

Article

Self-Supervised Bidirectional State Space Modeling for Voiceprint Feature Representation and Recognition

Junju Lai ¹, Wei Wang ¹, Guangyao Li ¹, Zhichong Kong ¹, Chao Yuan ^{2,*}  and Qian Zhou ²

¹ Qingyuan Power Supply Bureau, Guangdong Power Grid Limited Liability Company, Qingyuan 511500, China; peixunxl@outlook.com (J.L.); pxxiaolou@outlook.com (W.W.); xiangmyong@outlook.com (G.L.); tougyong@outlook.com (Z.K.)

² College of Electrical and Information Engineering, Hunan University, Changsha 410082, China; zhouq@hnu.edu.cn

* Correspondence: yc2300@hnu.edu.cn

Abstract

As substation equipment continues to evolve toward higher voltage levels, larger capacities, and more complex operating conditions, voiceprint signals exhibit greater sensitivity and observability during the early stages of faults. However, traditional modeling approaches still suffer from limitations in capturing long-range temporal dependencies, suppressing noise interference, and adapting to unlabeled data. To address these issues, a state space model-based Mamba self-supervised voiceprint framework, termed MSANet, is proposed. A bidirectional state space scanning mechanism is introduced into the network architecture to avoid the high computational complexity of attention mechanisms while simultaneously preserving both global contextual correlations and local detail representations of voiceprint signals. In addition, a spectrum block masking-based self-supervised learning strategy is incorporated, enabling the model to extract stable time–frequency structural features even under unlabeled or limited labeled samples. Experimental results demonstrate that MSANet achieves high accuracy in voiceprint-related tasks. Furthermore, the lightweight version of the model maintains competitive performance while significantly reducing computational and storage overhead, indicating its feasibility for deployment on edge devices in resource-constrained scenarios such as substation environments. The proposed method provides a potential methodological basis for enhancing fault-related voiceprint feature extraction, representation learning, and future practical engineering deployment.

Keywords: voiceprint recognition; state space model; self-supervised learning; edge computing



Academic Editor: Silvia Liberata Ullo

Received: 27 May 2026

Revised: 12 June 2026

Accepted: 23 June 2026

Published: 29 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Substation equipment serves as critical infrastructure for power transmission and distribution and is widely deployed throughout modern power systems. However, prolonged operation under harsh conditions characterized by high voltage, large current, and strong electromagnetic interference can easily lead to abnormal equipment faults [1]. Failures in substation equipment may cause significant safety risks and economic losses to the power system. Existing inspection approaches mainly rely on infrared temperature measurement, partial discharge monitoring, and manual inspection, which struggle to satisfy the requirements of online monitoring, early fault warning, and reliable fault identification [2]. In contrast, the acoustic signals generated during equipment operation contain multi-source information associated with mechanical vibration, magnetostriction, and dielectric discharge. The corresponding voiceprint information is highly sensitive to weak early-stage

faults and possesses the advantages of non-contact sensing, low cost, and convenient deployment [3]. Therefore, voiceprint-based fault detection for substation equipment is of great significance for improving the safety and reliability of power systems [4].

For accurate recognition and fault monitoring of substation equipment voiceprint signals, it is crucial to develop an efficient voiceprint feature modeling method with low computational complexity, high accuracy, and strong generalization capability [5]. Among the major challenges in the modeling process are the simultaneous capture of both short-range and long-range temporal dependencies, as well as the effective extraction of deep semantic information while suppressing noise interference [6]. Due to the multi-frequency components and nonlinear characteristics of substation equipment voiceprint signals, accurate measurement and feature extraction are key factors in achieving reliable voiceprint monitoring [7].

At present, numerous methods have been proposed for fault detection in substation equipment; however, many of these approaches still exhibit certain limitations. For example, Peng et al. proposed a transformer voiceprint recognition method based on LBT-ODF feature space and a multi-channel network, which extracts local time-domain features from acoustic signals and uses multiple 1D-CNN modules for mechanical fault identification [8]. Li et al. proposed a converter transformer voiceprint fault diagnosis method that combines multi-strategy improved MFCC features with a temporal convolutional network [9]. Nevertheless, these deep learning-based methods still depend on complex feature extraction or relatively heavy network structures, making them less suitable for deployment on edge devices in resource-constrained environments such as on-site power grid applications. Wan et al. proposed a power-transformer fault voiceprint diagnosis method that combines Mel time–frequency feature extraction, Mixup data enhancement, and CNN-based fault classification [10]. Chen et al. proposed a neural network-based transformer voiceprint fault pattern recognition and prediction method, but such data-driven methods still depend heavily on labeled fault samples [11]. Yu et al. developed an automatic condition monitoring and fault diagnosis system for power transformers based on voiceprint recognition, which uses acoustic features to identify different operating and fault conditions [12]. To address noise interference, Zhang et al. proposed a power transformer fault diagnosis method based on voiceprint recognition considering multi-type noises [13].

It should be noted that Mamba-based sequence modeling, bidirectional scanning, and masked spectrogram learning have been explored in existing audio self-supervised learning and sequence modeling studies. For example, SSAST and Audio-MAE mainly adopt Transformer-based architectures for masked spectrogram representation learning, while Wav2Vec 2.0 and HuBERT focus on self-supervised speech representation learning from raw waveforms or latent speech units. In addition, recent Mamba-based audio models have demonstrated the potential of state space models for efficient audio modeling. Different from these studies, the technical contribution of MSANet lies in integrating bidirectional selective state space modeling with masked spectrogram reconstruction and contrastive discrimination for fault-related equipment voiceprint feature extraction.

To address the aforementioned issues, this paper proposes a Mamba-based self-supervised voiceprint representation learning network, termed MSANet. Unlike Transformer-based masked audio modeling methods, MSANet adopts a bidirectional selective state space encoder as the main sequence modeling backbone, which reduces the dependence on self-attention while maintaining long-range temporal modeling capability. In addition, masked spectrogram reconstruction and contrastive discrimination are jointly introduced to improve the representation ability of equipment voiceprint features under limited labeled samples. First, a voiceprint recognition framework based on the State Space Model (SSM) is introduced, which performs sequence modeling with linear computational complexity. The proposed

framework significantly reduces network complexity and computational overhead while preserving long-range temporal modeling capability and global dependency representation. Second, by incorporating a self-supervised pretraining mechanism, the model can be effectively applied under conditions with limited or unlabeled samples. The proposed approach learns generalized feature representations from large-scale unlabeled voiceprint data, thereby alleviating the challenges of sample scarcity and annotation difficulty. A masked spectrogram segment pretraining strategy is further adopted to reduce the dependence on labeled data. Moreover, a bidirectional selective state scanning structure is introduced to achieve a unified representation of global contextual modeling and local detail preservation for voiceprint signals, thereby further enhancing the recognition capability for abnormal acoustic patterns. Finally, experiments conducted on a single-label audio event classification dataset demonstrate that the proposed method substantially reduces computational and storage requirements while maintaining high detection accuracy. In addition, the inference speed is significantly improved compared with conventional Transformer-based models. Owing to its lightweight architecture, the proposed model exhibits strong feasibility and practical engineering value for deployment on edge devices in resource-constrained on-site power grid environments.

2. Self-Supervised Mamba-Based Voiceprint Modeling Method

This section first introduces the mathematical foundation of the Mamba model, with a primary focus on the State Space Model (SSM) framework and its efficiency in capturing long-range temporal dependencies. Rather than directly applying existing audio Mamba models, MSANet adapts bidirectional selective state space modeling to spectrogram patch sequences and combines it with self-supervised masked learning for equipment voiceprint representation. Subsequently, a self-supervised audio Mamba architecture is presented, which utilizes a bidirectional SSM scanning mechanism to achieve robust contextual modeling of audio features. Finally, the adopted self-supervised learning framework is described, where a masked spectrogram segment strategy is employed to reduce the dependence on labeled training samples.

2.1. Mamba State Space Model

State space models originate from Kalman filters and are inspired by continuous systems. As a fundamental framework for sequence modeling, an SSM maps the input sequence to the output through the intermediate hidden state $h(t) \in \mathbb{R}^N$. This process is formulated using the key evolution parameter $A \in \mathbb{R}^{N \times N}$, the projection parameter $B \in \mathbb{R}^{N \times 1}$, and $C \in \mathbb{R}^{1 \times N}$. For clarity, the notation used in the state space formulation is defined as follows. Let $x(t) \in \mathbb{R}^{d_{in}}$ denote the input feature at time t , state, and $y(t) \in \mathbb{R}^{d_{out}}$ denote the output feature. These definitions are used consistently in the following continuous and discrete state-space equations. The state space equations in continuous-time systems are expressed as follows:

$$h'(t) = Ah(t) + Bx(t) \quad (1)$$

$$y(t) = Ch(t) \quad (2)$$

In practical applications, the input sequences fed into the model are typically discrete. Therefore, the continuous system needs to be discretized. In this work, the zero-order hold (ZOH) method is adopted, where the time-scale parameter Δ is used to discretize A and B .

$$\bar{A} = \exp(\Delta A) \quad (3)$$

$$\bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B \quad (4)$$

Here, $\bar{A} \in \mathbb{R}^{N \times N}$ and $\bar{B} \in \mathbb{R}^{N \times d_{in}}$ denote the discretized forms of A and B , respectively. The parameter Δ denotes the discretization step size. After discretization, the discrete input sequence is represented as $x_k \in \mathbb{R}^{d_{in}}$, the hidden state as $h_k \in \mathbb{R}^N$, and the output as $y_k \in \mathbb{R}^{d_{out}}$, where k is the discrete time index. The state space equations for the discrete signal obtained with step size Δ are given as follows:

$$h_d = \bar{A}h_{d-1} + \bar{B}x_d \quad (5)$$

$$y_d = Ch_d \quad (6)$$

In the above state equations, the three discrete parameters $\bar{A}$, $\bar{B}$, and C are constant after discretization. Therefore, when computing y , the output can be efficiently obtained in a convolutional form using the input sequence x and the structured convolution kernel $\bar{K}$. The convolution kernel is computed from the matrices $\bar{A}$, $\bar{B}$, and C , and is defined as follows:

$$\bar{K} = (C\bar{B}, C\bar{A}\bar{B}, \dots, C\bar{A}^{D-1}\bar{B}) \quad (7)$$

$$y = \bar{K}x \quad (8)$$

Here, D denotes the sequence length of the input x , and $\bar{K} \in \mathbb{R}^D$ represents the adopted structured convolution kernel. In the discretized Mamba framework, the parameters $\bar{B}$, C , and Δ are functions of the input sequence x . These parameters are dynamically updated at each step d , and different parameter values are computed during inference for different inputs. As a result, the model possesses content-aware and selective modeling capabilities, enabling the dynamic selection of representative features from the input sequence. Meanwhile, in the Mamba model, a selective parallel scan algorithm is employed to implement the convolution computation mechanism. Based on dynamically varying parameters, the model maintains both the efficiency and accuracy of sequence modeling under linear computational complexity.

2.2. MSANet Model Architecture

Figure 1 illustrates the overall architecture of the proposed MSANet framework. The network mainly consists of four components: spectrogram sequence input, flattening and linear projection mapping, positional encoding, and a bidirectional state space model encoder.

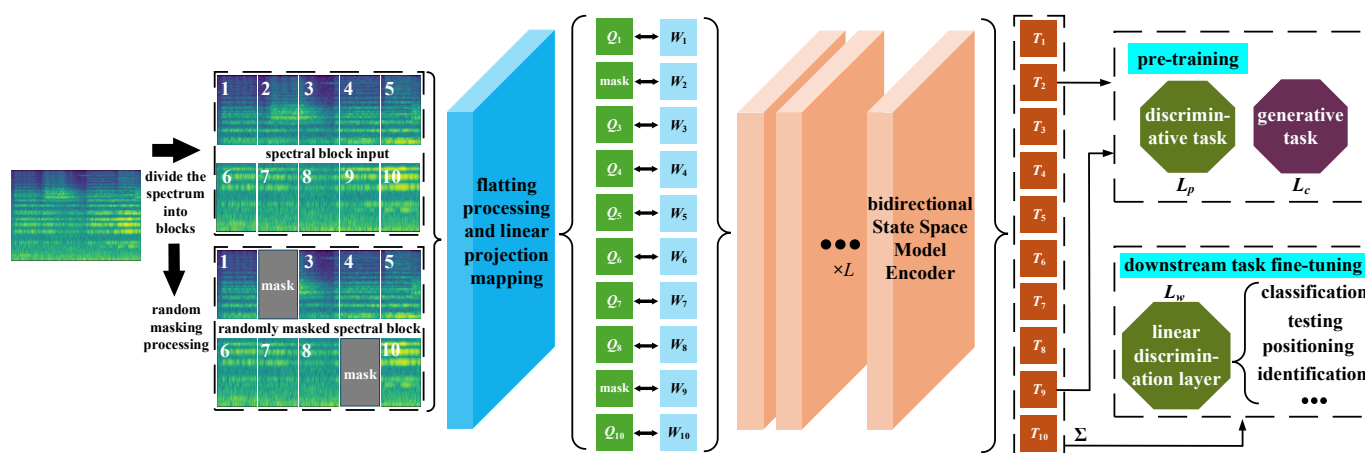


Figure 1. Overall Architecture of the Self-Supervised MSANet Framework.

2.2.1. Spectrogram Sequence Input

Before being fed into the model, the audio waveform is first transformed into a spectrogram representation, which characterizes the signal in both the time and frequency domains. This transformation is achieved through the Short-Time Fourier Transform (STFT), and the corresponding formulation is given as follows:

$$X(\tau, f) = \sum_{n=-\infty}^{+\infty} x[n]w[n - \tau]e^{-j2\pi fn} \quad (9)$$

In the above equation, $x[n]$ denotes the discrete-time speech signal, while $w[n - \tau]$ represents the analysis window function used to capture local temporal segments. In this study, a 25 ms Hamming window with a frame shift of 10 ms is adopted for computation. Here, τ denotes the time-frame index, and f represents the frequency variable. The term $e^{-j2\pi fn}$ corresponds to the Fourier kernel, and $X(\tau, f)$ denotes the complex-valued spectrum obtained through the Short-Time Fourier Transform. Subsequently, 128 logarithmic Mel filter banks are applied to the generated magnitude spectrum for feature extraction, thereby constructing a 128-dimensional spectrogram, denoted as S . The logarithmic Mel filter bank feature is formulated as follows:

$$S(\tau, m) = \log \left(\sum_{k=0}^{K-1} H_m(f_k) |X(\tau, f_k)| + \varepsilon \right) \quad (10)$$

In the above equation, $H_m(f_k)$ denotes the weighting function of the m -th Mel filter at frequency f_k . Here, $m = 1, 2, \dots, M$, where M represents the number of filters and is set to 128 in this work. K denotes the number of frequency sampling points, and ε is a small constant introduced to prevent numerical instability during logarithmic computation. The generated spectrogram has a dimension of $128 \times 100t$, where 128 represents the number of frequency resolution units, denoted as F , and $100t$ corresponds to the number of temporal frames for an audio signal with duration t seconds, denoted as T . Finally, the spectrogram is partitioned into 16×16 patches with a stride of 10 in the downstream experiments. Taking a 10 s audio signal as an example, approximately 500 spectrogram patches can be obtained, which provide inputs for subsequent feature embedding and encoding processes.

2.2.2. Flattening and Linear Projection Mapping

After obtaining the spectrogram, the local temporal and frequency-domain information must be transformed into feature vectors that can be processed by the model. Specifically, each spectrogram patch S_i is first flattened, converting the two-dimensional matrix into a one-dimensional feature sequence. Subsequently, a linear transformation layer is employed to project the sequence into a high-dimensional feature space, thereby obtaining the corresponding feature representation Q_i with dimensionality D . The purpose of this module is to enhance feature representation capability and provide a foundation for the subsequent model to capture global correlations and contextual information.

2.2.3. Positional Encoding

To preserve the temporal and spatial relationships among spectrogram patches during subsequent modeling, a learnable positional encoding W_i is introduced for each feature embedding vector Q_i . The dimensionality of the positional encoding is maintained consistent with that of the embedded features to facilitate subsequent training. Through parameterized learning, the positional encoding enables the model to perceive the relative ordering and spatial distribution of spectrogram patches within the overall spectrogram, thereby preventing the loss of positional information during the flattening and projection processes.

2.2.4. Bidirectional State Space Model Encoder

The core component of the model is an L -layer Mamba encoder, which consists of a bidirectional state space structure, as illustrated in Figure 2. The encoder takes the feature vector with positional encoding, $Q_i + W_i$, as its input and performs feature modeling along both the forward and backward directions of the temporal sequence, thereby capturing bidirectional dependencies in speech signals. Assume that the spectrogram is divided into N_p patches. The embedded patch sequence is denoted as $Q \in \mathbb{R}^{N_p \times D}$, where D is the embedding dimension. The learnable positional encoding is denoted as $W \in \mathbb{R}^{N_p \times D}$, and the input of the l -th encoder layer is denoted as $H^{(l)} \in \mathbb{R}^{N_p \times D}$. Therefore, the element-wise addition between the patch embedding and the positional encoding is dimensionally consistent. Compared with conventional SSMs that only support unidirectional modeling, the bidirectional structure can process information from both temporal directions while preserving sequence consistency. Therefore, it more fully exploits contextual information and achieves a comprehensive representation of the overall audio features.

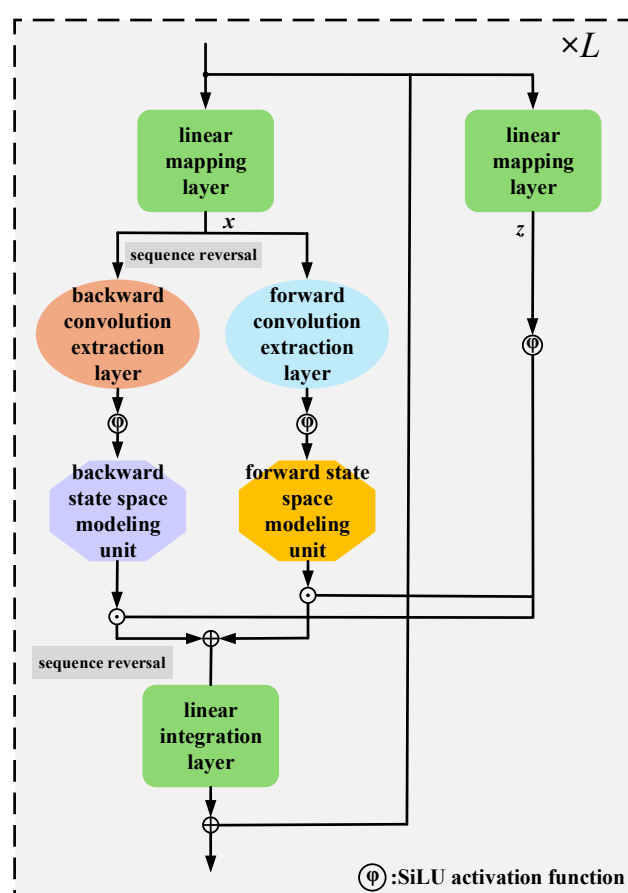


Figure 2. Architecture of the Bidirectional State Space Model Encoder.

The input features are first combined with the positional encoding through an element-wise addition operation, which is defined as follows:

$$Q'_i = \begin{cases} Q_i + W_i, & \text{when used as the first-level input} \\ T_i, & \text{when it comes to the subsequent layer} \end{cases} \quad (11)$$

Here, Q_i denotes the embedding vector of the i -th spectrogram patch, W_i represents the learnable positional encoding, and T_i denotes the output of the previous layer. Through this operation, temporal positional information is introduced into the feature input, thereby

providing guidance for subsequent state space modeling. The aggregated features are then linearly projected and transformed into a representation space suitable for state space modeling, as expressed below:

$$\mathbf{x} = E_x \mathbf{Q}'_i + \mathbf{b}_x, \mathbf{z} = E_z \mathbf{Q}'_i + \mathbf{b}_z \quad (12)$$

In the above equation, E_x and E_z denote learnable linear transformation matrices, while $\mathbf{b}_x$ and $\mathbf{b}_z$ represent the corresponding bias terms. Subsequently, state space modeling is performed separately along both the forward and backward directions to simultaneously capture dynamic features from different temporal directions. The computation process is formulated as follows:

$$\mathbf{x}'_o = \text{SiLU}(\text{Conv1D}_o(\mathbf{x})) \quad (13)$$

$$\mathbf{B}_o = E_{B_o} \mathbf{x}'_o, \mathbf{C}_o = E_{C_o} \mathbf{x}'_o \quad (14)$$

$$\Delta_o = \log(1 + \exp(E_{\Delta_o} \mathbf{x}'_o)) \quad (15)$$

Here, o denotes the forward and backward directions, respectively. Conv1D represents a one-dimensional convolution operation used to capture local temporal contextual information, while SiLU($\cdot$) denotes the activation function introduced to incorporate nonlinear features. E_{B_o} and E_{C_o} are learnable linear projection matrices, whereas $\mathbf{B}_o$ and $\mathbf{C}_o$ denote the parameters of the state space model. Δ_o represents a dynamic gating factor used to regulate the degree of state updating.

$$\mathbf{A}_o = \Delta_o \odot E_{A_o} \quad (16)$$

$$\mathbf{y}_o = \text{SSM}(\mathbf{A}_o, \mathbf{B}_o, \mathbf{C}_o)(\mathbf{x}'_o) \quad (17)$$

Here, $\mathbf{A}_o$ denotes the parameters of the state space model, while E_{A_o} represents a learnable linear projection matrix. The symbol $\odot$ denotes element-wise multiplication. By performing the above computation process separately in both directions, bidirectional state feature sequences $\mathbf{y}_{\text{fwd}}$ and $\mathbf{y}_{\text{bwd}}$ can be obtained. Finally, the forward and backward features are fused to generate the final output embedding:

$$\mathbf{y}'_{\text{fwd}} = \mathbf{y}_{\text{fwd}} \odot \text{SiLU}(\mathbf{z}) \quad (18)$$

$$\mathbf{y}'_{\text{bwd}} = \mathbf{y}_{\text{bwd}} \odot \text{SiLU}(\mathbf{z}) \quad (19)$$

$$\mathbf{T}_i = E_D(\mathbf{y}'_{\text{fwd}} + \mathbf{y}'_{\text{bwd}}) + \mathbf{Q}'_i \quad (20)$$

In the above equation, $i \in \{1, 2, \dots, D\}$ and E_D denotes a linear transformation matrix used to fuse the bidirectional features and restore them to the original dimensionality. This fusion mechanism preserves semantic information from both the forward and backward directions while enhancing the model's global feature perception capability, thereby providing more discriminative time–frequency feature representations for downstream tasks. The variable $\mathbf{z}$ serves as a gating feature, which introduces nonlinear modulation into the bidirectional state space outputs and enables adaptive control over feature intensity.

2.3. MSANet Learning Task Framework

The architecture adopted in this paper employs a self-supervised pretraining framework. In the self-supervised learning task, robust audio features can be learned through joint task optimization based on target generation and discrimination. The functions and roles of the related modules are described in detail in the following sections.

2.3.1. Random Spectrogram Patch Masking Strategy

After obtaining the spectrogram S , it is partitioned into a series of spectrogram patches to facilitate the model's capture of local time–frequency features. Each spectrogram patch is denoted as S_i , whose dimensions in the frequency and temporal domains are represented by P_f and J_f , respectively. During the pretraining stage, in order to enhance the model's structural modeling capability, a portion of the spectrogram patches is randomly masked and represented using masking indicators. The model is then required to infer the features of the masked regions based on the unmasked spectrogram patches, thereby learning the latent time–frequency structure and contextual correlations of audio signals during the prediction process.

In this study, the input log-Mel spectrogram is divided into patches with a size of 16×16 . The patch stride is set to 10 in the downstream experiments to increase the overlap between adjacent time–frequency regions and preserve more local contextual information. During the masked spectrogram patch modeling stage, a relatively high masking ratio is adopted to encourage the model to infer missing time–frequency structures from the visible patches. In the experiments, 400 spectrogram patches are randomly masked during representation learning. This setting increases the difficulty of the pretext task and helps the model learn more robust contextual representations from audio spectrograms.

2.3.2. Pretraining Task

During the model training stage, MSANet adopts a joint optimization strategy that integrates discriminative tasks with generative tasks to fully exploit the structural characteristics and semantic information of audio spectrograms. This design is related to masked spectrogram modeling methods such as SSAST and Audio-MAE, but differs in that the pretraining objective is built upon a bidirectional SSM encoder rather than a Transformer encoder. Through this multi-objective joint training paradigm, the model is capable of not only learning the reconstruction relationships of local time–frequency features, but also capturing more discriminative speech representations during the classification process. The overall training objective mainly consists of two types of subtasks, corresponding to generative reconstruction constraints and discriminative recognition constraints, respectively.

Discriminative Task: The objective of the model is to accurately identify the masked spectrogram patches. The classification head generates a corresponding vector representation for each masked feature patch and compares it with the embedding vectors of other spectrogram patches within the same batch to compute the similarity-based contrastive loss. This process adopts the InfoNCE (Info Noise Contrastive Estimation) criterion, which maximizes feature similarity among samples of the same category while minimizing the similarity among different categories. The corresponding loss function is formulated as follows:

$$L_p = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{\exp(\langle m_i, x_i \rangle)}{\sum_{j=1}^N \exp(\langle m_i, x_j \rangle)} \right) \quad (21)$$

In the above equation, m_i represents the discriminative output corresponding to the i -th masked patch, while x_i denotes the original input of the i -th masked patch. N represents the total number of masked patches.

Generative Task: This task is designed to recover the original content of the masked spectrogram patches. Based on the feature representations learned by the model, the reconstruction head predicts the masked spectrogram regions to approximate their corresponding ground-truth spectrogram patches. This process is achieved by minimizing the

mean squared error between the predicted results and the original spectrogram patches. The corresponding loss function can be expressed as follows:

$$L_c = \frac{1}{N} \sum_{i=1}^N \|\hat{x}_i - x_i\|^2 \quad (22)$$

Here, $\hat{x}_i$ denotes the predicted output of the masked spectrogram patch, while x_i represents the corresponding ground-truth spectrogram patch input.

To simultaneously balance the optimization effects of the discriminative and generative tasks, the overall loss function of the model is formulated as a weighted summation. By introducing a balancing coefficient β , the relative importance of the two tasks during training can be regulated. The overall loss function can be expressed as follows:

$$L = L_p + \beta L_c \quad (23)$$

In the joint learning objective, the discriminative loss and the generative reconstruction loss are optimized together. The reconstruction loss is assigned a relatively larger weight to strengthen the recovery of masked spectrogram regions, and the balancing coefficient β is set to 10 in the experiments. The contrastive discrimination branch is optimized based on the similarity between masked patch representations and their corresponding original patch embeddings. In the adopted implementation, no additional temperature hyperparameter is introduced for the contrastive loss. The model is trained using an Adam-based optimizer, and the learning rate is adjusted according to the validation performance during training.

3. Experimental Analysis

3.1. Preliminary Self-Supervised Pretraining Before Experimental Evaluation

Before conducting the self-supervised pretraining of MSANet, a training set covering diverse acoustic scenarios is required, enabling the model to learn stable voiceprint representations under unlabeled conditions. Publicly available datasets are selected, and the training process is designed to be independent of any category labels, thereby simulating the demand for voiceprint representation learning in real-world acoustic scenarios where annotations are unavailable. During the pretraining stage, all audio samples are used only in the form of their original waveform signals, without involving task-specific labels. This ensures that the model focuses on extracting time–frequency structural characteristics from voiceprints and learns generalizable audio representations, providing a more stable feature foundation for subsequent tasks.

To improve the reproducibility of the experiments, the details of the datasets used in this study are further described as follows. Speech Commands v0.02 is a public keyword spotting dataset containing 105,829 short audio clips from 35 spoken command categories. Each audio sample is approximately 1 s long and is recorded at a sampling rate of 16 kHz. In this study, the official training, validation, and testing split is adopted, containing 84,848, 9982, and 4890 samples, respectively. ESC-50 is a public environmental sound classification dataset containing 2000 audio clips from 50 semantic classes, with 40 samples in each class. Each clip has a duration of 5 s and a sampling rate of 44.1 kHz. The dataset is pre-arranged into five folds, and four folds are used for training while the remaining fold is used for validation/testing in each round of five-fold cross-validation.

For preprocessing, all audio waveforms are first converted into log-Mel spectrograms. Specifically, the short-time Fourier transform is performed using a 25 ms Hamming window with a frame shift of 10 ms, and 128 Mel filter banks are then applied to obtain 128-dimensional log-Mel spectrogram features. For Speech Commands, the original 16 kHz sampling rate is retained. For ESC-50, the audio signals are converted into the same feature

extraction format before being fed into the network. The generated spectrograms are divided into 16×16 patches with a stride of 10, which are then used as the input sequence of MSANet. During the self-supervised pretraining stage, random spectrogram patch masking is used as the pretext task. No additional data augmentation, such as noise mixing, pitch shifting, or time stretching, is used in the experiments.

This study uses Speech Commands and ESC datasets as public benchmark datasets to evaluate the general audio representation capability of the proposed model. Although these datasets are not collected from substation equipment, they contain diverse acoustic events, background noise, and time–frequency patterns, which can be used to verify the ability of MSANet to extract robust audio features, model temporal dependencies, and learn discriminative representations under limited labeled data. Therefore, the experiments in this paper should be regarded as a preliminary methodological validation of the proposed self-supervised voiceprint modeling framework, rather than a complete field validation for substation fault diagnosis. In future work, we will further collect real substation equipment voiceprint data and evaluate the proposed method under practical fault monitoring scenarios.

3.2. Performance Evaluation and Validation of MSANet

To evaluate the performance of the proposed MSANet model, experiments are conducted on the ESC-50 and Speech Commands datasets. The main experimental settings are as follows. For ESC-50, the target spectrogram length is set to 512, the batch size is set to 20, the learning rate is set to 2×10^{-4} , and the model is trained for 100 epochs. For Speech Commands, the target spectrogram length is set to 128, the batch size is set to 96, the learning rate is set to 2×10^{-4} , and the model is trained for 50 epochs. For both datasets, the number of Mel filter banks is set to 128, the patch size is set to 16×16 , and the stride is set to 10. Accuracy is adopted as the main evaluation metric. These public datasets are used because they provide standardized audio samples with diverse acoustic categories and background conditions, which are suitable for evaluating the robustness and generalization ability of audio representation learning methods. These datasets cover multiple representative environmental sound sources and are characterized by diverse categories and well-structured data organization. Serving as the primary evaluation platforms for downstream tasks, they are used to systematically assess the representation capability of the proposed model in acoustic event recognition scenarios.

During the fine-tuning stage, supervised training is adopted without applying the masking strategy. The complete temporal sequence features output by the encoder are aggregated to form a global voiceprint representation. Specifically, average pooling is performed on the encoded sequence features, followed by a classification layer to produce the final category prediction.

For the ESC dataset, a five-fold cross-validation protocol is employed, where four folds are used for training, and the remaining fold is used for validation. The final performance is obtained by averaging the results over multiple training runs. This strategy reduces the perturbations caused by random initialization, improves the stability of the experimental results, and ensures the consistency and reliability of the model in processing long audio sequences.

3.2.1. Performance Validation of the Model on the Speech Commands Dataset

Figure 3 presents the comparison between MSANet and several representative self-supervised acoustic pretraining methods, including APC, Wav2Vec, Wav2Vec 2.0, and HuBERT. It should be noted that the results of these baseline models are taken from previously reported results in the literature. The results indicate that MSANet achieves com-

petitive performance among representative self-supervised audio representation learning methods. By employing a bidirectional scanning Mamba-based framework, the proposed model can effectively process complex audio information, further validating its feature representation capability.

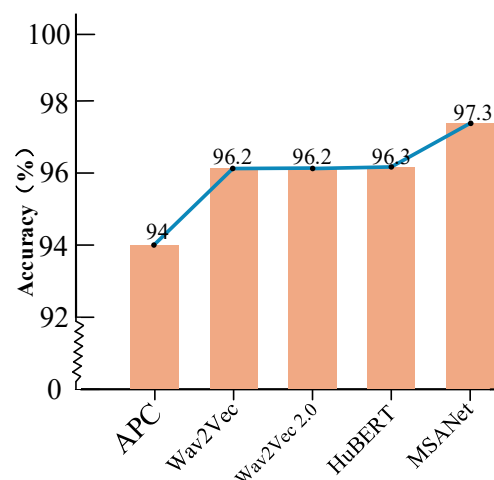


Figure 3. Comparison Results Between MSANet and Representative Model Frameworks.

Figure 4 illustrates the variation curve of the model accuracy on the validation set as the number of training iterations gradually increases. The accuracy of the model on the Speech Commands dataset exhibits a trend of rapid improvement followed by gradual stabilization. During the early stage of training, the accuracy quickly rises from a relatively low level to above 0.85, indicating that the model is capable of learning critical voiceprint features within a small number of iterations. Subsequently, the accuracy curve gradually approaches a stable region and remains at a relatively high level with only minor fluctuations. The model demonstrates stable generalization performance on the validation set without showing obvious signs of overfitting. These results indicate that the proposed MSANet possesses favorable convergence characteristics and strong feature representation capability.

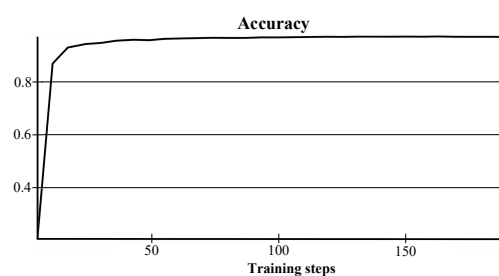


Figure 4. Variation Trend of Accuracy with Increasing Training Iterations.

Figure 5 illustrates the variation trend of the model loss on the validation set, which exhibits a convergent behavior from high to low values as the training iterations progress. During the initial stage of training, the validation loss remains at a relatively high level and then decreases rapidly, indicating that the model can efficiently capture critical acoustic features within a short training period. Subsequently, the loss curve gradually stabilizes and remains within a relatively low range, with only minor fluctuations observed throughout the training process. This behavior suggests that the parameter updates of the model become progressively stable without severe oscillatory variations.

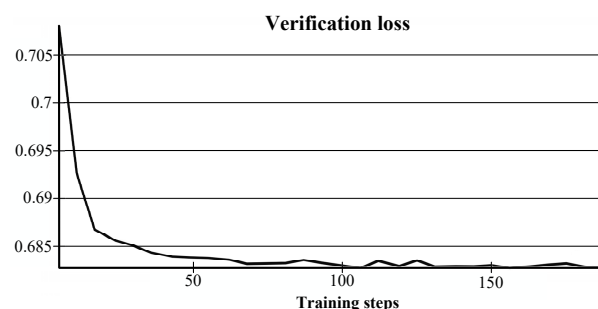


Figure 5. Variation Trend of Validation Loss Values.

3.2.2. Performance Validation of the Model on the ESC Dataset

Figure 6 presents the performance of MSANet on the ESC-50 dataset under a five-fold cross-validation setting. As observed from the figure, the overall performance remains relatively stable, with the accuracy of each fold ranging from 78% to 88%. Among them, the fourth fold achieves the highest accuracy of 87.7%, while the results of the first, second, and third folds exhibit only minor differences, indicating that the proposed model possesses consistent feature learning capability and strong feature extraction performance under different data partitions. In contrast, the accuracy of the fifth fold is slightly lower, which may be attributed to the relatively more complex category distribution within that fold. By averaging the results of all five folds, the overall accuracy of the model reaches 84.24%, demonstrating that MSANet exhibits favorable generalization ability and robustness in audio classification tasks. Furthermore, the proposed model is capable of effectively extracting voiceprint features and achieving competitive performance across diverse voiceprint-related tasks.

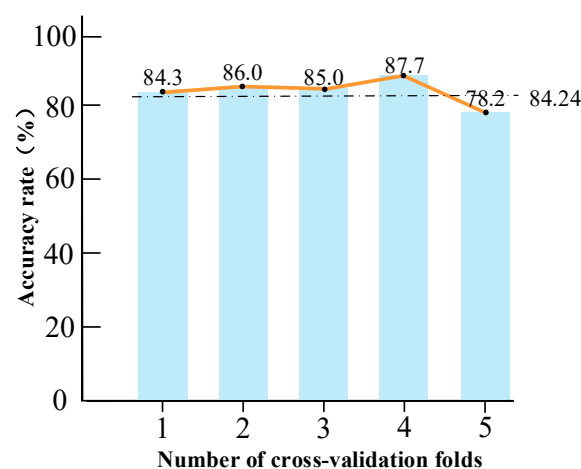


Figure 6. Accuracy Variation in Cross-Validation.

To further validate the stability of the model under different data partitions, the curves of mAP variation with respect to training iterations for each fold during cross-validation are illustrated in Figure 7. The five curves exhibit a highly consistent overall trend. During the early stage of training, the mAP increases rapidly, indicating that the model can capture the primary acoustic discriminative information within a relatively small number of training iterations. In the middle stage of training, the mAP curves of all folds gradually become stable and converge around 0.90 without exhibiting obvious performance fluctuations. Although slight differences exist among the folds, the overall convergence speed and final performance levels remain highly consistent, demonstrating favorable stability and robustness of the proposed model.

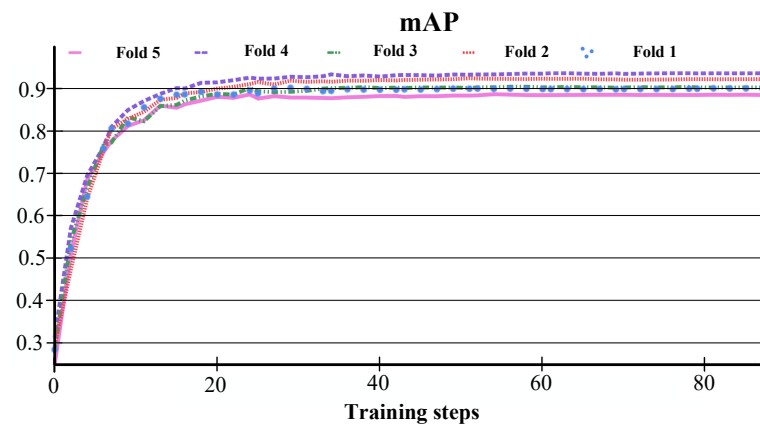


Figure 7. mAP Curves of Each Fold with Respect to Training Iterations During Cross-Validation.

3.3. Analysis of the Impact of Model Scale on Performance

To further evaluate the performance differences among models of varying scales in environmental sound recognition tasks, five-fold cross-validation experiments are conducted on three different configurations of MSANet based on the ESC dataset. The three model variants differ in terms of network depth and parameter scale, thereby reflecting the influence of model scale on feature extraction capability and classification performance. Table 1 presents the classification accuracies and average results of the three models under five-fold validation, providing an intuitive comparison of the stability and overall performance of different network scales in practical tasks.

Table 1. Comparison of Performance Among Different Model Scales.

	MSANet-Tiny	MSANet-Small	MSANet-Base
Fold 1 (Accuracy/%)	69.8	81.3	84.3
Fold 2 (Accuracy/%)	68.2	78.0	86.0
Fold 3 (Accuracy/%)	71.3	78.5	85.0
Fold 4 (Accuracy/%)	74.0	82.5	87.7
Fold 5 (Accuracy/%)	65.2	78.7	78.2
Average (Accuracy/%)	69.7	79.8	84.24

As shown in Table 1, the performance of MSANet on the ESC dataset exhibits a consistent improvement trend as the model scale increases. The Tiny model shows relatively limited overall performance, while the accuracy of the Small model increases to over 78%, indicating that moderately increasing the model capacity can significantly enhance its ability to represent voiceprint details. The Base model maintains high accuracy across all folds, achieving an average accuracy of 84.24%, and demonstrates the most stable and outstanding performance among the three model variants. These results indicate that, with enhanced feature extraction capability, MSANet gains clear advantages in voiceprint characterization, category discrimination, and pattern recognition in complex acoustic environments. They also verify the favorable scalability of the proposed architecture when the model scale is enlarged. Meanwhile, the small-scale model can still achieve satisfactory recognition accuracy while maintaining low computational overhead, providing a feasible basis for its deployment on resource-constrained edge devices for substation equipment fault detection.

In addition, the computational complexity of MSANet-Tiny is evaluated under the ESC-50 input setting. The model contains 6.4M parameters, with a storage size of 24.6 MB and a computational cost of 1.4 GFLOPs. The average inference time is 5.1 ms per sample,

and the peak CUDA memory consumption is 43.2 MB. These results further indicate that MSANet-Tiny has relatively low computational and storage overhead.

4. Conclusions

Taking the monitoring requirements of substation equipment voiceprint signals as the research background, this paper focuses on the problems of unstable voiceprint feature representation, difficulty in modeling long-range temporal dependencies, and insufficient labeled samples. To address these issues, a self-supervised voiceprint modeling model based on the state space model, termed MSANet, is proposed. By introducing a bidirectional state scanning structure, the proposed model achieves a balanced representation of the global structure and local details of voiceprint signals without relying on attention mechanisms. Combined with a spectrogram patch masking-based pretraining strategy, the network can learn time–frequency structural features under unlabeled conditions, thereby maintaining strong feature extraction capability even with limited data. Experimental results demonstrate that MSANet achieves high overall performance and exhibits favorable robustness and stability. Moreover, MSANet effectively models the temporal structure of voiceprint signals while maintaining low computational complexity, making it suitable for application scenarios with limited computational resources. Therefore, the experimental results provide preliminary evidence for the feasibility of deploying the proposed algorithm on edge devices in substation sites. However, since the current experiments are based on public audio datasets rather than real substation equipment fault voiceprint data, further field validation is still required. Future work will focus on collecting real substation equipment voiceprint data, constructing fault-related acoustic datasets, and evaluating the robustness of MSANet under practical multi-noise and multi-source monitoring environments.

Author Contributions: Conceptualization, J.L. and C.Y.; methodology, C.Y.; software, W.W.; validation, J.L., G.L. and Z.K.; formal analysis, Q.Z.; investigation, J.L.; resources, Z.K.; data curation, G.L.; writing—original draft preparation, J.L.; writing—review and editing, C.Y. and Q.Z.; visualization, W.W.; supervision, C.Y.; project administration, J.L.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the China Southern Power Grid Science and Technology Project, Research and Application of Voiceprint Monitoring Methods for Abnormal States of Substation Equipment Based on Multi-Source Data Fusion (Grant No. GDKJXM20240405).

Data Availability Statement: The datasets analyzed in this study are publicly available. The experimental results and processed data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: Authors Junju Lai, Wei Wang, Guangyao Li and Zhichong Kong were employed by Qingyuan Power Supply Bureau, Guangdong Power Grid Limited Liability Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Wang, C.; Gao, X.; Long, H. Carbon Emission Accounting for 220 kV Transmission and Transformation Projects Based on a Whole Life Cycle Improvement Method. *Energies* **2025**, *18*, 912. [\[CrossRef\]](#)
2. Zhang, Z.; Zhang, G.; Zhang, P.; Chen, J.; Liu, X.; Peng, G.; Zhou, F. Unsupervised Online Learning for Substation Equipment Defect Identification Based on SoftHebb. *Energy Internet* **2025**, *2*, 287–296. [\[CrossRef\]](#)
3. Du, X.; Lei, X.; Ye, T.; Sun, Y.; Shang, Z.; Liu, Z.; Xu, T. A review of research on intelligent fault detection of power equipment based on infrared and voiceprint: Methods, applications and challenges. *Glob. Energy Interconnect.* **2025**, *8*, 821–846. [\[CrossRef\]](#)
4. Zhou, Y.; He, Y.; Xing, Z.; Wang, L.; Shao, K.; Lei, L.; Li, Z. Vibration Signal-Based Fusion Residual Attention Model for Power Transformer Fault Diagnosis. *IEEE Sens. J.* **2024**, *24*, 17231–17242. [\[CrossRef\]](#)

5. Thinh, T.N.H.; Lam, P.D.; Tran, H.Q.; Tien, L.H.C.; Thai, P.H. Transformer Vibration and Noise Monitoring System Using Internet of Things. *IET Commun.* **2023**, *17*, 815–828. [[CrossRef](#)]
6. Bin, L.; Sheng, S.; Yuan, L.; Qiang, F.; Yunfeng, X.; Guanke, L. Spectro-Temporal Self-Similarity Based Identification of Corrupted Acoustic Signal of Distribution Transformer in Noisy Environment. *IEEE Trans. Power Deliv.* **2023**, *38*, 105–116. [[CrossRef](#)]
7. Wang, K.; Pang, Y.; Wang, W.; Yu, W.; Liang, J.; Liu, T.; An, Y. Switching Error Detection of Rheological Tap-Changer for UHVDC Transmission Based on Voice Print Feature. *Results Eng.* **2025**, *28*, 107730. [[CrossRef](#)]
8. Peng, J.; Ma, Y.; Ye, H.; Che, X.; Li, S.; Ai, B. Voiceprint recognition method of transformer based on LBT-ODF and MVN. *J. Meas. Eng.* **2025**, *13*, 387–404. [[CrossRef](#)]
9. Li, H.; Yao, Q.; Li, X. Voiceprint Fault Diagnosis of Converter Transformer under Load Influence Based on Multi-Strategy Improved Mel-Frequency Spectrum Coefficient and Temporal Convolutional Network. *Sensors* **2024**, *24*, 757. [[CrossRef](#)] [[PubMed](#)]
10. Wan, S.; Dong, F.; Zhang, X.; Wu, W.; Li, J. Fault Voiceprint Signal Diagnosis Method of Power Transformer Based on Mixup Data Enhancement. *Sensors* **2023**, *23*, 3341. [[CrossRef](#)] [[PubMed](#)]
11. Chen, J.; Zhao, T.; Wang, Y.; Jin, P.; Lin, C.; Zhang, H. Fault Pattern Recognition and Prediction of Transformer Voicing Based on Neural Network. *Diagnostyka* **2025**, *26*, 2025315. [[CrossRef](#)]
12. Yu, Z.; Wei, Y.; Niu, B.; Zhang, X. Automatic Condition Monitoring and Fault Diagnosis System for Power Transformers Based on Voiceprint Recognition. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 9600411. [[CrossRef](#)]
13. Zhang, K.; Lu, H.; Han, S.; Zhao, X. A Novel Fault Diagnosis Method for Power Transformers Based on Voiceprint Recognition Considering Multi-Type Noises. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 3541311. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.