Hindawi
Structural Control and Health Monitoring
Volume 2024, Article ID 1270912, 23 pages
https://doi.org/10.1155/2024/1270912

Research Article

Fault Detection of In-Service Bridge Expansion Joint Based on Voiceprint Recognition

Yiqing Dong,[1,2] **Dalei Wang**,[1,3] **Yue Pan**,[1] **Jin Di**,[4] **and Airong Chen**[1]

[1]*College of Civil Engineering, Tongji University, Shanghai 200092, China*
[2]*School of Civil and Environmental Engineering, Nanyang Technological University, Singapore 639815*
[3]*Key Laboratory of Performance Evolution and Control for Engineering Structures (Ministry of Education), Tongji University, Shanghai 200092, China*
[4]*School of Civil Engineering, Chongqing University, Chongqing 400044, China*

Correspondence should be addressed to Yue Pan; pan_yue@tongji.edu.cn

Received 3 December 2023; Revised 15 January 2024; Accepted 2 March 2024; Published 15 March 2024

Academic Editor: Wenai Shen

Bridge expansion joints (BEJs) in service are susceptible to damage from various factors such as fatigue, impact, and environmental conditions. While visual inspection is the most common approach for inspecting BEJs, it is subjective and labor-intensive. In this paper, we propose a novel methodology for detecting the fault status of BEJs, inspired by voiceprint recognition (VPR) based on audio signals. We establish an Artificial Neural Network to filter nonevent segments from low signal-to-noise ratio signals, achieving an AuC value of 0.981. We design and improve ConFormer VPR models with a multifeature aggregation strategy and cascade them to realize fault detection of BEJs. For three successive tasks in classifying environment sound types, vehicle impact types, and faults, the ConFormer VPR models achieve AuC values of 0.975, 0.925, and 0.886, respectively, demonstrating the feasibility of our methods for unmanned inspection of BEJs. In future research, the introduction of multiple types of damage and the implementation of benchmarking tests are planned to further enhance the capabilities of the system.

1. Introduction

Bridge expansion joints (BEJs) are structural components that allow bridges to accommodate thermal movements and vibrations caused by traffic loads, wind forces, and seismic events. However, BEJs are also vulnerable to deterioration and damage due to various factors, such as fatigue, corrosion, impact, and environmental conditions [1, 2]. Damaged expansion joints can compromise the structural integrity of the bridge, create noise and vibration problems, and pose a hazard to traffic [3, 4]. Therefore, inspecting BEJs is essential for ensuring the safety and functionality of the bridge, as well as preventing costly and disruptive repairs in the future.

Currently, the inspection of BEJs primarily relies on two methods: manual visual inspection and nondestructive testing (NDT) techniques [5, 6]. Trained professionals conduct visual inspections, employing methods such as hammer testing for bolt looseness, magnifying glasses or microscopes for crack detection, and micrometers to assess beam gap distances [7]. Despite its simplicity and efficiency, manual visual inspection is inherently subjective, contingent upon the inspector's expertise, skill, and fatigue. Furthermore, this approach has limitations, as it may overlook concealed defects beneath the surface or within the joint.

Consequently, NDT methods are being explored for a more objective and precise inspection of BEJs. Accelerometers or displacement gauges are installed to capture the dynamic response of BEJs. This allows for the extraction of vibration characteristics induced by moving vehicles, changes which are detected for fault identification in damaged BEJ steel fingers [8]. A wireless installation, based on Internet-of-Things technology, has also been proposed to enhance the flexibility of this approach [9, 10]. Ultrasonic testing, proposed for inspecting steel conditions [11–13], can be employed to pinpoint the damaged areas of steel

components in BEJs. X-ray Computed Tomography (X-ray CT) is being investigated to inspect the corrosion zones of BEJ steel specimens in a laboratory setting [14]. Corrosion products are observed in X-ray CT scanning images across different vertical sections. Electromagnetic testing, a novel NDT technique for detecting defects in steel components [15], can be utilized to inspect support bars in BEJs. However, these methods necessitate expensive specialized equipment and trained personnel. Some methods also require lane closures and traffic control, thereby escalating the cost and duration of the inspection process.

Previous research has investigated the sound produced by vehicle impact on BEJs, with a focus on understanding its generation [16] and devising methods to mitigate it for environmental conservation [17, 18]. Findings suggest that the audio response of BEJs contains information about their operational status, as corroborated by experienced inspectors who have leveraged these sounds to pinpoint anomalous BEJs [19]. Consequently, sound signals elicited by vehicle impact hold the potential for detecting BEJ damage. However, there is a paucity of studies on BEJ fault inspection that utilize audio signals.

In recent decades, sound signals have been increasingly utilized for detecting damage or anomalies in infrastructures that are in service. To acquire these sound signals, a microphone or an array of microphones is required. Research on auditory perception has shown that most of the information conveyed by sound signals is below 10 kHz [20]. As a result, a sampling rate of 16 kHz is commonly used in most scenarios [21]. However, for applications that prioritize efficiency and are based on audio, a sampling rate of 8 kHz can also be practical [22, 23]. Then, the original audio signals are processed using digital signal processing (DSP) techniques, such as Fourier Transform, to extract high-level features. Given the unique characteristics of audio signals, specific features have also been defined, including Mel frequency, Fbank [24], Gammatone [25], and so on.

After signal preprocessing, machine learning methods can be used to detect faults in these infrastructures. In the early stages, Hidden Markov Models (HMMs) are used to classify audio signals by obtaining Mel cepstrum coefficients, resulting in a 58% accuracy rate for 18 types of sounds [26, 27]. For more generalized usage, statistical learning models are applied in fault detection. The Support Vector Machine (SVM) is widely used for detecting damage to structures or equipment, such as bolt loosening [28], pipeline cracking [29, 30], bearing faults [31], ratchet faults [32], and turbine blade damage [33]. A decision tree is designed to classify damage events of wind turbine rotor blades using airborne sound, and the results show the high precision of the algorithm [34]. Artificial Neural Network (ANN) model is selected to analyze in-pipe leak detection and bearing fault inspection, resulting in high precision [35]. In recent years, deep learning models have been introduced into fault detection [36, 37] as they can obtain deeper features of signals with greater efficiency. One-Dimensional Convolution Neural Network (1D-CNN) is first applied to audio signal classification [38, 39]. To exploit the features of time-series signals, a recurrent module is added to the 1D-CNN model, improving its classification accuracy [40]. Additionally, a Long Short-Term Memory model is utilized for fault detection in additive manufacturing in industrial applications [41]. By applying WaveNet [42], the accuracy of fault detection is improved compared to the LSTM model [43]. For special occasions with few fault data, unsupervised methods are proposed to monitor the operation state of machines [44–47]. These approaches are available in non-interrupting surroundings where only the sounds of monitored devices are acquired.

In conclusion, existing research suggests that manual visual inspection remains the most prevalent method for inspecting in-service BEJs. Techniques such as accelerometers, ultrasonic sensors, and electromagnetic sensors have been explored for detecting BEJ faults, aiming to inspect the BEJs in an NDT manner. However, these techniques are labor-intensive and cost-intensive and necessitate professional training or practical analysis. Given its proficiency in anomaly detection tasks and the application of cutting-edge deep learning techniques for audio analysis, the audio signal-based approach shows promise as a more cost-effective and automated method for detecting faults in in-service BEJs. For BEJs that are in service, the environment is noisier and more complex than in a factory or laboratory. As a result, more robust and feasible event segmentation methods and fault detection models are required. Recent advancements in speech recognition and voiceprint recognition (VPR) for human voices can be utilized for improvement, especially VPR which also focuses on the audio signal classification problem.

In this study, we propose a novel framework for fault detection of in-service BEJs based on VPR. First, microphones are deployed under the BEJs to acquire audio data. Then, an Artificial Neural Network (ANN) classification model is established to filter nonevent segments from low SNR audio signals in the BEJ environment. Subsequently, the ConFormer VPR model is designed and improved for general audio event classification. Finally, a cascading approach is proposed, consisting of three successive ConFormer VPR models, to separate vehicle impact audio, distinguish the detailed type of vehicle, and ultimately detect fault status. The main contributions of this work are fourfold. (1) Acoustic sensors and VPR algorithms are introduced for the first time in the fault detection of BEJs. (2) A machine learning model is applied for audio signal event segmentation, achieving an AuC of 0.981. (3) The original ConFormer is modified for VPR and improved with a multifeature aggregation strategy. (4) A cascading approach is proposed by combining ConFormer models for fault detection of BEJs.

The remainder of this work is organized as follows. Section "Overview of our methodology" presents our methodology for fault detection of BEJs based on the VPR technique. In Section "Audio event segmentation based on Fbank-ANN model," the Fbank feature and ANN model for audio signal event segmentation are introduced. Section "Fault detection of BEJs by cascading ConFormer VPR models" proposes the ConFormer structure for VPR and a cascading approach consisting of ConFormer models for

complete fault detection. Furthermore, a case study is conducted and the results are discussed in Section "Case study." Finally, conclusions are drawn in Section "Conclusions."

2. Overview of Our Methodology

In this paper, we propose a framework for fault detection of in-service BEJs based on audio processing techniques. As shown in Figure 1, the methodology consists of four major parts.

First, a microphone is deployed under the BEJ to acquire audio data. Additionally, some auxiliary sensors are necessary for annotation purposes, such as cameras to annotate actual passing vehicles for sound collection. The audio data are then preprocessed using the Fbank approach, which includes preemphasis, framing and windowing, Short-Term Fourier Transform (STFT), and Mel filters. The resulting Fbank feature maps are used to train an ANN classification model for audio event segmentation in actual applications. Finally, fault detection can be achieved by cascading ConFormer VPR models. The first ConFormer is applied for environment VPR to separate vehicle impact audio signals. The second ConFormer is used to distinguish the detailed type of vehicle impact. The last ConFormer serves for the final fault detection under the same vehicle impact type.

3. Audio Event Segmentation Based on Fbank-ANN Model

Exposed to the outdoor environment, BEJs are surrounded by multiple types of sounds caused by four factors, as listed in Table 1.

Nonevent segments in audio signals represent meaningless information and should be eliminated. Threshold-based methods, such as Short-Term Energy (STE) and Short-Term Cross-Zero Rate (STCZR), are popular approaches for audio event segmentation. However, in low signal-to-noise ratio (SNR) environments such as BEJs, these methods may not perform well for event segmentation [48, 49].

In this paper, we apply a machine learning (ML) approach to event segmentation of audio signals by treating it as a classification problem. Firstly, Fbank feature extraction is used for audio preprocessing, with event and nonevent samples manually annotated. Then, an ANN model can be trained to distinguish event features from nonevent features.

3.1. Fbank Feature. The audio signal data have ultra-high frequency and rich semantics, making its feature extraction approach more complex than other signal data in the bridge monitoring system. By investigating the theory of auditory perception, the Fbank feature descriptor has been proposed and realized through four calculation procedures [24].

Step 1. Preemphasis. The high-frequency components of audio signals are significant in VPR. However, the uniform discretization process of signal sampling can cause attenuation of energy in high-frequency regions. Therefore, a preemphasis procedure is required to boost energy and highlight resonant peaks in these regions. Specifically, a first-order digital filter can be used to achieve this compensation, as follows:

$$x_p(n) = x(n) - \alpha x(n-1), \tag{1}$$

where n is discrete time coordinate, $x_p(*)$ and $x(*)$ are audio signals after and before the preemphasis procedure, respectively, and α is the factor of emphasis.

Step 2. Framing and Windowing. The process of audio signal generation has inertia, indicating that the audio signal is stationary in a short time period and exhibits short-time stationarity. Therefore, the original high-frequency signal needs to be segmented into multiple short-duration audio segments called speech frames. Frame-length and frame-shift should be determined in the process. The former refers to the duration of the audio frame, while the latter defines the overlap between adjacent frames. Since framing is a truncation of signals that can cause spectral leakage, audio frames need to be windowed to reduce edge weighting and avoid the Gibbs phenomenon. A commonly used window function for audio signals is the Hamming window:

$$w(n) = 0.54 - 0.46\cos\frac{2\pi n}{N-1}, \tag{2}$$

where $w(*)$ is the window function and N is the sample number of the audio frame.

Step 3. STFT. The audio frames are processed by K-point STFT to obtain the frequency-domain responses of audio frames, as shown in equation (3). Here $x_t(n)$ represents the n-th signal of the t-th frame. Then, the power spectrum of audio frames can be further calculated using equation (4).

$$X_t(k) = \sum_{n=0}^{K-1} x_t(n)\exp\left\{-j\frac{2\pi nk}{K}\right\},\ k = 0, 1, \ldots, K-1, \tag{3}$$

$$P_t(k) = \frac{|X_t(k)|^2}{K}, k = 1, 2, \ldots, \frac{K}{2} + 1. \tag{4}$$

Step 4. Mel Filter. The frequency-domain responses of audio frames are further processed by Mel filters. Mel filters are mathematically triangular bandpass filters, set one by one according to linear intervals of Mel frequency. The center frequency interval between adjacent triangular filters can be obtained using the following equation:

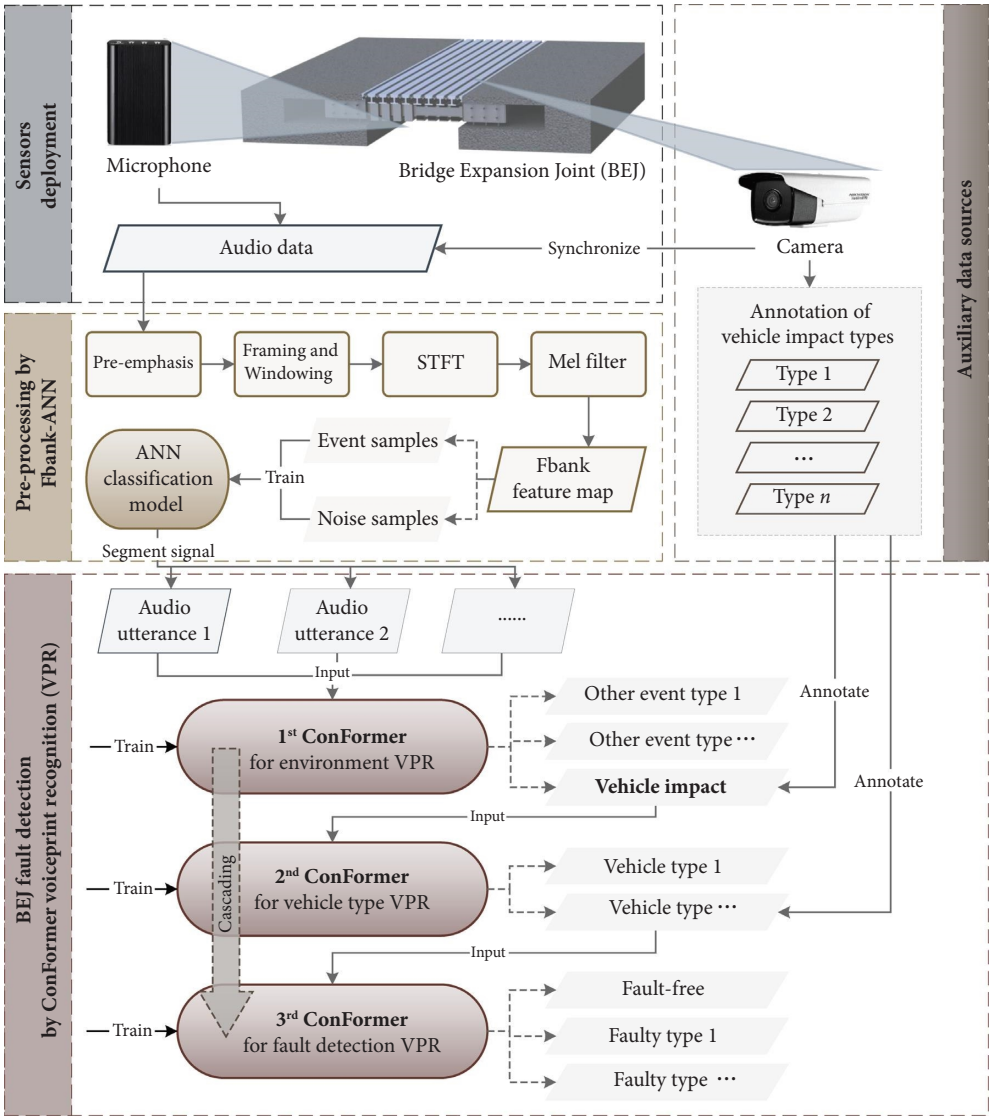


Figure 1: Overview of our methodology.

Table 1: Sounds in BEJ environment.

Factor	Description
Vehicle factors	Audio events generated by vehicles passing over BEJs include wheel impact sounds, vehicle horn sounds, and more
Human factors	Audio events, such as the sound of personnel speaking, can occur due to the passage of inspectors near BEJs
Bird and animal factors	Audio events, such as bird sounds, can occur due to birds and animals passing near the expansion joints
Noise factors	Noise can be generated by the aerodynamic factors of the environment in which it is located, such as urban ambient noise and beach ambient noise

$$\Delta = \frac{\mathrm{Mel}(f_{\max}) - \mathrm{Mel}(f_{\min})}{M+1}\mathrm{Mel}(f) = 2595\log_{10}\left(1+\frac{f}{700}\right), \quad (5)$$

where M is the number of triangular filters and $f_{\max}$ and $f_{\min}$ are the maximum and minimum real frequencies, respectively. The lower frequency $f_{\mathrm{L}}(m)$, center frequency $f_{\mathrm{mid}}(m)$, and higher frequency $f_{\mathrm{H}}(m)$ of the m-th triangular filter are defined as follows:

$$\begin{aligned} f_L &= f_{\min} + (m-1) \times \mathrm{Mel}^{-1}\Delta, \\ f_H &= f_{\min} + (m+1) \times \mathrm{Mel}^{-1}\Delta \\ f_{\mathrm{mid}} &= f_{\min} + m \times \mathrm{Mel}^{-1}\Delta. \end{aligned} \tag{6}$$

Here $m = 1, 2, \ldots, M$, so the response of the m-th triangular filter in Mel filter bank at frequency γ is

$$H_m(\gamma) = \begin{cases} 0, & \gamma < f_L(m), \\ \dfrac{\gamma - f_{\mathrm{mid}}}{f_{\mathrm{mid}} - f_L}, & f_L \le \gamma \le f_{\mathrm{mid}}(m), \\ \dfrac{f_{\mathrm{mid}} - \gamma}{f_H - f_{\mathrm{mid}}}, & f_{\mathrm{mid}} \le \gamma \le f_H(m), \\ 0, & \gamma > f_H(m), \end{cases} \tag{7}$$

where γ is in the range $[f_{\min}, f_{\max}]$. Figure 2 shows the Mel filter bank with $M = 10$. It can be seen that the filtering range widens at high frequencies. The upper limit amplitudes of the filter are the same to simultaneously retain low and high frequency information of the audio.

Then, the power spectrum of the audio signal after STFT is input into the Mel filter bank to extract the Mel power spectrum of the audio, calculated using the following equation:

$$\begin{aligned} C_t(m) &= \sum_{k=1}^{K/2+1} P_t(k) H_m(\phi(k)), m = 1, 2, \ldots, M, \\ \phi(k) &= (k-1) \times \frac{f_s}{K}, k = 1, 2, \ldots, \frac{K}{2} + 1, \end{aligned} \tag{8}$$

where t and m are the number of frames and triangular filters, separately.

Furthermore, the obtained Mel power spectrum is subjected to a logarithmic operation to obtain the final Fbank feature of the t-th frame along the m-th dimension.

$$h_t(m) = \ln C_t(m), m = 1, 2, \ldots, M. \tag{9}$$

An illustration of Fbank feature acquisition is shown in Figure 3. The 2-second audio signals are processed into 200 frames with a frame-length of 25 ms and a frame-shift of 10 ms. After STFT, nonevent segments of the audio signals have low responses, as shown in Figure 3(b). Then, an 80-dimensional Mel filter bank is used to obtain Fbank features of the audio, as shown in Figure 3(c). The features will be used for further applications.

3.2. ANN Model for Distinguishing between Events and Nonevents. Audio signals are preprocessed to obtain Fbank features. A classification model is then required to distinguish nonevent features from event features. ANN is a type of ML model that can learn any nonlinear function. For an ANN to work, audio signals are annotated into event and nonevent segments and processed to obtain Fbank features frame by frame. The model is then built and trained to classify Fbank features of event or nonevent frames. Finally, the model is applied for event frame recognition, which is the purpose of event segmentation.

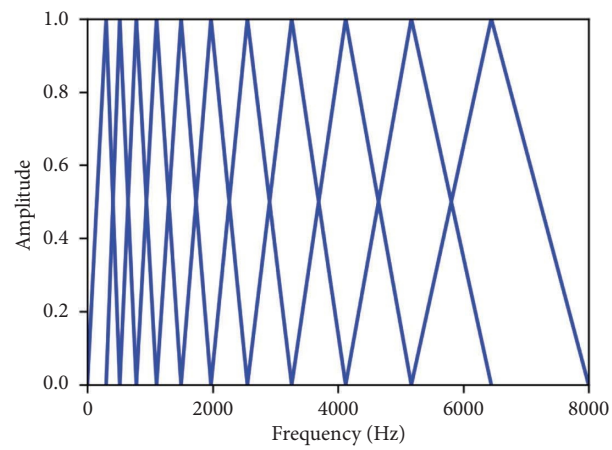


Figure 2: Mel filter bank composed of 10 triangular bandpass filters.

An ANN model is composed of an input layer, one or more hidden layers, and an output layer. The number of elements in the input and output layers is determined by the dimension of the feature vector and the number of prediction categories, respectively. The number of elements in the hidden layer(s) is determined through hyperparameter tuning. Mathematically, each layer consists of two algorithmic steps: linear weighted summation and nonlinear activation, as represented by the following equation:

$$\begin{aligned} \mathbf{m} &= \mathbf{w}\mathbf{x} + b, \\ \mathbf{a} &= \mathrm{Act}(\mathbf{m}) = \max \quad (\mathbf{0}, \mathbf{x}), \end{aligned} \tag{10}$$

where $\mathbf{x}$ represents the input vector, $\mathbf{w}$ and b are weight parameter vectors to be learned, $\mathrm{Act}(*)$ represents the activation function, which in this case is the ReLU function, and $\mathbf{a}$ is the output result of this layer and serves as the input vector for the next layer.

An ANN forms a complex nonlinear cascading structure through its multiple layers, enabling it to fit the mapping relationship between input and output. Specifically, the back-propagation mechanism [50] is used to tune the model by minimizing the error between the model's output and the true result, namely, the loss function $L(*)$:

$$L(\theta) = \frac{1}{n} \sum_{r=1}^{n} \| g(\mathbf{x}^r; \theta_i) - \mathbf{y}^r \|, \tag{11}$$

where θ represents the weighting parameters to be learned, n is the number of samples, $\mathbf{y}$ is the final output vector, and $g(*)$ represents the entire ANN model.

Furthermore, during the training process of an ANN, errors in the loss function with respect to each parameter are calculated. The error δ in the output layer is given by equation (12), while the errors of the parameters in the hidden and input layers are obtained through back-propagation using chain rule differentiation.

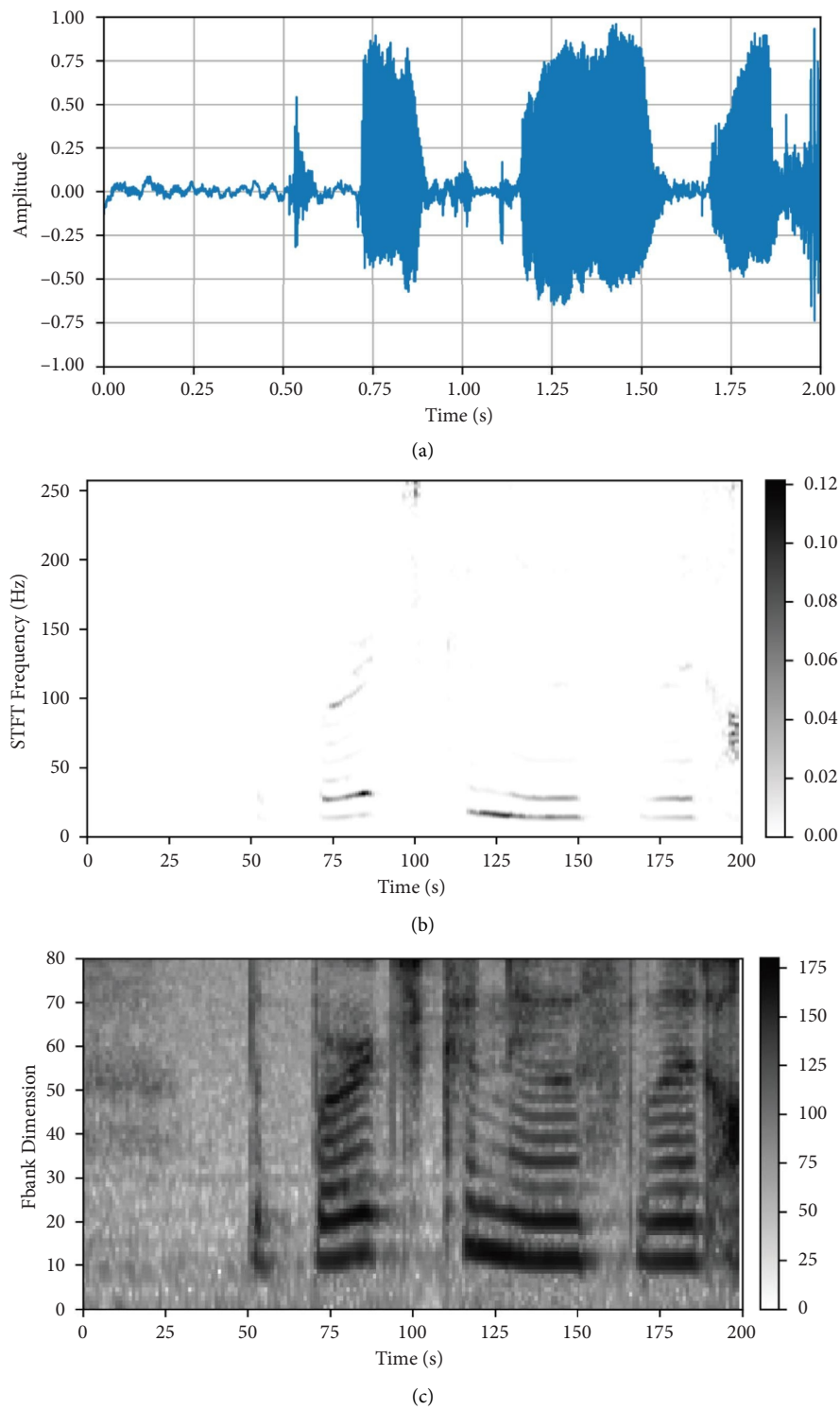


Figure 3: Illustration of Fbank feature acquisition. (a) Original audio signals. (b) Power spectrum by STFT (unit: dB). (c) Fbank feature (unit: dB).

$$\delta = \frac{\partial L(\theta)}{\partial \theta_i} = \frac{\partial L(\theta)}{\partial a(\theta_i)} \frac{\partial a(\theta_i)}{\partial \theta_i} = \frac{\partial L(\theta)}{\partial a(\theta_i)} \frac{\partial \mathrm{Act}(a(\theta_i))}{\partial \theta_i} = \frac{\partial L(\theta)}{\partial a(\theta_i)} \mathrm{Act}'(\theta_i). \tag{12}$$

Finally, the parameters are optimized using the Gradient Descent algorithm in each iteration. The coefficient α, defined in the range (0,1], represents the learning rate. The parameter optimization process is given by the following equation:

$$\theta := \theta - \alpha\delta = \theta - \alpha \frac{\partial L}{\partial \theta}. \tag{13}$$

An ANN for event segmentation is a binary classification model. A confusion matrix, defined in Table 2, is used to evaluate the performance of the ANN. The Accuracy, True Positive Rate (TPR), and False Positive Rate (FPR) of the model can be calculated via equation (14) by setting different classification thresholds. The Receiver Operating Characteristic (ROC) curve is obtained by plotting FPR on the horizontal axis and TPR on the vertical axis. The closer the ROC curve is to the upper left corner, the greater the model's ability to distinguish between different types of features, indicating higher generalization and robustness. The performance of the model is evaluated using the Area under Curve (AuC) value, which is a metric calculated from the ROC curve. The AuC represents the area enclosed under the ROC curve and provides a quantitative measure of the model's ability to distinguish between different classes [51, 52].

$$\begin{aligned} \mathrm{Accuracy} &= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} = \frac{\text{Number of correct identifications}}{\text{Total number of tests}}, \\ \mathrm{TPR} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{\text{Number of true positives}}{\text{Total number of positive samples}}, \\ \mathrm{FPR} &= \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = \frac{\text{Number of false positives}}{\text{Total number of negative samples}}. \end{aligned} \tag{14}$$

4. Fault Detection of BEJs by Cascading ConFormer VPR Models

Fault detection of BEJs can be treated as a voiceprint recognition (VPR) task, where audio signals are inputs and the normal/faulty status of BEJs are outputs. A VPR model is required to classify audio signals by BEJs to detect their status.

4.1. ConFormer Model with MFA Module. The Convolution-augmented Transformer network (ConFormer) was first proposed by Gulati [53] for speech recognition applications. The design idea of ConFormer is to combine the advantages of Transformer and CNN models. On the one hand, the encoder-decoder structure of the Transformer makes it effective at capturing global features based on content. On the other hand, CNNs excel at local feature extraction through multiple convolution-pooling operations. Therefore, ConFormer combines the benefits of both models to achieve unified modeling of global and local features in audio data analysis while minimizing the number of parameters.

In this paper, we modify the structure of ConFormer to fulfill VPR usage and achieve better performance. Firstly, after obtaining the speaker embedding, a fully connected layer is added to reduce the embedding dimensions and obtain the final classification results. Secondly, the Multi-scale Feature Aggregation strategy is introduced for ConFormer basic blocks to improve its ability to extract features at different depths of layers.

As shown in Figure 4, the ConFormer VPR model is composed of five main parts.

4.1.1. Fbank Feature Extraction. As mentioned earlier, the Fbank can preprocess audio signals to extract their nonlinear frequency-domain features. For a typical 2-second audio utterance with a 16 kHz sampling rate, 200 audio frames are obtained after windowing with a frame-length of 25 ms and a frame-shift of 10 ms. Therefore, by setting the dimension of Mel filters M to 80, a 200×80 sized Fbank feature map can be acquired.

4.1.2. Convolution for Downsampling. As shown in part two of Figure 4, the two-dimensional Fbank feature map is essentially a digital image. Thus, a two-dimensional convolutional layer is first introduced to perform spatial downsampling on the audio features to accelerate the inference procedure and achieve higher spatial dimensions. Then, a linear layer is connected to unfold the feature map obtained by the last operation, realizing dimension reduction along the depth channel. A dropout layer is then connected to remove the weights of randomly selected neurons, intended to avoid model overfitting during training.

4.1.3. ConFormer Block. As depicted in the third part of Figure 4, the ConFormer block is a core component of ConFormer VPR, consisting of four modules: a Feed-

Table 2: Confusion matrix of ANN.

	Predict: nonevent	Predict: event
Truth: nonevent	True Negative (TN)	False Positive (FP)
Truth: event	False Negative (FN)	True Positive (TP)

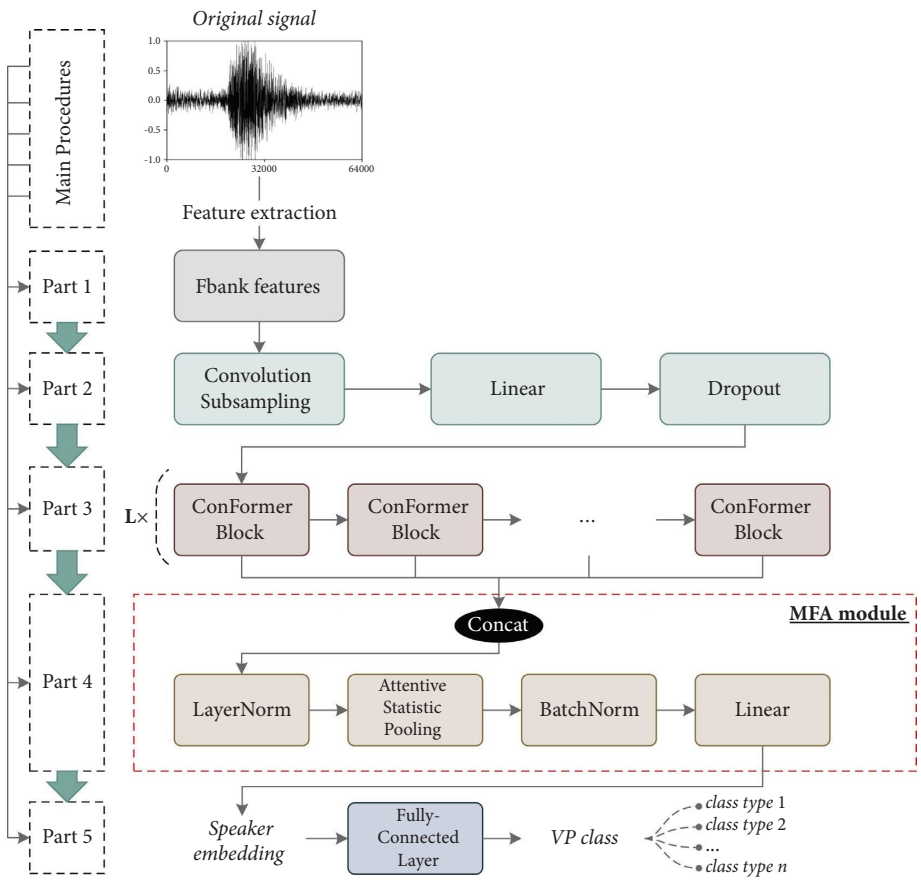


Figure 4: ConFormer VPR model.

Forward Network (FFN) module, a Multihead Self-Attention (MHSA) module, a Convolution (Conv) module, and a second FFN module at the end. These modules form a macaron structure with the same head and tail. Figure 5 shows the computation process of the ConFormer block. Mathematically, for input feature h_{i-1}, the i-th ConFormer block performs layer-by-layer operations according to the following equation:

$$\begin{aligned} \tilde{h}_i &= h_{i-1} + \frac{1}{2}\mathrm{FFN}\left(h_{i-1}\right), \\ h'_i &= \tilde{h}_i + \mathrm{MHSA}\left(\tilde{h}_i\right), \\ h''_i &= h'_i + \mathrm{Conv}\left(h'_i\right), \\ h_i &= \mathrm{LayerNorm}\left(h''_i + \frac{1}{2}\mathrm{FNN}\left(h''_i\right)\right). \end{aligned} \tag{15}$$

4.1.4. MFA Strategy. Concatenating different layers of feature maps can improve the performance of deep learning-based VPR models [54, 55]. Therefore, we introduce the MFA strategy to connect the features extracted by L ConFormer blocks. Specifically, an attentive statistic pooling layer is used to provide different trainable weights for the output feature map of each ConFormer block. After batch normalization and a linear layer, a one-dimensional vector speaker embedding is acquired, which is usually a 1×192 vector in speech recognition tasks.

4.1.5. Fully Connected Layer. As shown in Figure 4, following the speaker embedding vector, a fully connected layer is constructed according to the final outputs. Here, the speaker embedding vector is dimensionally reduced to the number of voiceprint types. The Additive Angular Margin Softmax (AAMSoftmax) [56] is applied to compute model

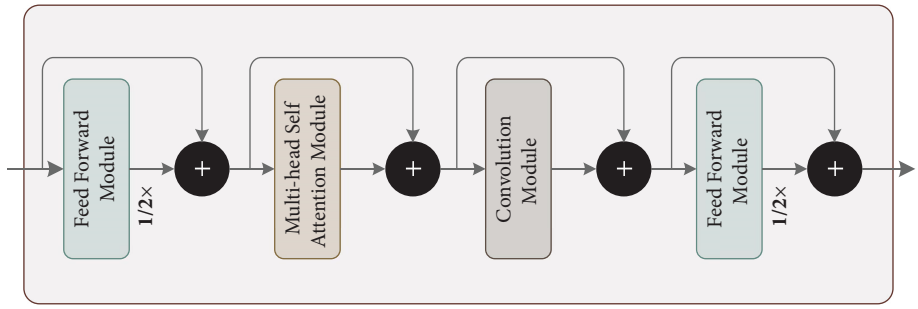


Figure 5: ConFormer block.

loss during training, defined in equation (16) Compared to traditional Softmax, AAMSoftmax can enhance the discriminative power of features and improve the performance of classification tasks. It can also reduce intraclass variance and increase interclass variance, making the learned features more compact and separable.

$$L = -\log \frac{e^{s\cos\left(\theta_{y_i}+m\right)}}{e^{s\cos\left(\theta_{y_i}+m\right)} + \sum_{j=1, j\neq y_i}^{N} e^{s\cos\theta_j}}, \tag{16}$$

where N is the number of voiceprint types, θ_j is the angle between the weight and feature, s is the scale factor used for normalization, and m is the margin penalty value between the weight and feature.

4.2. Cascading Approach for BEJ Fault Detection. Compared with sensing methods such as vision and radar, audio acquisition requires active excitation sources. According to research by Nishikawa [19], experienced inspectors can determine the fault status of BEJs by the sound of vehicle impact, indicating that vehicle impact is a reliable excitation source for audio-based fault detection of BEJs.

Therefore, as shown in Figure 6, a cascading approach is established using three ConFormer models, each for a different VPR task, to ultimately achieve fault detection of BEJs. Firstly, environmental event VPR is performed to select vehicle impact events from all audio events. Secondly, to explicitly recognize the distinction between fault-free and faulty status of BEJs, vehicle impact events are finely classified according to different vehicle types. Thirdly, for the same type of vehicle impact audio, fault identification of BEJ component states is performed.

In the application phase, we introduce a decision-making module that combines all outcomes deduced from the four ConFormer models to enhance the reliability of the judgment. For each outcome from the third-level ConFormer model, both fault-free and faulty detections are logged, leading to an accumulation of hits for both categories. The status of the BEJ is subsequently determined based on the ratio of faulty hits: if it falls below a predetermined threshold, the BEJ is classified as fault-free, whereas if it surpasses the threshold, the BEJ is considered faulty.

4.3. Dataset Creation. Comprehensive datasets are essential for data-driven VPR applications. As shown in Figure 1, videos recorded by an in-field camera can be used to annotate specific vehicle impacts. Besides, sound events such as vehicle horns, human sounds, and bird sounds are manually differentiated and annotated.

The duration of audio signals in Table 1 may be imbalanced for practical applications. To supplement the in situ audio data, Open-Access (OA) datasets can be used. However, environmental noise can affect the robustness of the VPR model [57], particularly in low SNR scenarios such as BEJ environments. Therefore, environmental noise enhancement should be applied to the audio segments of OA datasets based on on-site measured signals, as shown in equation (17) and Figure 7.

$$S_a(x) = S(x) + \alpha N(x), \tag{17}$$

where $S(x)$ and $S_a(x)$ represent the audio signals before and after noise enhancement, respectively, α is the enhancement coefficient, set to 1 in this case, and $N(x)$ represents noise signals, which are randomly obtained from labeled noise signal segments in the BEJ environment.

Typical audio signal segments in the BEJ environment are shown in Figure 8. Both fault-free and faulty status audios are generated by vehicle impact, making it difficult to directly detect faults in BEJs. Besides, prominent segments of bird sounds are relatively short and distinct from other types of audio. However, the remaining types of audio are difficult to distinguish simply by using thresholds or time-to-peak. Therefore, a VPR model is necessary to achieve sound classification through its ability to extract high-dimensional features.

5. Case Study

5.1. Basic Information. An in situ experiment was conducted on the Jiangyin Bridge, a suspension bridge that contains two modular BEJs on each side. The experimental layout is shown in Figure 9. A microphone was fixed on a tripod under the main girder to capture audio data surrounding the BEJ. Above the main girder, a camera was temporarily installed to annotate moment and type of each vehicle passing over the BEJ. The main configurations of the two experimental sensors are listed in Table 3.

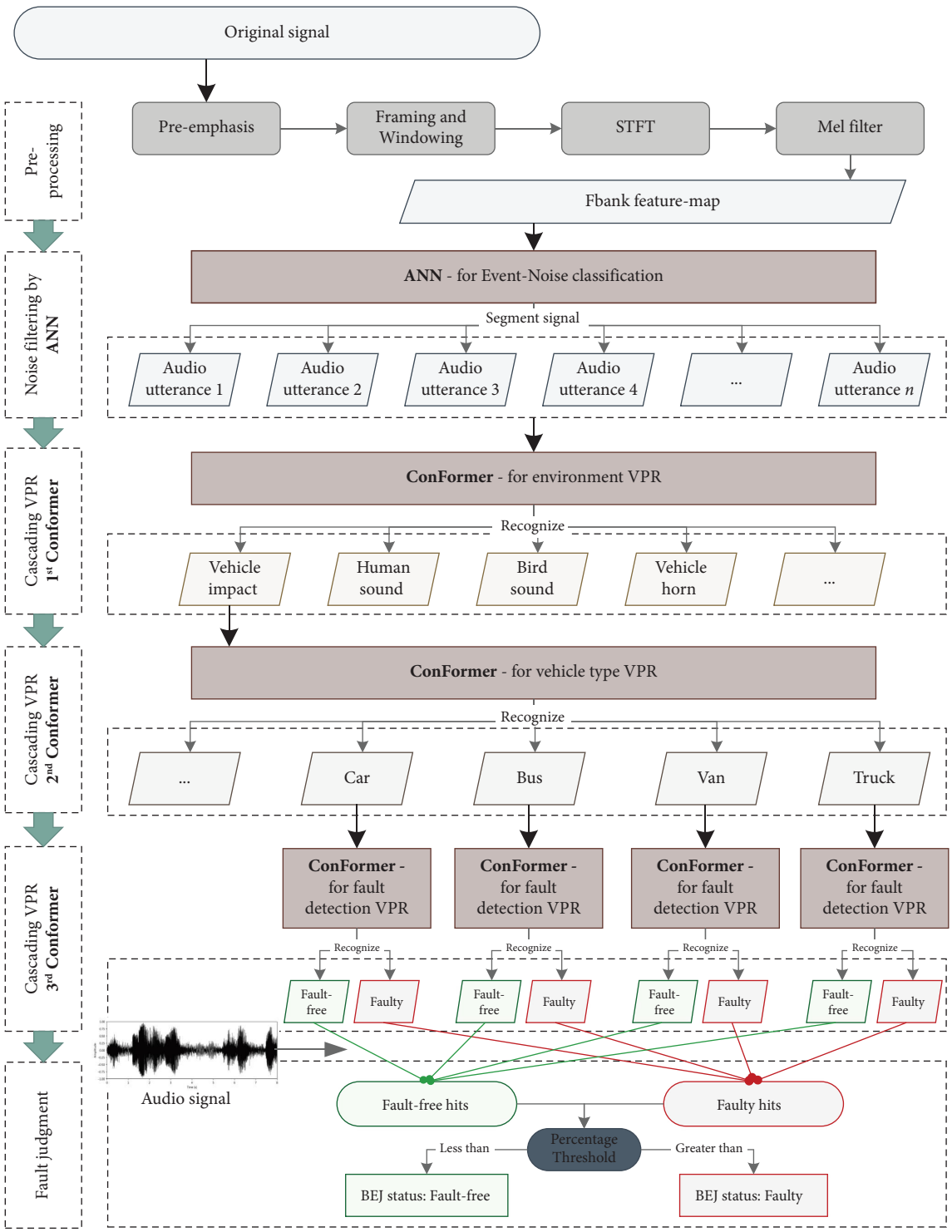


Figure 6: Cascading approach for BEJ fault detection.

For data processing and model testing, we used a desktop PC (CPU: Intel® CoreTM i7-6800k; RAM: 32GB; and GPU: NVIDIA GeForce GTX 1080Ti) with the support of CUDA v10.2 and cuDNN v8.2. The PyTorch deep learning framework was utilized to accomplish model training and evaluation.

5.2. Damage Types of BEJ Faults. During their service, BEJs often experience damage and deterioration within a local range due to the prolonged influence of various factors such as load and environmental conditions (including vehicular traffic, corrosive actions, and temperature), leading to the failure of specific components of BEJs. Common forms of

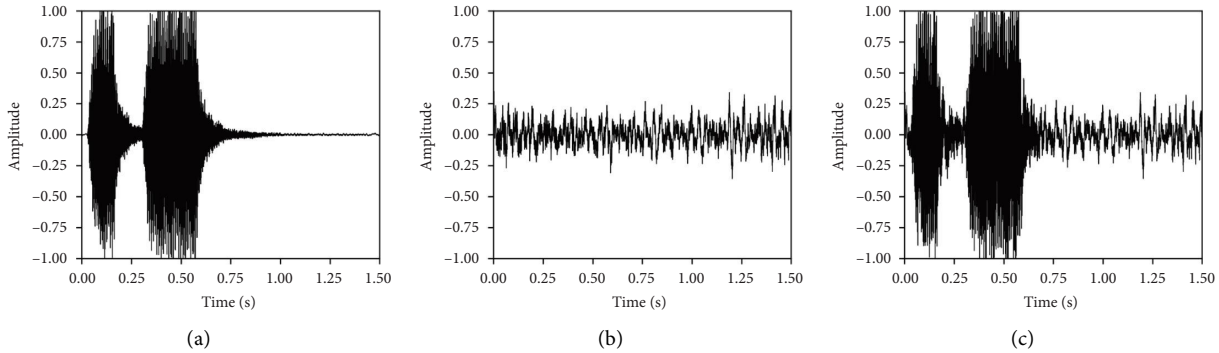


Figure 7: Noise enhancement of OA dataset. (a) Original signal (IUSD dataset). (b) Noise signal (in situ BEJ field). (c) Signal after noise enhancement.

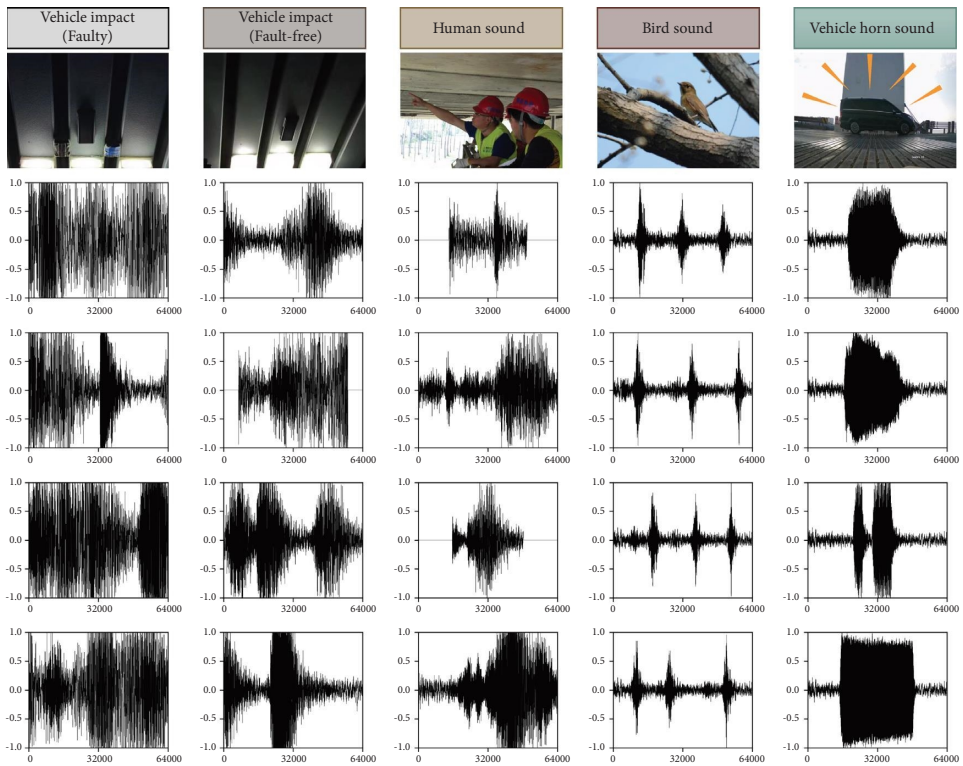


Figure 8: Typical audio signals surrounding BEJs.

damage include the following [5, 58]: (1) local congestion, twisting deformation, or even breakage of the center steel beams; (2) long-term wear of the sliding bearing resulting in sliding failure; (3) aging of the sealing rubber, loss of elasticity in the rubber spring, or even fatigue cracking; (4) and corrosion and detachment of the welding points in some parts of the hanger.

Due to the distinct changes in BEJ components caused by damage, the fault-free state and various damage modes will manifest different acoustic features under the influence of vehicular impacts. Therefore, it is logical to categorize different kinds of damage based on impacting audio signals.

In this paper, to fundamentally verify this idea, we selected a fault-free BEJ of Jiangyin Bridge and collected one hour of audio data. Subsequently, we introduced simulated faults by using two steel shims to congest two center steel beams. One hour of vehicle-induced audio signals under this faulty BEJ was then recorded. The practical scenario is illustrated in Figure 10.

5.3. Event Segmentation of Audio Data. To train the ANN classification model for event segmentation, we annotated audio signals using Praat software [59]. A total of 2010.438s

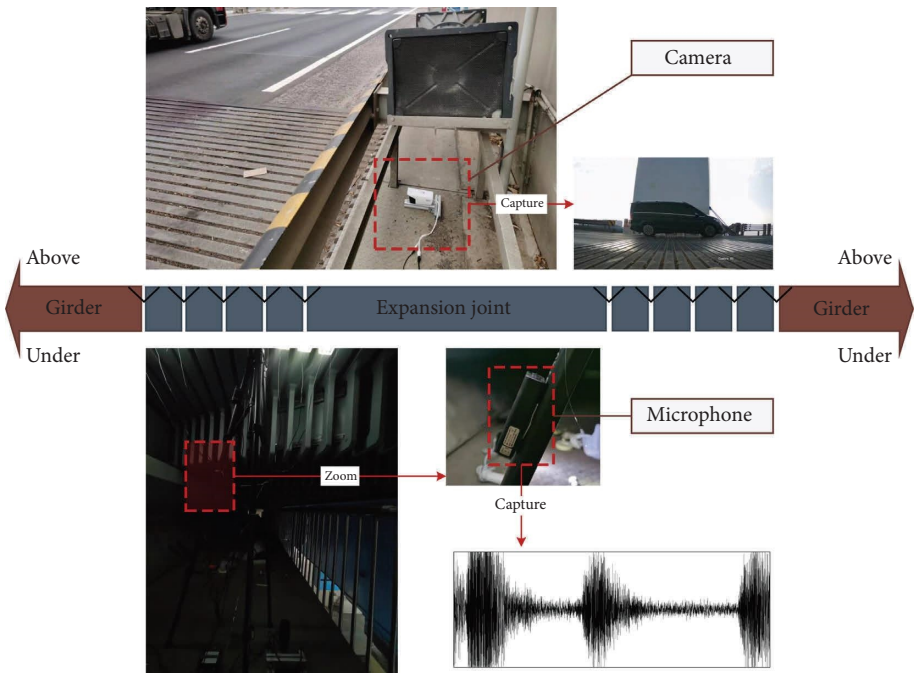


Figure 9: The experimental layout.

Table 3: Main configurations of the devices.

Device	Parameter	Unit	Value
Microphone	Size	mm	78.5 × 39.7×18.4
	Sampling rate	Hz	32,000
	Bit depth	bit	16
	Data format	—	WAV
Camera	Size	mm	87 × 82.8×169.9
	Image size	px	1,920 × 1,080
	Image resolution	px	2,073,600
	Frame per second	—	25
	Data format	—	MP4

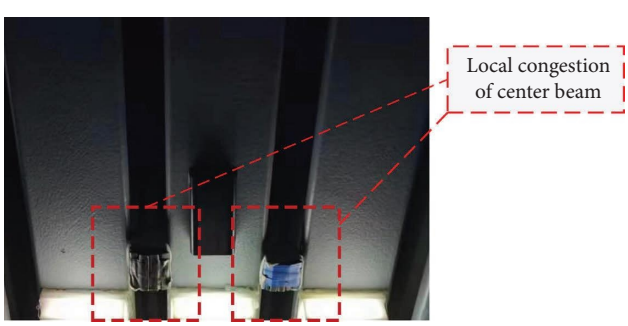


Figure 10: Fault condition setting: local congestion of center steel beams.

of signals were obtained, including 1035.562s of event signals and 974.876s of nonevent signals, as shown in Table 4. The dataset was split into training and test sets at a ratio of 80% and 20%, respectively.

The proposed ANN and other comparison models for audio event segmentation were applied, including the ML model Support Vector Machine (SVM) and threshold-based models STE and STCZR. The performance of different models was evaluated using accuracy as the metric. Hyperparameter optimization of the different models was first conducted to determine their sensitivity and stability in event segmentation. The results are shown in Figure 11.

It can be seen that ML methods have higher accuracy than threshold-based methods. Among them, the ANN and SVM models are insensitive to their hyperparameters and achieve accuracies of 93.5% ~ 93.7% and 93.0% ~ 93.1%, respectively. The STE model is sensitive to the predefined threshold and achieves an accuracy of 79.1% ~ 90.0%, so a precise threshold setting is required. The STCZR model is insensitive to the threshold, but its accuracy is only slightly above 50%, indicating little disparity between event and nonevent audio signals in STCZR under the BEJ environment.

An 8-second audio data sample was selected to apply the event segmentation procedures using the four models mentioned above. As shown in Figure 12, the ANN, SVM, and STE models can accurately segment event signals, with only about 0.05s of misidentification locally at the beginning

Table 4: Self-annotated dataset for audio event segmentation.

Class	Duration of signals (s)	No. of frames/total	No. of frames/for training	No. of frames/for test
Event	1035.562	82845	66276	16569
Nonevent	974.876	77990	62392	15598
Total	2010.438	160835	128668	32167

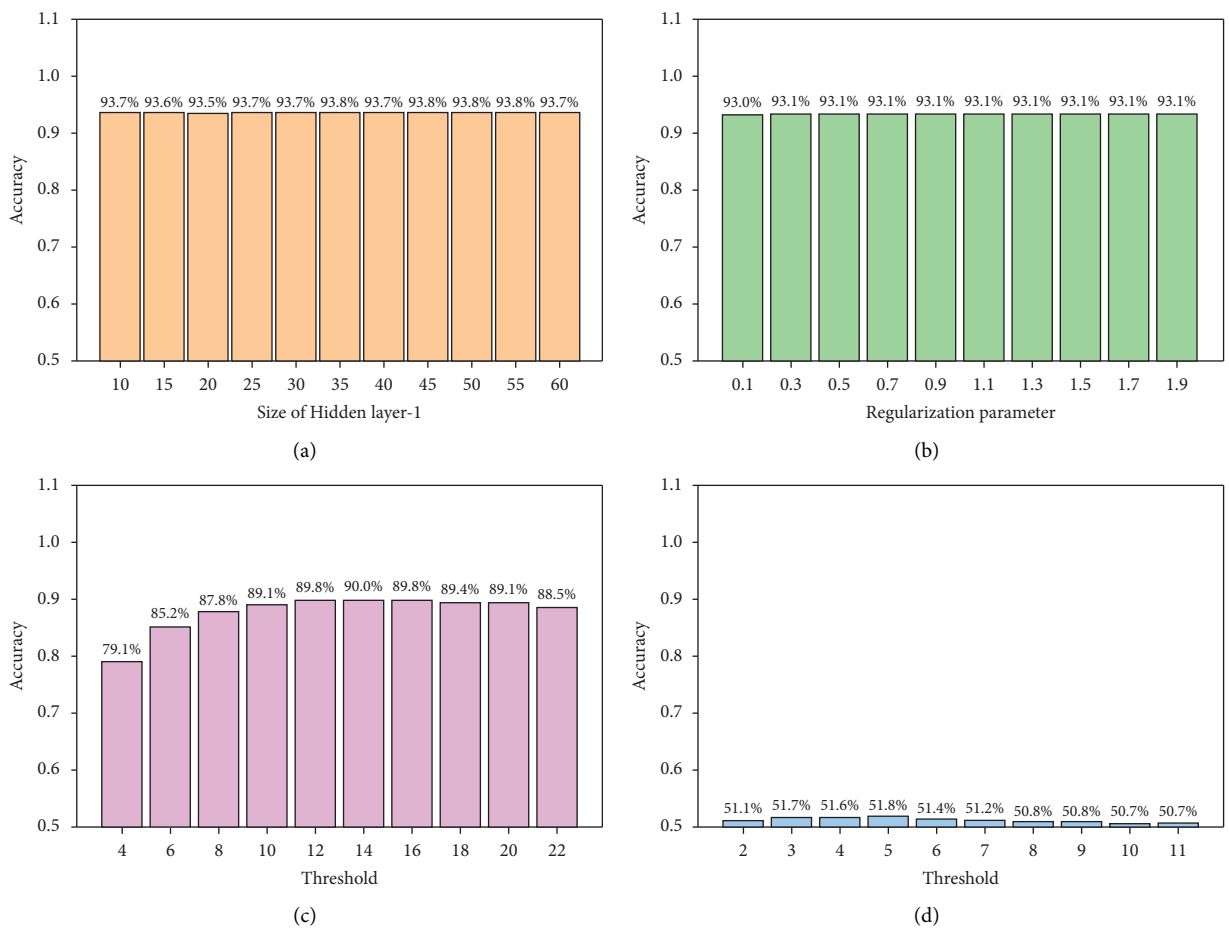


Figure 11: Hyperparameter optimization of each event segmentation model. (a) ANN (hyperparam: size of hidden layer). (b) SVM (hyperparam: regularization parameter). (c) STE (hyperparam: segmentation threshold). (d) STCZ (hyperparam: segmentation threshold).

and end of events. Their main errors involve identifying low-amplitude event signals as nonevent, which has little impact on subsequent VPR applications. Compared to the SVM and STE models, false predictions occur less frequently for the ANN model, indicating its stronger adaptability and reduced need for postprocessing. Consistent with previous conclusions, the STCZR model almost misidentifies all nonevent segments as event segments, so it cannot segment event audio.

Additionally, for the ML methods of the ANN and SVM models, the ROC curves under optimal hyperparameters are plotted as shown in Figure 13. The ROC curves show a good "Γ" shape, indicating that the models have good generalization performance, with AuC values of 0.981 and 0.964, respectively. As listed in Table 5, in terms of model training time and inference processing efficiency, the threshold-based method does not require data training. Since the SVM model is based on matrix operations for tuning, training time can be very long for large sample sizes. The ANN model has clear advantages in both training and inference efficiency, exceeding other algorithms by 20 times in terms of segmentation and recognition efficiency of the original audio. It only takes 10.5 ms to segment and infer 1 s of original audio data, making it suitable for preprocessing BEJ environment audio.

5.4. Fault Detection of BEJs

5.4.1. Dataset Preparation. In this work, in addition to the in-field audio data, we used the DCASE [60], ESC [61], and IUSD [62] OA datasets for dataset expansion. A total of 2103 audio segments were divided into training and test sets at

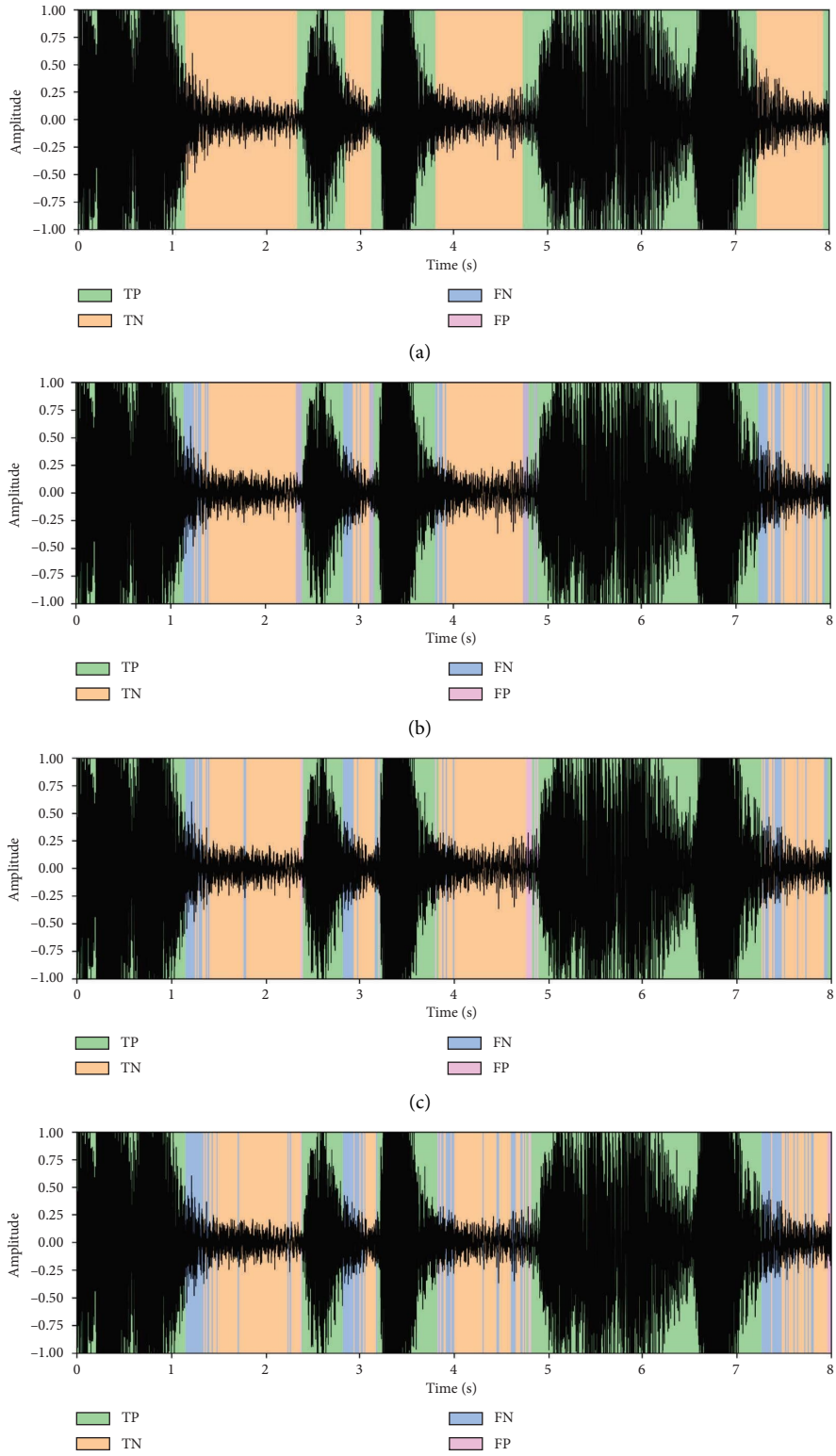
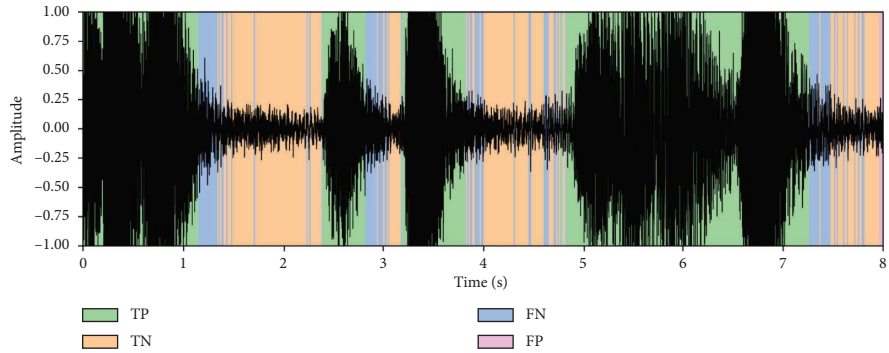
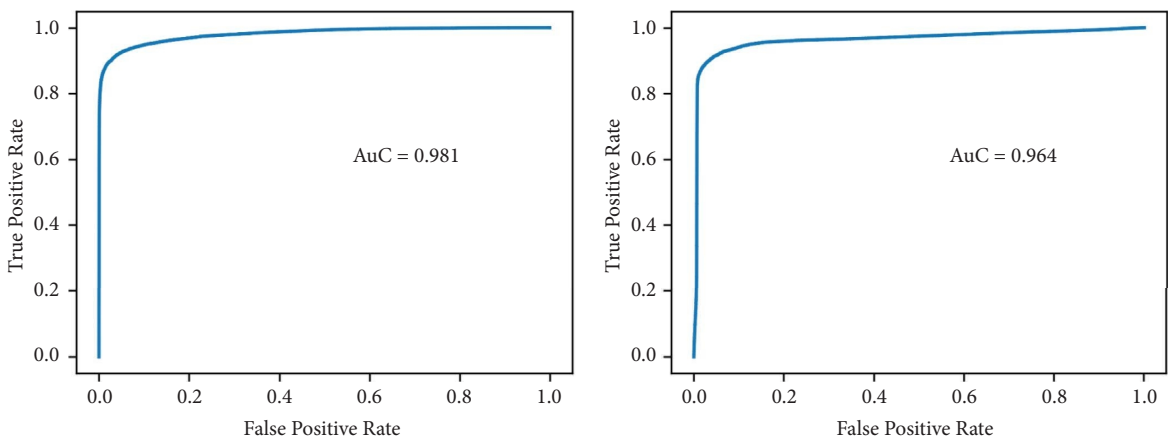


Figure 12: Continued.



(e)

Figure 12: Result comparison of event segmentation models. (a) Ground truth. (b) Segmentation results by ANN. (c) Segmentation results by SVM. (d) Segmentation results by STE. (e) Segmentation results by STCZ.



(a) (b)

Figure 13: ROC curves of ML models for event segmentation. (a) ANN model. (b) SVM model.

Table 5: Performance comparison of event segmentation models.

Item	Unit	ML models		Threshold models	
		ANN	SVM	STE	STCZ
Optimized hyperparameter	—	35	0.5	8	6
Accuracy	%	93.8	93.1	90.0	51.8
AuC	—	0.981	0.964	—	—
Time of training	s	36.7	3613.9	0	0
Time of inference	ms/s	10.5	242.9	208.1	250.5

a ratio of 0.8 : 0.2, resulting in a benchmark dataset for BEJ fault detection, with details listed in Table 6. Specifically, for fine-grained recognition of vehicle impact events, four vehicle classes—Car, Bus, Van, and Truck—were labeled based on definitions in bridge standards [63]. A total of 4123 audio segments were obtained, as shown in Table 7.

5.4.2. Cascading ConFormer Model Training and Evaluation for BEJ Fault Detection. Based on the proposed cascading approach shown in Figure 6, three types of ConFormer models were established: environment VPR, vehicle type VPR, and fault detection VPR. A length restriction of 2 seconds was applied to regularize all audio segments to balance efficiency and precision [64]. Therefore, random cropping or zero-padding operations were applied to segments shorter or longer than 2 seconds, respectively.

The training parameters in this work are listed as follows. For Fbank feature extraction, the number of Mel banks M was set to 80. The initial learning rate was set to 0.01 and reduced to 97% of its original value at the end of each epoch to achieve optimal model performance. Considering memory limitations, the batch size for a training epoch was

Table 6: Benchmark dataset for BEJ fault detection.

Audio type	Training set	Test set	Total
Vehicle impact (faulty/fault-free)	229/527	58/132	287/659
Human sound	73	19	92
Bird sound	459	115	574
Vehicle horn	392	99	491
Total	1680	423	2103

Table 7: Fine-grained vehicle types of impact events.

Vehicle impact audio type	Training set	Test set	Total	Signal statistics (mean) (ms)	Signal statistics (standard deviation) (ms)
Car	2412	603	3015	276.0	56.1
Bus	118	29	147	449.2	100.4
Van	442	111	553	375.6	63.9
Truck	326	82	408	721.3	333.7
Total	3298	825	4123	390.9	260.6

set to 100. The scale factor s of the AAMSoftmax loss function (equation (16)) was set to 30 and the margin penalty m was set to 0.2. The total number of training epochs was set to 100.

First, the ConFormer VPR model with MFA module was trained and evaluated, as shown in Figure 14. As indicated by the loss and accuracy curves in Figure 14(a), the model's performance gradually improved and quickly stabilized at around 30 epochs, where the loss was close to 0 and accuracy was close to 100%. In subsequent rounds, the model's loss and accuracy only oscillated slightly, indicating that the training process was complete. As seen from the ROC curves in Figure 14(b), the model performed well for all four types of audio, achieving a mean average AuC (maAuC) of 0.975. Among them, AuCs for bird sounds and vehicle horn sounds almost reached 1.0. Meanwhile, AuCs for human sounds and vehicle impact sounds were 0.966 and 0.939, respectively. The model's performance can be further improved by adding more working conditions for more data.

The model shown in Figure 14 was obtained after hyperparameter tuning, shown in Figure 15. Three hyperparameters were used for comparison: (1) L—the number of Transformer layers in the ConFormer block; (2) D_{en}—the encoder dimension of the ConFormer block; and (3) D_{em}—the voiceprint embedding dimension after the MFA module.

From Figure 15, it can be seen that L has a significant impact on both accuracy in the training set and maAuC in the test set. When $L > 3$, classification performance decreases significantly and brings a large amount of computation, so the optimal value for L should be 3. Meanwhile, D_{en} and D_{em} have no obvious impact on the model's accuracy and maAuC. Therefore, considering a balance between computational efficiency and classification performance, the two hyperparameters are set to 128 and 64, respectively.

Afterward, a second ConFormer model was trained to classify the types of vehicle impact events, with results plotted in Figure 16. From the curves in Figure 16(a), after 80 epochs, the model's loss and accuracy were close to 0 and 1, respectively. This indicates that the training was complete and the model was stable, achieving high classification accuracy under reasonable thresholds. Additionally, Figure 16(b) shows the ROC curve for each vehicle type, with an overall performance maAuC value of 0.925, slightly lower than the performance of the previous model.

Specifically, the model has an AuC value of nearly 1.0 for Car and Truck, indicating that it can perfectly distinguish between these two types of vehicle impact sounds. However, for Bus, the AuC value is 0.764, indicating medium classification accuracy based on the definition of the ROC curve [51]. According to Table 7 measurements, the number of Bus is much lower than the other three types of vehicle impacts. Therefore, the model's classification accuracy for Bus could be improved by increasing the number of Bus samples.

A ConFormer model is cascaded to complete the fault detection of BEJs. As shown in Figure 6, a specific ConFormer model can be applied for each vehicle impact type. In this work, we use Car and Van as two examples for the fault-free and faulty status classification of BEJs, with results shown in Figures 17 and 18. Both models have great classification performance for fault detection, with AuC values of 0.886 and 0.910 respectively. Compared to light passenger cars, vans provide greater differences for the status judgment of BEJs, indicating that stronger wheel impact serves as a better criterion for fault detection of BEJs.

5.4.3. VPR Model Comparison. To verify the performance advantages of our model, we conducted a comparison study with other mainstream deep learning-based VPR models, including the Transformer model [65], the ECAPA-TDNN model [54], and the ResNet18 and ResNet34 models [66]. All models were trained on the environment VPR dataset shown in Table 6. The performance of the models in terms of both precision and efficiency was compared, with results shown in Table 8 and Figure 19.

All models achieved accuracy close to 100%, indicating that they have high classification ability under appropriate thresholds. However, for maAuC evaluation, our

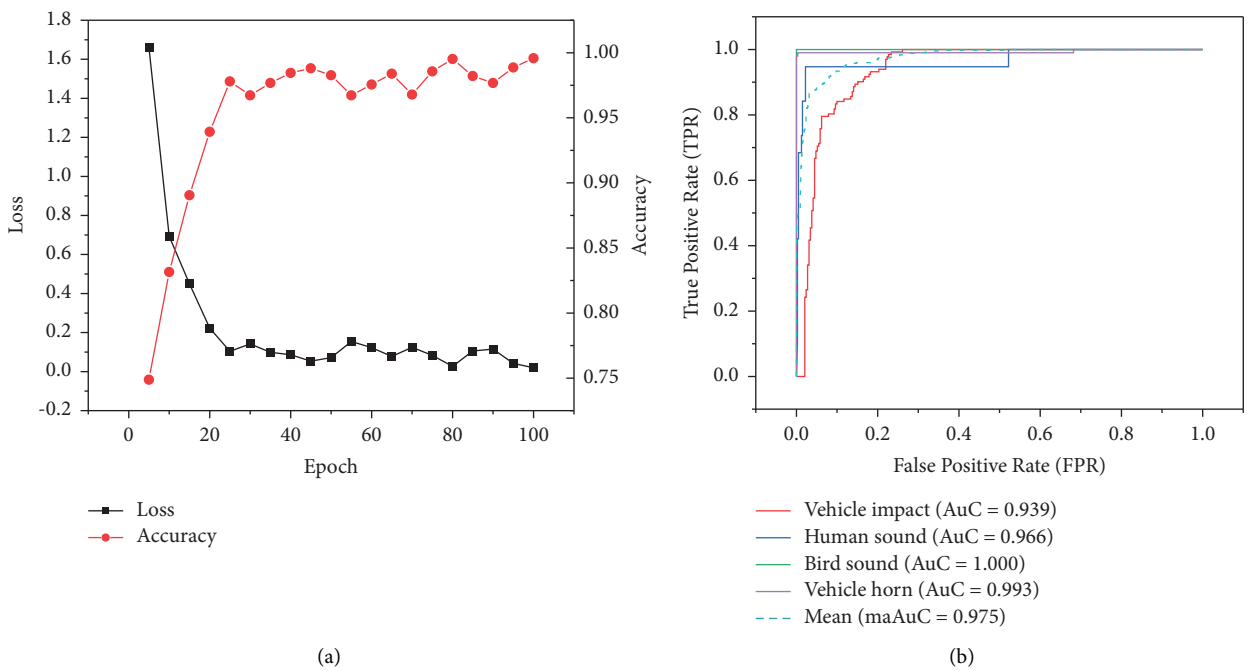


Figure 14: Training and evaluation of ConFormer VPR model for environmental sound classification. (a) Loss curve and accuracy curve. (b) ROC curve and AuC value.

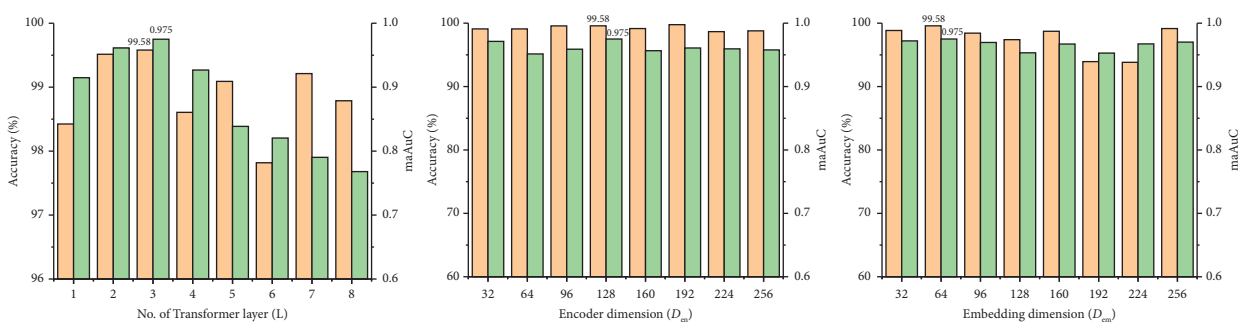


Figure 15: Hyperparameter tuning of ConFormer VPR model.

ConFormer VPR models had a significant performance advantage over other models, indicating that they have the best classification reliability and robustness. The MFA module further improved the maAuC by 0.01. Due to the introduction of self-attention mechanisms, the Transformer and ECAPA-TDNN models achieved the second-best maAuC values of 0.929 and 0.900, respectively. In contrast, the two ResNet models, which only have basic convolution and residual operations, achieved maAuC values of only 0.771 and 0.758. Thus, the self-attention mechanism has been proven to be an important component for improving the performance of BEJ VPR applications.

In terms of inference speed, as shown in Table 8, using a GPU can significantly accelerate the process. The ConFormer VPR model is the slowest, requiring 16.9 ms on GPU hardware to infer 2s of audio data, while ResNet18, which primarily employs convolution operations, only requires 4.5 ms for inference. On the other hand, considering that embedded CPU devices are cheaper and lighter than GPU-based equipment, we also conducted benchmark tests for these models using a CPU. As listed in Table 8, the ConFormer VPR model has medium-level efficiency. The ECAPA-TDNN model, due to parameter pruning optimization, has the highest inference speed with a CPU, requiring only 32.8 ms to classify a 2-second utterance. Therefore, this model can be applied to edge computing scenarios with lower accuracy requirements.

5.4.4. Discussion of Fault Judgment. As outlined in the "Dataset Preparation" section, the audio data collected over a two-hour period, which include one-hour audio signals each for fault-free and faulty BEJs, are utilized for the discussion of fault judgment using the combined approach proposed in Figure 6. The audio signal is segmented into 12 parts for analysis, with each segment lasting for 5 minutes.

The outcomes for each audio segment are illustrated in Figure 20. For an individual audio segment, the classification of vehicle impact is accomplished by the second-

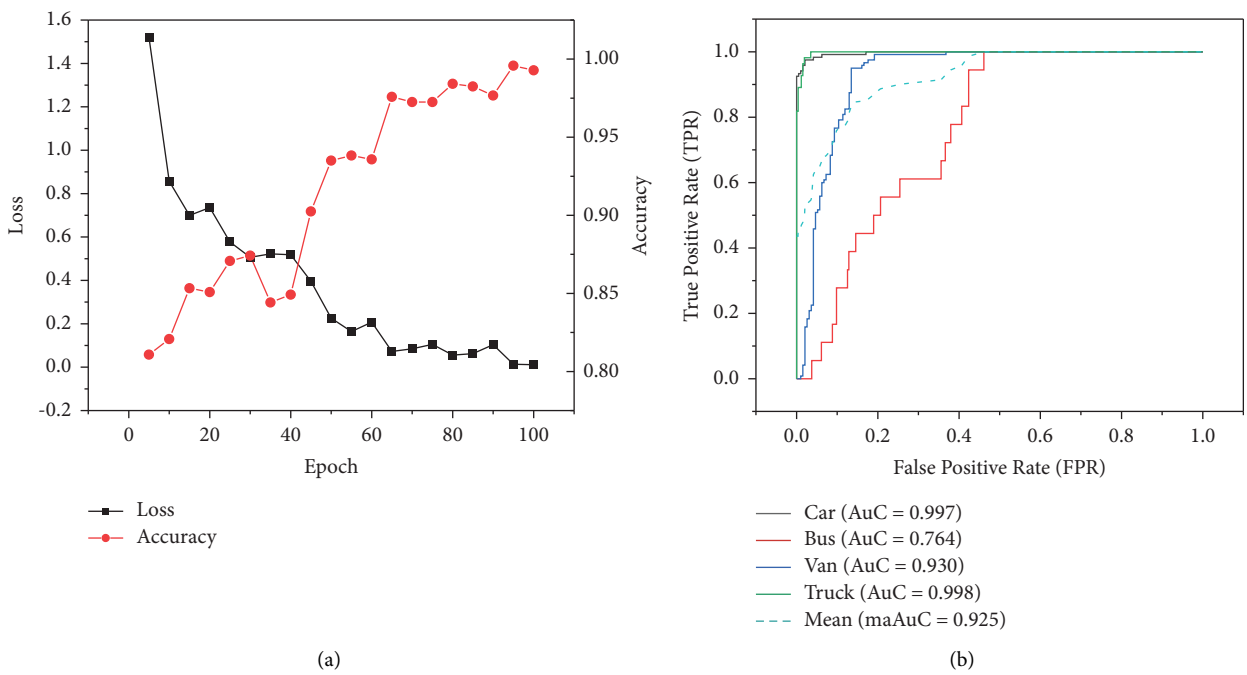


Figure 16: Training and evaluation of ConFormer VPR model for vehicle impact sound classification. (a) Loss curve and accuracy curve. (b) ROC curve and AuC value.

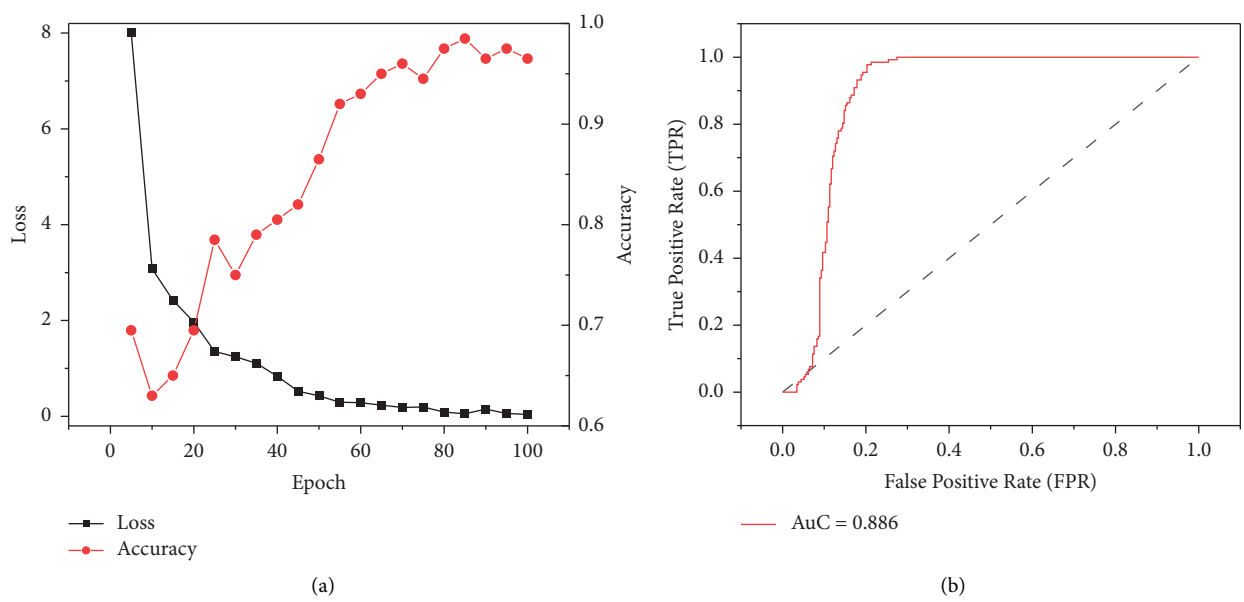


Figure 17: Training and evaluation of ConFormer VPR model for BEJ fault detection (based on car type). (a) Loss curve and accuracy curve. (b) ROC curve and AuC value.

level ConFormer model, with the results displayed in Figure 20(a). Subsequently, for each vehicle audio utterance, a fault is inferred by its corresponding type-specific third-level ConFormer model, resulting in a detection of either fault-free or faulty. Ultimately, the proportion of faulty detections is computed for this audio segment. Figure 20(b) plots the proportion of faulty detections for each audio segment.

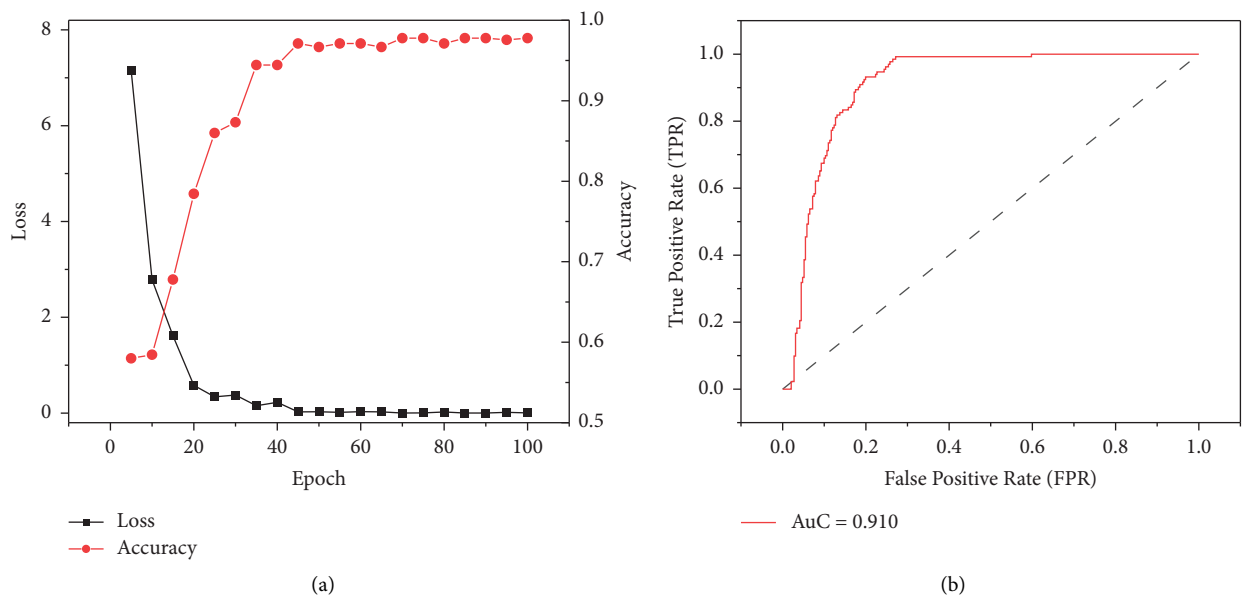


Figure 18: Training and evaluation of ConFormer VPR model for BEJ fault detection (based on van type). (a) Loss curve and accuracy curve. (b) ROC curve and AuC value.

Table 8: Performance comparison of VPR models.

Model	Accuracy (%)	maAuC	Inference time with GPU (ms)	Inference time with CPU (ms)
ConFormer VPR (w/MFA)	99.58	0.975	16.9	127.4
ConFormer VPR (w/o MFA)	98.55	0.965	16.8	115.1
Transformer	98.79	0.929	10.1	70.8
ECAPA-TDNN	99.70	0.900	11.2	32.8
ResNet18	97.15	0.771	4.5	196.9
ResNet34	98.67	0.758	7.2	139.6

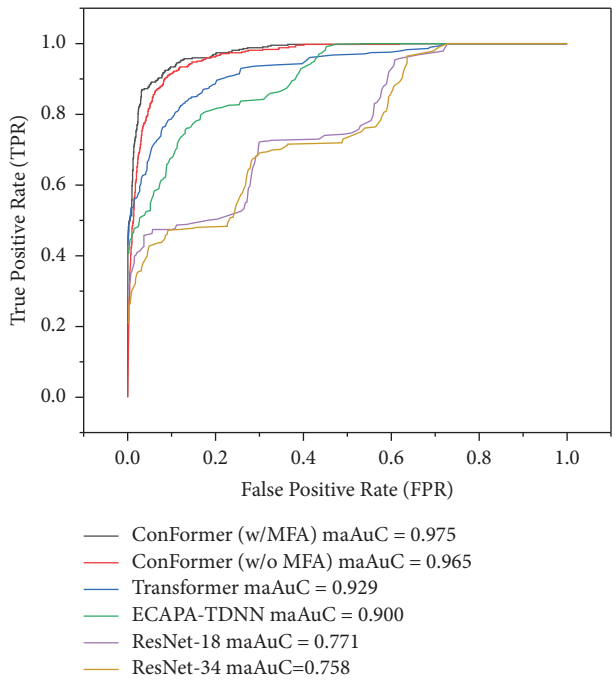


Figure 19: Deep learning-based VPR model comparison.

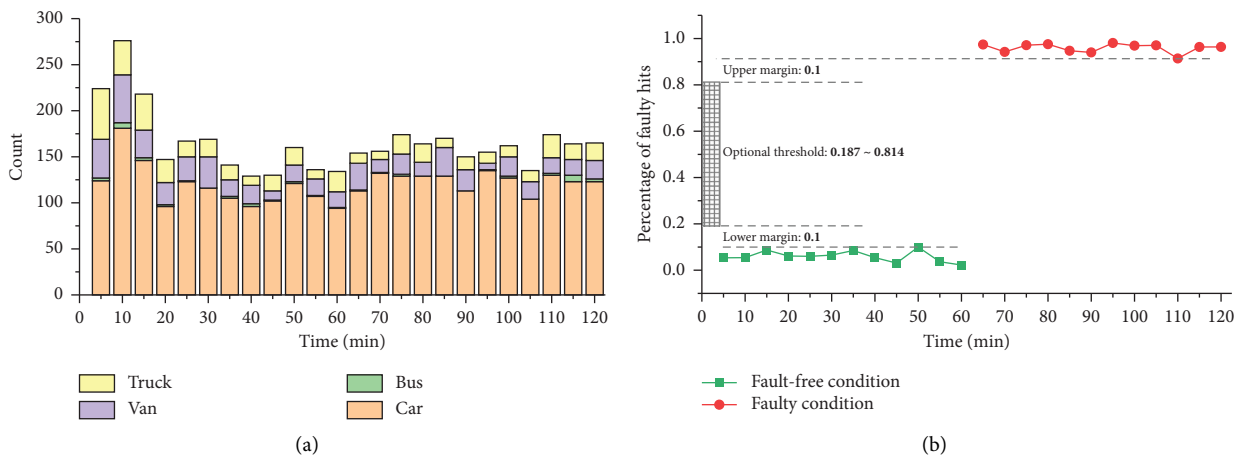


Figure 20: Fault judgment using the proposed combination approach. (a) Detected vehicles. (b) Proportion of faulty hits.

The results in Figure 20(b) demonstrate that the proportions under the fault-free condition are low, while those under the faulty condition are high. This is a reasonable outcome and signifies the practicality of the proposed method's application. Furthermore, both the upper and lower margins are set at 0.1 to establish the separation threshold for fault judgment. In this study, the final optional threshold can be chosen from a range of 0.187 to 0.814.

6. Conclusions

Effective inspection and monitoring of BEJs are crucial for bridge maintenance and management. This paper presents a novel methodology for detecting faults of in-service BEJs through VPR. Microphones are placed beneath the BEJs to capture audio data. An ANN model is then employed to segregate nonevent audio utterances from the original signals. Subsequently, the ConFormer VPR model is enhanced for general audio event classification. Finally, a cascading approach is proposed, involving three successive ConFormer VPR models, aiming to discern vehicle impact audio, identify specific vehicle types, and ultimately detect fault status. Additionally, a case study on an in situ bridge has been conducted to verify the proposed method. Conclusions are drawn as follows:

(1) The proposed methodology is a new endeavor to broaden the spectrum of inspection methods for BEJs. The use of a consumer-grade microphone sensor in this study proves to be a cost-effective alternative when compared to more sophisticated NDT sensors such as accelerometers [8], ultrasonic sensors [11], and electromagnetic sensors [15] used in prior research. From a long-term perspective, the audio-based method offers a more economical solution than manual inspections conducted by inspectors.

(2) An ANN classification model is devised and trained for audio event segmentation, achieving an AuC of 0.981. The ANN model exhibits superior accuracy and stability in event segmentation when compared to threshold-based methods.

(3) The original ConFormer model is adapted for VPR and fortified with an MFA strategy. In response to the intricate sound factors associated with BEJs, we adopt a cascading approach using consecutive ConFormer models for classifying environmental sound types, vehicle impact types, and faults. The trained models achieve AuCs of 0.975, 0.925, and 0.886, respectively, demonstrating the feasibility of detecting faults in BEJs based on audio signals. Notably, the enhanced ConFormer VPR models surpass other VPR models such as Transformer and CNN-based models in terms of performance.

In future research, we aim to further advance this novel field and build upon this foundational work. Our plans include the continuous expansion of the audio signal dataset to encompass multiple types of BEJ faults, as well as faults in other bridge components. Additionally, we intend to conduct benchmarking tests to achieve a more detailed localization and classification of damage based on audio signals.

Data Availability

The data used to support the findings of this study can be obtained from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (grant nos. 52208198, 52238005, 52192663, and 51978514), Interdisciplinary Project in Ocean Research of Tongji University (grant no. 2023-1-YB-03), and National Key Research and Development Program of China (grant no. 2021YFF0501004).

References

[1] P. Paultre, O. Chaallal, and J. Proulx, "Bridge dynamics and dynamic amplification factors—a review of analytical and experimental findings," *Canadian Journal of Civil Engineering*, vol. 19, no. 2, pp. 260–278, 1992.

[2] J. Z. Li, T. B. Peng, and Y. Xu, "Damage investigation of girder bridges under the wenchuan earthquake and corresponding seismic design recommendations," *Earthquake Engineering and Engineering Vibration*, vol. 7, no. 4, pp. 337–344, 2008.

[3] T. Guo, L. Y. Huang, J. Liu, and Y. Zou, "Damage mechanism of control springs in modular expansion joints of long-span bridges," *Journal of Bridge Engineering*, vol. 23, no. 7, Article ID 04018038, 2018.

[4] J. H. Hu, L. H. Wang, X. P. Song, Z. Sun, J. Cui, and G. Huang, "Field monitoring and response characteristics of longitudinal movements of expansion joints in long-span suspension bridges," *Measurement*, vol. 162, Article ID 107933, 2020.

[5] J. Marques Lima and J. de Brito, "Inspection survey of 150 expansion joints in road bridges," *Engineering Structures*, vol. 31, no. 5, pp. 1077–1084, 2009.

[6] Z. Sun, Z. L. Zou, and Y. F. Zhang, "Utilization of structural health monitoring in long-span bridges: case studies," *Structural Control and Health Monitoring*, vol. 24, no. 10, Article ID e1979, 2017.

[7] S. Agnisarman, S. Lopes, K. Chalil Madathil, K. Piratla, and A. Gramopadhye, "A survey of automation-enabled human-in-the-loop systems for infrastructure visual inspection," *Automation in Construction*, vol. 97, pp. 52–76, 2019.

[8] H. Iwabuki, A. Yabe, S. Ono, and S. Tanaka, "A study on fabrications and vibration characteristics of steel finger joints simulating damage stages," in *Bridge Maintenance, Safety, Management, Life-Cycle Sustainability and Innovations*, pp. 2347–2352, CRC Press, Boca Raton, FL, USA, 2021.

[9] J. S. Park, H. M. Ham, and Y. H. Ahn, "Expansion joints risk prediction system based on iot displacement device," *Electronics*, vol. 12, no. 12, p. 2713, 2023.

[10] S. Jang, S. Dahal, and J. Li, "Rapid full-scale expansion joint monitoring using wireless hybrid sensor," *Smart Structures and Systems*, vol. 12, no. 3_4, pp. 415–426, 2013.

[11] T. Hosman, M. Yeary, and J. K. Antonio, "Design and characterization of an mfsk-based transmitter/receiver for ultrasonic communication through metallic structures," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 12, pp. 3767–3774, 2011.

[12] B. L. He, H. P. Deng, M. M. Jiang, K. Wei, and L. Li, "Effect of ultrasonic impact treatment on the ultra high cycle fatigue properties of sma490bw steel welded joints," *The International Journal of Advanced Manufacturing Technology*, vol. 96, no. 5-8, pp. 1571–1577, 2018.

[13] N. K. Mutlib, S. B. Baharom, A. El-Shafie, and M. Z. Nuawi, "Ultrasonic health monitoring in structural engineering: buildings and bridges," *Structural Control and Health Monitoring*, vol. 23, no. 3, pp. 409–422, 2016.

[14] Y. Z. Wang, Y. X. Zhao, L. G. Peng, and D. D. Xu, "Steel corrosion in precast reinforced concrete column-beam joint with grouted sleeve connections under chloride-rich environment," *Journal of Building Engineering*, vol. 77, Article ID 107533, 2023.

[15] T. H. Lin, A. Putranto, P. H. Chen, Y. Z. Teng, and L. Chen, "High-mobility inchworm climbing robot for steel bridge inspection," *Automation in Construction*, vol. 152, Article ID 104905, 2023.

[16] K. A. Ravshanovich, H. Yamaguchi, Y. Matsumoto, N. Tomida, and S. Uno, "Mechanism of noise generation from a modular expansion joint under vehicle passage," *Engineering Structures*, vol. 29, no. 9, pp. 2206–2218, 2007.

[17] J. P. Ghimire, Y. Matsumoto, H. Yamaguchi, and I. Kurahashi, "Numerical investigation of noise generation and radiation from an existing modular expansion joint between prestressed concrete bridges," *Journal of Sound and Vibration*, vol. 328, no. 1-2, pp. 129–147, 2009.

[18] J. Bohatkiewicz, M. Jukowski, M. Halucha, and M. Debinski, "Influence of the acoustic cover of the modular expansion joint on the acoustic climate in the bridge structure surroundings," *Materials*, vol. 13, no. 12, p. 2842, 2020.

[19] Y. Nishikawa, K. Taniguchi, L. H. Ichinose, S. Tsukamoto, and T. Yamagami, "Development of a damage detection system for expansion joints of highway bridges applying acoustic method," in *Proceedings of 6th International Conference on Bridge Maintenance, Safety and Management (IABMAS)*, pp. 165–172, New York, NY, USA, January 2012.

[20] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Pearson College Div, Victoria, Canada, 1st edition, 1993.

[21] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*, Springer, Berlin, Germany, 1st edition, 2015.

[22] S. Abdoli, P. Cardinal, and A. Lameiras Koerich, "End-to-end environmental sound classification using a 1d convolutional neural network," *Expert Systems with Applications*, vol. 136, pp. 252–263, 2019.

[23] F. Demir, M. Turkoglu, M. Aslan, and A. Sengur, "A new pyramidal concatenated cnn approach for environmental sound classification," *Applied Acoustics*, vol. 170, Article ID 107520, 2020.

[24] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[25] A. M. Darling, *Properties and Implementation of the Gammatone Filter: A Tutorial. Speech Hearing and Language*, University College London, London, UK, 1st edition, 1991.

[26] A. Eronen, J. Tuomi, and A. Klapuri, "Audio-based context awareness-acoustic modeling and perceptual evaluation," in *Proceedings of 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, pp. 529–532, Hong Kong, China, April 2003.

[27] A. J. Eronen, V. T. Peltonen, J. T. Tuomi et al., "Audio-based context recognition," *IEEE Transactions on Audio Speech and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

[28] D. B. Zhuo and H. Cao, "Damage identification of bolt connection in steel truss structures by using sound signals," *Structural Health Monitoring*, vol. 21, no. 2, pp. 501–517, 2022.

[29] C. F. Wan and A. Mita, "Recognition of potential danger to buried pipelines based on sounds," *Structural Control and Health Monitoring*, vol. 17, no. 3, pp. 317–337, 2008.

[30] H. Fan, S. Tariq, and T. Zayed, "Acoustic leak detection approaches for water pipelines," *Automation in Construction*, vol. 138, Article ID 104226, 2022.

[31] R. K. Jha and P. D. Swami, "Fault diagnosis and severity analysis of rolling bearings using vibration image texture enhancement and multiclass support vector machines," *Applied Acoustics*, vol. 182, Article ID 108243, 2021.

[32] B. Polok and P. Bilski, "Intelligent diagnostic system for the rachet mechanism faults detection using acoustic analysis," *Measurement*, vol. 183, Article ID 109637, 2021.

[33] B. Chen, S. H. Yu, Y. Yu, and Y. L. Zhou, "Acoustical damage detection of wind turbine blade using the improved incremental support vector data description," *Renewable Energy*, vol. 156, pp. 548–557, 2020.

[34] T. Krause and J. Ostermann, "Damage detection for wind turbine rotor blades using airborne sound," *Structural Control and Health Monitoring*, vol. 27, no. 5, 2020.

[35] W. M. Wang, X. X. Mao, H. G. Liang, D. Yang, J. Zhang, and S. Liu, "Experimental research on in-pipe leaks detection of acoustic signature in gas pipelines based on the artificial neural network," *Measurement*, vol. 183, Article ID 109875, 2021.

[36] Z. Mnasri, S. Rovetta, and F. Masulli, "Anomalous sound event detection: a survey of machine learning based methods and applications," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5537–5586, 2022.

[37] B. Rezaeianjouybari and Y. Shang, "Deep learning for prognostics and health management: state of the art, challenges, and opportunities," *Measurement*, vol. 163, Article ID 107929, 2020.

[38] M. T. Nguyen and J. H. Huang, "Fault detection in water pumps based on sound analysis using a deep learning technique," *Proceedings of the Institution of Mechanical Engineers-Part E: Journal of Process Mechanical Engineering*, vol. 236, no. 2, pp. 298–307, 2022.

[39] S. Dorafshan and H. Azari, "Evaluation of bridge decks with overlays using impact echo, a deep learning approach," *Automation in Construction*, vol. 113, Article ID 103133, 2020.

[40] C. C. Kao, W. R. Wang, M. Sun, and C. Wang, "R-CRNN: region-based convolutional recurrent neural network for audio event detection," *Proceedings of Interspeech*, pp. 1358–1362, 2018.

[41] P. Becker, C. Roth, A. Roennau, and R. Dillmann, "Acoustic anomaly detection in additive manufacturing with long short-term memory neural networks," in *Proceedings of 2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA)*, pp. 921–926, Bangkok, Thailand, April 2020.

[42] A. Oord, S. Dieleman, and H. Zen, "Wavenet: a generative model for raw audio," 2016, https://arxiv.org/abs/1609.03499.

[43] T. Komatsu, T. Hayashiy, R. Kondo, T. Todaz, and K. Takeday, "Scene-dependent anomalous acoustic-event detection based on conditional wavenet and i-vector," in *Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 870–874, Brighton, UK, May 2019.

[44] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 212–224, 2019.

[45] Y. G. Wang, Y. H. Zheng, Y. X. Zhang et al., "Unsupervised anomalous sound detection for machine condition monitoring using classification-based methods," *Applied Sciences*, vol. 11, no. 23, Article ID 11128, 2021.

[46] B. Bayram, T. B. Duman, and G. Ince, "Real time detection of acoustic anomalies in industrial processes using sequential autoencoders," *Expert Systems*, vol. 38, no. 1, 2021.

[47] J. Wu, F. Yang, and W. K. Hu, "Unsupervised anomalous sound detection for industrial monitoring based on arcface classifier and Gaussian mixture model," *Applied Acoustics*, vol. 203, Article ID 109188, 2023.

[48] P. Teng and Y. D. Jia, "Voice activity detection via noise reducing using non-negative sparse coding," *IEEE Signal Processing Letters*, vol. 20, no. 5, pp. 475–478, 2013.

[49] R. Colak and R. Akdeniz, "A novel voice activity detection for multi-channel noise reduction," *IEEE Access*, vol. 9, pp. 91017–91026, 2021.

[50] Y. LeCun, B. Boser, J. S. Denker et al., "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[51] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[52] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, Hoboken, NJ, USA, 2nd edition, 2000.

[53] A. Gulati, J. Qin, and C. C. Chiu, "Conformer: convolution-augmented transformer for speech recognition," in *Proceedings of Interspeech 2020*, pp. 5036–5040, Shanghai, China, October 2020.

[54] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-Tdnn: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification," in *Proceedings of Interspeech 2020*, pp. 3830–3834, Shanghai, China, October 2020.

[55] Y. Zhang, Z. Q. Lv, and H. B. Wu, "MFA-conformer: multi-scale feature aggregation conformer for automatic speaker verification," 2022, https://arxiv.org/abs/2203.15249.

[56] J. K. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "Arcface: additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 5962–5979, 2022.

[57] L. Nanni, G. Maguolo, and M. Paci, "Data augmentation approaches for improving animal audio classification," *Ecological Informatics*, vol. 57, Article ID 101084, 2020.

[58] S. Lee and N. Kalos, "Bridge inspection practices using non-destructive testing methods," *Journal of Civil Engineering and Management*, vol. 21, no. 5, pp. 654–665, 2015.

[59] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2023, https://www.fon.hum.uva.nl/praat/.

[60] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[61] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, Association for Computing Machinery, Brisbane, Australia, October 2015.

[62] J. R. Gloaguen, A. Can, M. Lagrange, and J. F. Petiot, "Creation of a corpus of realistic urban sound scenes with controlled acoustic properties," *Journal of the Acoustical Society of America*, vol. 141, no. 5_Supplement, p. 4044, 2017.

[63] Ministry of Transport of the People's Republic of China, *Technical Standards of Highway Engineering (JTG B01-2014)*, Chinese Ministry of Communications, Beijing, China, 2014.

[64] C. W. Sun, Y. X. Yang, C. Wen, K. Xie, and F. Q. Wen, "Voiceprint identification for limited dataset using the deep migration hybrid model based on transfer learning," *Sensors*, vol. 18, no. 7, p. 2399, 2018.

[65] Y. Q. Wang, A. Mohamed, and D. Le, "Transformer-based acoustic modeling for hybrid speech recognition," in *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6874–6878, Barcelona, Spain, May 2020.

[66] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.