most widely used nonlinear activation functions, was introduced to address the vanishing gradient problem in deep neural networks. ReLU outputs the input directly if it is positive; otherwise, it outputs zero. This design not only enhances the network's nonlinear fitting capability but also mitigates the gradient vanishing issue that occurs in the Sigmoid activation function due to the cumulative multiplication of gradients. However, when the input value is negative, the derivative of ReLU is zero, which can hinder efficient gradient updates and potentially lead to the dying ReLU problem.

$$g(x) = \begin{cases} 0, x < 0 \\ x, x \ge 0 \end{cases}, g'(x) = \begin{cases} 0, x < 0 \\ 1, x \ge 0 \end{cases} \tag{12}$$

The GLU activation function, introduced by Google in 2017, is a gated mechanism that selectively filters input signals through a gating process, enhancing model generalisation. This function splits the input into two parts, applying a sigmoid activation to one part to gate the other. The computational expression for GLU is shown in equation (13), where $x \in \mathbb{R}^{k\times 1}$ is a vector of real numbers, and x_1, x_2 $x \in \mathbb{R}^{2\times 1}$, $\sigma(\cdot)$, are the Sigmoid activation functions.

$$GLU(x) = x_1 \odot \sigma(x_2), x = [x_1; x_2] \tag{13}$$

The SoftMax activation function converts a real vector $x \in \mathbb{R}^{k\times 1}$ into a probability distribution. In classification tasks, SoftMax converts the model's output vector into a probability distribution across categories. The category with the highest probability is then selected as the model's predicted classification result. The calculation procedure of SoftMax is shown in equation (14).

$$softmax(x)_i = \frac{\exp(x_i)}{\sum_{j=0}^{k-1} \exp(x_j)} \tag{14}$$

The fully connected layer is a fundamental component in neural networks, where each input neuron is connected to each output neuron. It performs a linear transformation of the input vector into a new feature space, followed by a nonlinear activation function to enable complex modelling. For a fully connected layer with an input vector $x \in \mathbb{R}^{M\times 1}$ and an output $y \in \mathbb{R}^{N\times 1}$, which has a weight matrix $W \in \mathbb{R}^{N\times M}$ and a bias vector $b \in \mathbb{R}^{N\times 1}$, the computational process can be expressed by matrix operations as:

$$y = g(Wx + b) \tag{15}$$

where $g(\cdot)$ is the activation function chosen, the weight matrix W and the bias vector b are the learnable parameters, optimised with the gradient descent algorithm trained on the dataset.

Excessive deviation in data or feature distribution can complicate model training, hinder convergence, or even trigger gradient vanishing or explosion. Introducing a normalisation layer can mitigate these issues by reducing the likelihood of gradient vanishing or explosion in neural networks. This not only simplifies the training process but also accelerates convergence. Additionally, normalisation enables the model to adapt to features of varying magnitudes and scales, thereby enhancing its generalisation ability and reducing the risk of overfitting. For an input $x \in \mathbb{R}^{N\times d}$, whose feature dimension size is d, x can be expressed as $x = \{x_0, x_1, x_2, \ldots, x_{d-1}\}$, where $x_i \in \mathbb{R}^{N\times 1}, 0 \le i < d$. The layer normalisation of x feature dimension is performed as follows:

$$\mu = \frac{1}{d}\sum_{i=0}^{d-1} x_i, \sigma^2 = \frac{1}{d}\sum_{i=0}^{d-1} (x_i - \mu) \odot (x_i - \mu) \tag{16}$$

The input data x feature dimensions are then normalised to obtain the output $\hat{x}$:

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \in}} \tag{17}$$

$\in = 10^{-6}$ is likewise a minimal value. The normalised data $\hat{x}$ is then scaled and translated in the feature dimension:

$$y_i = \gamma \hat{x}_i + \beta \tag{18}$$

The final input to the LN is obtained as $y = \{y_0, y_1, \ldots, y_{d-1}\}, y \in \mathbb{R}^{N\times d}$.

3 Acoustic noise identification method for ancient architecture

3.1 Mel frequency

A sound signal can be viewed as being generated by a source excitation through a linear system of sound channels, and in speech signal processing, a speech signal is considered to be a random signal through a linear system such as:

$$y(n) = x(n) * h(n) \tag{19}$$

where $y(n)$ is the speech signal, $x(n)$ is the source excitation, and $h(n)$ is the impulse response of the vocal tract, the speech signal can be described as the convolution of the source excitation and the vocal tract's impulse response.

In acoustics, pitch is a measure of how high or low a sound appears to the human ear. There's a nonlinear relationship between pitch and frequency. As frequency increases, pitch initially rises but at a decreasing rate. This means that while higher frequency sounds have higher pitch, the perceived increase in pitch slows down as frequency continues to rise. This is due to the physiological structure of the cochlea; the basilar membrane in the cochlea can perceive different frequencies, and the resonance points on the basilar membrane are proportional to the distance from the cochlear aperture, which creates a nonlinear relationship between the pitch perceived by the human ear and the frequency of the sound. In order to describe this nonlinear relationship, the Mel frequency was introduced, and the unit of pitch was specified as Mel (Shan et al., 2022). the Mel frequency has a nonlinear correspondence with the sound frequency, and the distance from the resonance point on the basilar membrane to the cochlear aperture is proportional to the sound frequency of its corresponding sound frequency, so the Mel frequency can reflect the pitch level.

$$Mel(f) = 2595 \lg\left(1 + \frac{f}{700}\right) \tag{20}$$

where f is the sound frequency and Mel(f) is the Mel frequency.

The Mel Cepstrum coefficients (MFCC) are based on the Mel frequencies and are proposed in conjunction with cepstrum analysis. The calculation of the MFCC utilises a Mel filter bank to simulate the nonlinear properties of the human ear (Deng et al., 2020). Because each band component acts superimposed in the human ear, it is also necessary to superimpose the energy within each filter band, so the Mel filter bank is designed as a set of filters superimposed on each other, the simplest of which is the interleaved rectangular filter, which treats each frequency component equally but in reality does not have the same effect on different frequency components (Zhu et al., 2017). A triangular filter weights the different frequency components, giving different weights to the different frequency components. The parameters of the triangular filter bank include the lower limit frequency f_l, the upper limit frequency f_h, and the centre frequency f_m. In the Mel filter bank, the lower limit frequency of a filter serves as the upper limit frequency of the preceding filter, while the upper limit frequency acts as the centre frequency of the subsequent filter. For the l-th filter in the Mel filter bank, let the lower limit frequency, upper limit frequency, and centre frequency be denoted as $f_l(i)$, $f_h(i)$ and $f_m(i)$, respectively, so that the Mel filter bank can be expressed as:

$$\begin{cases} f_l(i) = f_m(i-1) = f_h(i-2) \\ f_h(i) = f_m(i+1) = f_l(i+2) \\ f_l(1) = 0 \\ f_h(L) = f_{max} \end{cases} \tag{21}$$

where L is the total number of filters and f_{max} is the highest frequency of the signal.

Due to the nonlinear and masking effects of human ear hearing, when the low frequency end, Mel frequency and sound frequency are roughly linear, while at the high frequency end Mel is logarithmic; at the low frequency end the critical bandwidth is linear with the centre frequency, and at the high frequency end it is nonlinear with the centre frequency. In the design of the Mel filter bank, the centre frequencies of the filters in the low-frequency range are distributed linearly, while at the high-frequency end, the relationship between the centre frequency and the filter index is exponential. Its frequency response is:

$$H_m(k) = \begin{cases} 0, k \le f_m(i-1) \\ \dfrac{k - f_m(i-1)}{f_m(i) - f_m(i-1)}, f_m(i-1) \le k \le f_m(i) \\ \dfrac{f_m(i+1) - k}{f_m(i+1) - f_m(i)}, f_m(i) \le k \le f_m(i+1) \\ 0, k > f_m(i+1) \end{cases}, 0 \le i \le M-1 \tag{22}$$

where M is the number of Mel filters and the centre frequency $f_m = \left(\frac{N}{f_s}\right) B^{-1}\left(B(f_1) + m\frac{B(f_h) - B(f_l)}{M+1}\right)$.

3.2 Noise classification methods

SVM is a classical machine learning method known for its simple construction, strong learning ability, and good generalisation performance. Its core idea is to map low-dimensional sample features to high-dimensional space using a kernel function, leveraging the advantages of high-dimensional space to solve classification difficulties in the low-dimensional feature space (Blanco et al., 2022). This process involves finding the optimal hyperplane in the high-dimensional space that maximises the margin between different data samples. Suppose the data samples are (x_i, y_i), $i = 1, 2, \ldots, N$, N is the number of samples, and the feature vector $x_i \in R^n$. The category label is y_i, $y_i \in \{-1, 1\}$ which represents two different classes respectively (Li et al., 2018). Then we need to find a classification surface $w^T x + b = 0$ can distinguish the data of the two classes of samples, and make the interval between these two classes of samples the maximum, i.e., $y_i(w^T x + b) \ge 1$, $i = 1, 2, \ldots, N$. And the classification interval of the two classes of samples is $d = 2/||w||$, to make the classification interval distance d maximum, equivalent to the minimum of , that is, to find the minimum value of $||w||^2$ or $||w||^2/2$, then the optimal hyperplane solution equations can be transformed into the optimal problem with constraints to be solved.

$$\begin{cases} min\dfrac{1}{2}\|w\|^2 \\ s.t.y_i(w^T x_i + b) \geq 1, i = 1, \ldots, N \end{cases} \tag{23}$$

Solve for the minimum value using Lagrangian:

$$L(w,a,b) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} a_i \left[y_i \left(wx_i + b \right) - 1 \right] \tag{24}$$

Find the partial derivatives of w and b for $L(w, a, b)$ and make them equal to 0, respectively, and then follow the KKT complementarity condition:

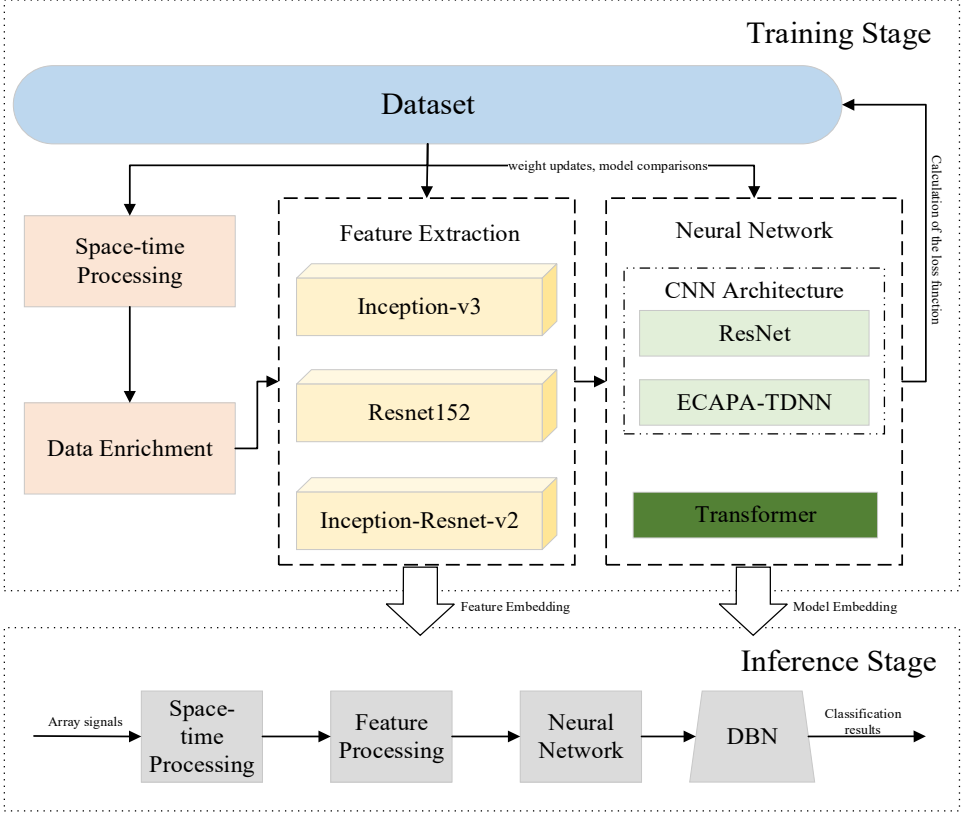
$$\begin{cases} w^* = \sum_{i=1}^{n} a_i^* y_i x_i \\ b^* = y_j - \sum_{i=1}^{n} a_i^* y_i x_i x_j \end{cases} \tag{25}$$

3.3 Acoustic noise identification model for ancient buildings

In the field of acoustic noise recognition of ancient buildings, facing the complex and changing acoustic environment and a wide variety of noise sources, the traditional single recognition method is often limited by low accuracy and insufficient robustness. In view of this, this paper innovatively proposes a recognition method framework based on space time joint processing and deep learning, aiming to give full play to the advantages of each technology to realise efficient and accurate recognition of acoustic noise of ancient buildings. First of all, to overcome the problem of inconspicuous target noise features and complex interference components in the original acoustic signals, MVDR spatial joint processing model is introduced to pre-process the acquired multi-channel acoustic signals. MVDR, with its powerful spatial filtering capability, can effectively enhance the target acoustic features, weaken the background noise and interference, and provide a clearer basis for analysing the acoustic features after the spatial joint processing. Then, using Mel frequency feature extraction technology, the pre-processed acoustic signal is converted to the Mel frequency domain, taking full account of the auditory characteristics of the human ear, highlighting the more recognisable feature components in the acoustic signal, so that the subsequent feature learning is more targeted. On this basis, the MFA-conformer network integrates Mel-frequency Cepstral coefficients with the conformer model. The conformer model's combination of convolution and self-attention mechanisms allows it to capture both local and global features of the audio signals. This network takes the Mel-frequency features extracted from the spatiotemporal processing as inputs. The convolution layers in the conformer capture local patterns, while the self-attention layers model long-range dependencies. This dual approach enables the network to comprehensively analyse the acoustic noise characteristics. On this basis, with the powerful feature learning capability of CNN, the deep complex patterns and associations in the acoustic features are automatically extracted, and the abstract representation of the features is gradually extracted by CNN through convolutional layers, pooling layers and other operations, and finally the mapping of feature vectors is

realised through the fully-connected layer. The feature vectors extracted by the CNN are fed into the SVM classifier, enabling precise classification of various noise types and thus completing the task of recognising the acoustic noise of ancient buildings. This method framework provides a new method for the recognition of acoustic noise of ancient buildings through the integration of multiple technologies, and its specific flow is shown in Figure 1. The connection between the steps is seamless, with data transferred directly from one stage to the next. Specifically, the pre-processed multi-channel audio signals from the MVDR model are first converted into Mel-frequency features. These features are then fed into the CNN, where they undergo convolution and pooling operations to extract deeper feature representations. The resulting feature vectors are subsequently transferred to the SVM classifier for final noise type identification. This streamlined data transfer ensures that each step builds upon the previous one, maintaining the integrity and continuity of the recognition process.

Figure 1 Framework diagram of acoustic noise identification method for ancient buildings (see online version for colours)



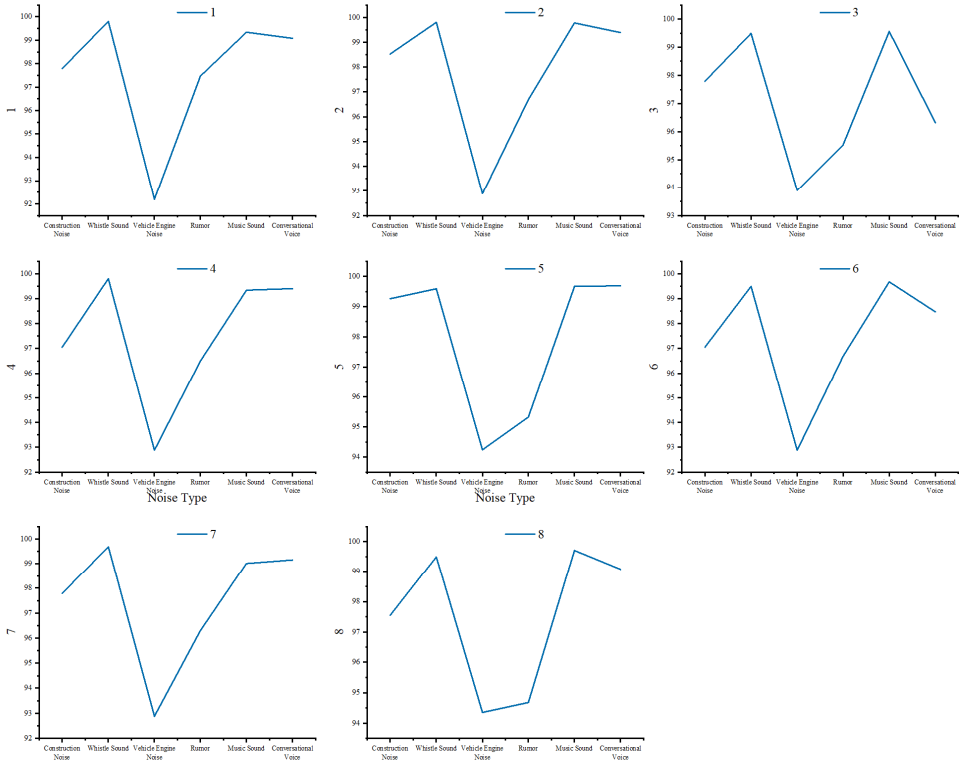
4 Experimental results and analyses

In the field of acoustic noise recognition of ancient buildings, in the face of the complex and changing acoustic environment and a wide variety of noise sources, the traditional single recognition method is often limited by low accuracy and insufficient robustness. To verify the effectiveness of the proposed noise recognition method for ancient buildings, this study conducts a series of experiments. A 3-layer RBM is employed to explore the impact of different hidden layer nodes on the results. Except for the number of hidden layer nodes, other parameters of the DBN are kept constant. The data augmentation methods used in this study include adding background noise and varying the speed of audio recordings. The regularisation technique is implemented using L2 regularisation with a weight decay parameter of 0.0001. Specifically, the network's loss function is set to the cross-entropy loss function, with a learning rate of 1e-3. The dropout rate is 0.12. The hidden layer uses a sigmoid activation function, while the output layer employs a softmax activation function. The network undergoes 30 iterations, utilising a small batch stochastic gradient descent algorithm with a batch size of 32. In the RBM training phase, the initial learning rate is 1e-3, and the number of iterations is 10. The parameter of the contrast-scattering algorithm is set to 1. The network weights are initialised randomly through a normal distribution with a standard deviation of $\sqrt{2/(n_v+n_h)}$, where n_v represents the number of visible layer nodes and n_h denotes the number of hidden layer nodes. Then set the number of nodes in the hidden layer separately, in this paper we set the number of nodes in the hidden layer as 300-200-100, 500-300-200, 750-500-250, 1000-500-250, 1200-600-300, 1500-750-350, 1800-900-450, 2000-1000-500 for experiments. The experimental data is collected from multiple ancient building sites, capturing various noise sources such as construction noise, vehicle engine noise, and conversational voices. These sounds vary in intensity, frequency, and propagation patterns due to the unique spatial structures and materials of ancient buildings. 1000-500 are experimented and the results are shown in Table 1 and Figure 2.

Table 1 Recognition rate of urban environmental noise with different numbers of hidden layer nodes in DBN

Noise type	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Construction noise	97.79	98.53	97.79	97.06	99.26	97.06	97.79	97.55
Whistle sound	99.80	99.80	99.49	99.80	99.59	99.49	99.69	99.49
Vehicle engine noise	92.20	92.88	93.90	92.88	94.24	92.88	92.88	94.35
Rumor	97.47	96.69	95.53	96.50	95.33	96.69	96.30	94.68
Music sound	99.34	99.78	99.56	99.34	99.67	99.67	99.01	99.71
Conversational voice	99.08	99.39	96.32	99.40	99.69	98.47	99.16	99.08

Figure 2 Recognition rate of urban environmental noise with different numbers of hidden layer nodes in DBN (see online version for colours)



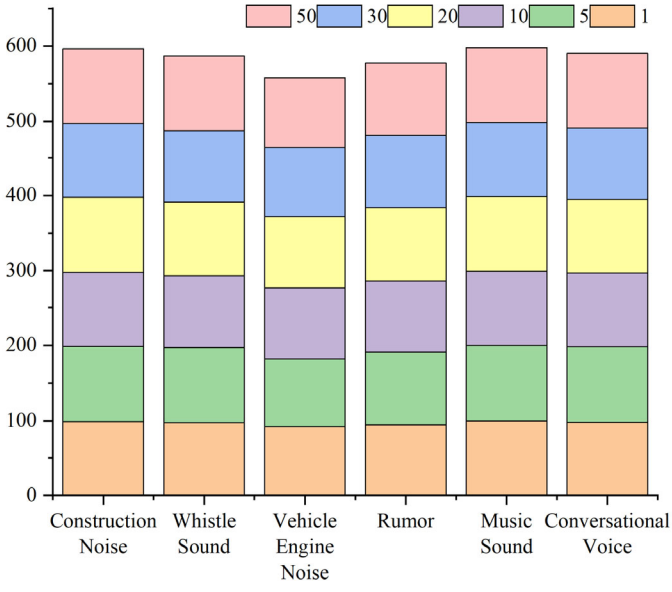
As the number of hidden layer nodes increases, the model's recognition accuracy for acoustic noise in ancient buildings initially rises gradually, then fluctuates after reaching a peak. This suggests that when the number of hidden layer nodes reaches a certain level, the model can better capture the complex features of acoustic noise and improve recognition performance. However, when the number of hidden layer nodes is too high, the recognition accuracy of some noise types may fluctuate or even slightly decrease. This could be due to overfitting, where the model overlearns the noise features in the training data, reducing its generalisation ability on test data. Overall, the number of hidden layer nodes significantly impacts model performance. A moderate number of nodes can balance model complexity and generalisation ability, achieving higher recognition accuracy. In contrast, too few or too many nodes may lead to decreased model performance.

In order to study whether the number of RBM iterations affects the classification results of urban acoustic signals, this paper analyses the effect of 1~50 iterations of experiments on the recognition of urban environmental noise, which are set respectively during the unsupervised training of the RBM, where the number of RBM iterations of the experiments are 1, 5, 10, 20, 30, and 50. Based on the aforementioned experimental outcomes, the hidden layer parameters of the Deep Belief Network are configured to 1500-750-350, while all other parameters remain constant. The corresponding results are presented in Table 2 and Figure 3.

Table 2 Recognition rate of urban environmental noise with different RBM iteration times

Noise type	*1*	*5*	*10*	*20*	*30*	*50*
Construction noise	99.31	99.21	99.51	99.51	99.51	99.51
Whistle sound	97.79	99.26	96.32	97.79	96.32	99.26
Vehicle engine noise	92.88	89.15	95.25	94.58	91.86	94.24
Rumor	94.94	96.11	95.33	97.67	97.28	96.30
Music sound	99.89	99.67	99.45	99.45	99.89	99.56
Conversational voice	98.16	99.69	99.08	98.16	96.01	99.39

Figure 3 Recognition rate of urban environmental noise with different RBM iteration times (see online version for colours)



Through the above experiments, it is found that the best result is achieved when the number of RBM iterations is 10, but it can also be seen that DBN is not sensitive to the number of RBM iterations in the classification of acoustic noise recognition of ancient buildings.

5 Conclusions

In this paper, we propose an innovative method that integrates space time joint processing and deep learning for the difficult problem of acoustic noise recognition of ancient buildings. Traditional noise recognition means are not precise enough for feature extraction in complex scenes, and the generalisation ability is insufficient. In this study, acoustic signals are processed finely through space time joint processing, and their space time correlation features are mined in depth, and then accurate classification and recognition is realised with the powerful ability of deep learning models. The

experimental results strongly prove the effectiveness and superiority of the method, which shows good robustness under different environmental conditions. This not only provides strong technical support for the acoustic protection of ancient buildings, but also opens up new ideas and methods in the field of cultural heritage protection, which can help to carry out the maintenance and repair of ancient buildings more scientifically and revitalise them in the modern society, and at the same time, the method has a certain degree of universality, which can be a useful reference for the recognition of noise in other complex acoustic environments.

Acknowledgements

This work is supported by the China Scholarship Council (No. 202206935003) and the Natural and Science foundation of Shanxi Province (No. 202203021222098).

Declarations

All authors declare that they have no conflicts of interest.

References

Barakat, A., Naqra, A. and Hussian, A.R. (2024) 'Measuring the accuracy of a voiceprint analysis system designed by applying the euclidean distance function and genetic algorithm', *International Journal of Electrical and Electronics Engineering*, Vol. 11, No. 3, pp.220–230.

Blanco, V., Japon, A. and Puerto, J. (2022) 'A mathematical programming approach to SVM-based classification with label noise', *Computers and Industrial Engineering*, Vol. 172, p.108611.

Dai, X., Nie, J., Chen, F. and Ou, G. (2017) 'Distortionless space-time adaptive processor based on MVDR beamformer for GNSS receiver', *IET Radar, Sonar and Navigation*, Vol. 11, No. 10, pp.1488–1494.

Demir, F., Abdullah, D.A. and Sengur, A. (2020) 'A new deep CNN model for environmental sound classification', *IEEE Access*, Vol. 8, pp.66529–66537.

Deng, M., Meng, T., Cao, J., Wang, S., Zhang, J. and Fan, H. (2020) 'Heart sound classification based on improved MFCC features and convolutional recurrent neural networks', *Neural Networks*, Vol. 130, pp.22–32.

Gibbs, B. (2010) 'Collected papers in building acoustics: room acoustics and environmental noise', *Journal of the Acoustical Society of America*, Vol. 128, No. 6, pp.3815–3815.

Jiang, W., Li, Z., Li, J., Zhu, Y. and Zhang, P. (2019) 'Study on a fault identification method of the hydraulic pump based on a combination of voiceprint characteristics and extreme learning machine', *Processes*, Vol. 7, No. 12, pp.894–894.

Jin, R., Ablimit, M. and Hamdulla, A. (2023) 'Speaker verification based on single channel speech separation', *IEEE Access*, Vol. 11, pp.112631–112638.

Khdier, H.Y., Jasim, W.M. and Aliesawi, S.A. (2021) 'Deep learning algorithms based voiceprint recognition system in noisy environment', *Journal of Physics: Conference Series*, Vol. 1804, No. 1, p.12042.

Klemm, R. (1999) 'Introduction to space-time adaptive processing', *Electronics and Communication Engineering Journal*, Vol. 11, No. 1, pp.5–12.

Kunjir, A. (2022) 'Voiceprint recognition based on machine learning methods', *Journal of Science and Engineering Research*, Vol. 1, No. 1, pp.29–46.

Li, X., Yang, X. and Zhou, L. (2020) 'Adaptive threshold estimation of open set voiceprint recognition based on OTSU and deep learning', *Journal of Applied Mathematics and Physics*, Vol. 8, No. 11, pp.2671–2682.

Li, X., Zhu, Y., Wang, J., Liu, Z., Liu, Y. and Zhang, M. (2018) 'On the soundness and security of privacy-preserving SVM for outsourcing data classification', *IEEE Transactions on Dependable and Secure Computing*, Vol. 15, No. 5, pp.906–912.

Shan, S., Liu, J., Wu, S., Shao, Y. and Li, H. (2022) 'A motor bearing fault voiceprint recognition method based on Mel-CNN model', *Measurement: Journal of the International Measurement Confederation*, Vol. 207, pp.112408.

Shaw, A., Smith, J. and Hassanien, A. (2021) 'MVDR beamformer design by imposing unit circle roots constraints for uniform linear arrays', *IEEE Transactions on Signal Processing*, Vol. 69, pp.6116–6130.

Tang, Y. and Sun, X. (2023) 'Research on voiceprint recognition based on densenet', *Frontiers in Computing and Intelligent Systems*, Vol. 6, No. 2, pp.80–85.

Yang, X., Zhu, X. and Zhu, X. (2016) 'Temperament accuracy judgment and intelligent processing based on voiceprint recognition', *TELKOMNIKA: Indonesian Journal of Electrical Engineering*, Vol. 14, No. 2A, p.357.

Yao, J.C., Zhang, J., Li, J.F. and Zhuo, L. (2021) 'Anchor voiceprint recognition in live streaming via RawNet-SA and gated recurrent unit', *Eurasip Journal on Audio, Speech, and Music Processing*, Vol. 2021, No. 1, pp.1–18.

Zhang, X., Tang, J., Cao, H., Wang, C., Shen, C. and Liu, J. (2024) 'Cascaded speech separation denoising and dereverberation using attention and TCN-WPE networks for speech devices', *IEEE Internet of Things Journal*, Vol. 11, No. 10, pp.18047–18058.

Zhou, R., Zhao, S., Luo, M., Meng, X., Ma, J. and Liu, J. (2024) 'MFCC based real-time speech reproduction and recognition using distributed acoustic sensing technology', *Optoelectronics Letters*, Vol. 20, No. 4, pp.222–227.

Zhu, G., Zhou, Y., Si, Z., Cheng, Y., Wu, F., Wang, H., Pan, Y., Xie, J., Li, C., Chen, A., Wang, R. and Sun, J. (2023) 'A multi-hole resonator enhanced acoustic energy harvester for ultra-high electrical output and machine-learning-assisted intelligent voice sensing', *Nano Energy*, Vol. 108, p.108237.

Zhu, L., Chen, L., Zhao, D., Zhou, J. and Zhang, W. (2017) 'Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN', *Sensors*, Vol. 17, No. 7, pp.1694–1694.