Audio Scanning Network: Bridging Time and Frequency Domains for Audio Classification

Liangwei Chen[1], **Xiren Zhou**[2*], **Huanhuan Chen**[2*]

[1]School of Data Science, University of Science and Technology of China
[2]School of Computer Science and Technology, University of Science and Technology of China
clw2018@mail.ustc.edu.cn, zhou0612@ustc.edu.cn, hchen@ustc.edu.cn

Abstract

With the rapid growth of audio data, there's a pressing need for automatic audio classification. As a type of time-series data, audio exhibits waveform fluctuations in both the time and frequency domains that evolve over time, with similar instances sharing consistent patterns. This study introduces the Audio Scanning Network (ASNet), designed to leverage abundant information for achieving stable and effective audio classification. ASNet captures real-time changes in audio waveforms across both time and frequency domains through reservoir computing, supported by Reservoir Kernel Canonical Correlation Analysis (RKCCA) to explore correlations between time-domain and frequency-domain waveform fluctuations. This innovative approach empowers ASNet to comprehensively capture the changes and inherent correlations within the audio waveform, and without the need for time-consuming iterative training. Instead of converting audio into spectrograms, ASNet directly utilizes audio feature sequences to uncover associations between time and frequency fluctuations. Experiments on environmental sound and music genre classification tasks demonstrate ASNet's comparable performance to state-of-the-art methods.

Introduction

The wide array of recording devices available today has caused a tremendous surge in recorded and stored audio. Audio labeling is pivotal for content analysis and retrieval. However, the workload associated with manual labeling is considerable, underscoring the need for automated audio classification systems. Central to these systems is the extraction of distinct features from raw audio. Proper extraction and integration of audio features typically yield richer information, enhancing the system's efficacy. Various audio classification tasks have driven research in audio feature extraction (Sharma, Umapathy, and Krishnan 2020), facilitating a profound comprehension of audio nature and offering crucial impetus for advancing the field of audio analysis.

Typically, audio features could be segmented into two domains: the time domain (i.e., temporal features) and the frequency domain (i.e., spectral features) (Panda, Malheiro, and Paiva 2020). Time domain analysis primarily focuses on the evolution of audio signals over time, providing intuitive insights. In contrast, frequency domain analysis necessitates methodologies such as the Fourier transform to delve deeper into the frequency components of the audio. For audio classification tasks, temporal feature extraction captures nuanced fluctuations of audio signals over time. Concurrently, spectral features are equally indispensable, allowing not only the conversion of audio signals into spectral data but also revealing intricate inter-frequency relationships. A holistic approach that integrates insights from both time and frequency domains (Marafioti et al. 2019) promises a more comprehensive and accurate depiction of audio recognition.

Recent advances have converted audio into image-like representations, such as spectrograms, characterized by time (x-axis) against frequency (y-axis). Visual models like Convolutional Neural Networks (CNNs) and Vision Transformers (ViT) (Gong et al. 2022) are then utilized for efficient audio classification. However, challenges persist: 1) the conversion to spectrograms might compromise some intricate temporal details intrinsic to the raw audio waveform; 2) besides, a systematic approach for analyzing the correlation between waveform fluctuations in the time and frequency domains is absent, potentially overlooking their correlations essential for audio discrimination.

Our audio classification framework, the Audio Scanning Network (ASNet), is illustrated in Figure 1, which incorporates both frequency and time-domain modules. In the frequency-domain module, ASNet segments the audio with a fixed-size window, transforming it into feature sequences. We employ Mel-frequency cepstral coefficients (MFCCs) (Wang et al. 2020) to capture the spectral envelope and fine details, culminating in a spectral waveform representation. This module organizes the MFCCs sequentially, reflecting waveform evolutions over time. Correspondingly, the time-domain module creates a Root-Mean-Square Energy (RMS) (Luo, Xu, and Chen 2019) sequence for amplitude dynamics, well-reflecting waveform fluctuations. Also, considering the rate of waveform change in the time domain, we establish a Zero Crossing Rate (ZCR) (Toffa and Mignotte 2020) sequence. Both the above two modules leverage Reservoir Computing for efficiently capturing dynamic features. This process transforms sequences into a more intricate "dynamic feature space", effectively representing waveform evolutions through the linear readout model, and without the need for

*Corresponding authors: Xiren Zhou and Huanhuan Chen.

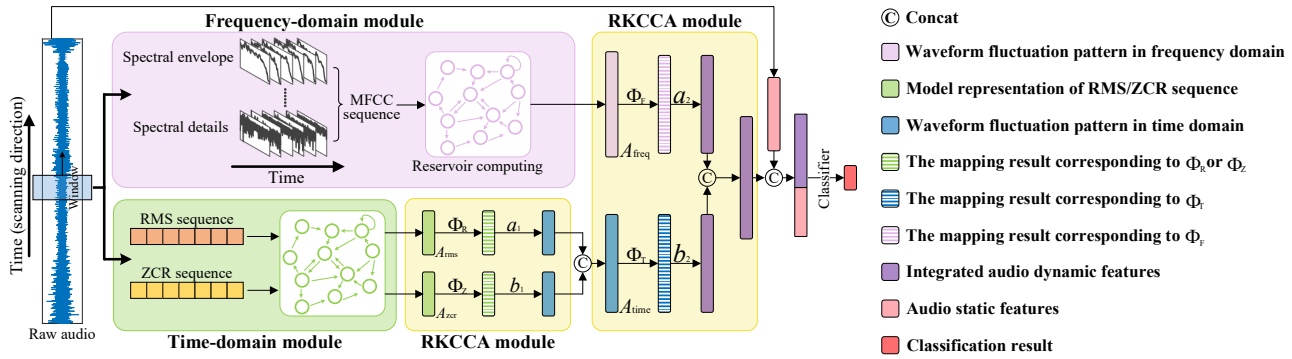


Figure 1: ASNet's framework comprises four modules: frequency and time-domain modules for capturing waveform fluctuation patterns, plus two RKCCA modules for inter-fluctuation correlation. Reservoir model processes RMS, ZCR, and MFCC sequences into readout models (A_{rms}, A_{zcr}, A_{freq}). A_{time} embodies fused A_{rms} and A_{zcr} through RKCCA. Nonlinear mapping Φ, linear transformations a and b are utilized. These are related to kernel methods, which will be detailed in the following text.

offline iterative training (Chen et al. 2015; Ma et al. 2020).

ASNet further focuses on exploring the correlation between the dynamic information present in both time-domain and frequency-domain[1]. Specifically, a novel Reservoir Kernel Canonical Correlation Analysis (RKCCA) is proposed within ASNet to adaptively uncover correlations between time-domain and frequency-domain signal waveform fluctuations[2]. As depicted in Figure 1, the RKCCA module builds the kernel by configuring the distance of the readout model obtained from both the time-domain and frequency-domain modules, resulting in an integrated audio dynamic feature vector. Additionally, underlying audio features, like MFCCs, RMS, and ZCR, termed as audio "static features", are also incorporated into ASNet, enhancing its classification prowess. Experimental results demonstrate that ASNet's classification performance is on par with state-of-the-art methods for Environmental Sound Classification (ESC) and Music Genre Classification (MGC) tasks.

The primary contributions of this paper are as follows:

- The time-domain and frequency-domain modules in ASNet capture audio waveform fluctuations in both domains respectively. Additionally, the RKCCA module within ASNet effectively analyzes and correlates waveform patterns between time and frequency domains. This process efficiently integrates the rich dynamic information of audio, enhancing the model's classification performance.
- When constructing audio feature sequences, ASNet scans audio using fixed-size windows. Subsequently, the reservoir computing model, employed to capture waveform fluctuations in ASNet, continues to utilize sliding windows with a fixed size. This design enables ASNet to naturally handle audio of varying lengths.
- The entire framework of ASNet does not involve any iterative training during the feature extraction process, and only matrix operations are required. This characteristic makes it an ideal choice with limited computing resources and training data.

[1]In this paper, the dynamic information of the feature sequence indicates the waveform fluctuation patterns.

[2]Kernel Canonical Correlation Analysis (KCCA) is employed to analyze nonlinear correlations between features and perform feature fusion (Akaho 2006).

Related Works

Audio covers an expansive range of categories, among which environmental sounds and music are two primary types. ESC tasks have broad applications, such as indoor monitoring and automation, robot audition, and hazard detection. Besides, MGC aids music retrieval systems in quickly identifying target tracks. This section presents research related to these two tasks and discusses some works relevant to Canonical Correlation Analysis.

ESC and MGC Tasks: Audio classification is generally bifurcated into two phases as suggested by (Cai and Zhang 2022): feature extraction and subsequent classification of these features. For MGC, a classifier was developed by using the Joint Sparse Low-Rank Representation (JSLRR) of the test set (Panagakis, Kotropoulos, and Arce 2014), relative to the training set, by resolving an appropriate convex problem. Another approach leveraged Instance-Specific Hidden Markov Models (ISHMMs) to derive representations (Chandrakala and Jayalakshmi 2019) for sound event recognition. In classifying environmental sounds, a fusion of local binary pattern features with traditional audio features proved effective (Toffa and Mignotte 2020).

Furthermore, several research works have utilized time-frequency diagrams as model input features to bridge the time and frequency domains. The Masked Conditional Neural Network (MCLNN) (Medhat, Chesmore, and Robinson 2020) was designed to capture time-frequency details and holistically learn features from both music and environmental sound datasets. Tools such as the Audio Spectrogram Transformer (AST) (Gong, Chung, and Glass 2021) and the audio MLP-mixer (AMM) network (Tripathi and Pandey 2023) have demonstrated notable classification results in the context of ESC. For MGC, there have been efforts towards advancing broadcast-based neural networks with the goal of bolstering localization and adaptability while maintaining

a compact parameter set (Heakl, Abdelgawad, and Parque 2022). Different from prior methods, our approach simultaneously considers waveform fluctuations in both time and frequency domains and explores their correlations.

Canonical Correlation Analysis: CCA (Hardoon, Szedmak, and Shawe-Taylor 2004) explores the correlation between two sets of variables, but its effectiveness is limited when the relationship between variables is nonlinear. KCCA (Akaho 2006) incorporated kernel techniques into CCA, enabling the incorporation of nonlinear relationships. CCA has been used in several tasks recently. Gabor features were extracted from the normalized face images and fused with HOG features using CCA (Haghighat, Abdel-Mottaleb, and Alhalabi 2016). CCA was combined with self-supervised learning on graphs to simplify model design (Zhang et al. 2021). In addition, a K-means clustering-based KCCA algorithm was proposed (Chen et al. 2022) for multimodal emotion recognition in human–robot interaction. In our work, the reservoir computing model combined with KCCA can handle variable-length audio data and improve the effect of audio classification.

Audio Scanning Network

The entire network consists of four modules: a frequency-domain module and a time-domain module for capturing dynamic audio feature information, and two RKCCA modules to explore the correlation between these dynamics.

Capturing Audio Waveform Fluctuation Patterns

In this section, we employ reservoir computing models to capture the dynamic information of audio feature sequences, specifically the fluctuation patterns of waveforms.

Building Feature Sequences by Scanning Audio: Given its nature as high-sampling-rate time-series data (Stan et al. 2011), directly capturing the audio's context can be challenging. Traditional feature extraction techniques use a sliding window to extract features from individual audio segments and then compute statistical attributes, such as mean and variance, across all segments. In this paper, ASNet utilizes fixed-size windows to scan the audio, preserving features in each segment for feature sequence reconstruction.

ASNet employs Mel-Frequency Cepstral Coefficients (MFCCs) sequences for audio frequency-domain waveform fluctuations. Besides, it integrates Root-Mean-Square Energy (RMS) and Zero-Crossing Rate (ZCR) sequences for time-domain waveform fluctuation patterns.

In the frequency-domain module, MFCCs are derived to illustrate energy distribution across frequency ranges. The MFCCs exhibit a hierarchy: lower-dimensional ones represent the spectral envelope, while higher-dimensional ones capture spectral details, thus the MFCC sequence depicts dynamic changes in spectral waveform shape over time.

Meanwhile, the time-domain module constructs RMS and ZCR sequences. RMS describes the amplitude envelope robustly, while ZCR captures the rate of amplitude changes. These two sequences encapsulate the fluctuation and fluctuation rate of the time-domain waveform, respectively.

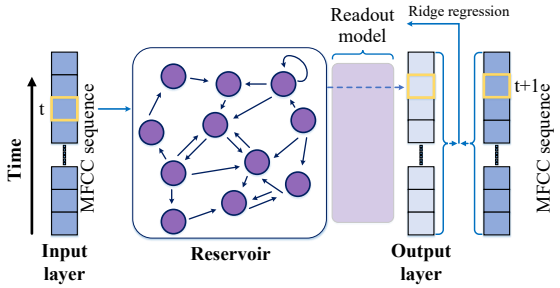


Figure 2: The figure shows the structure of the Echo State Network (ESN). The ESN consists of the input layer, the reservoir network, and the output layer. This work adopts the next-item prediction task to train the ESN.

Representing Waveform Fluctuation Patterns with Readout Models: For the feature sequences constructed above, we fit these sequences using the Echo State Network (ESN) to capture the waveform fluctuation patterns. ESN (Jaeger 2001), as a reservoir computing model, proves to be efficient in fitting sequential data and capturing its intrinsic dynamic information without the need for iterative training (Wu, Zhou, and Chen 2022; Gong et al. 2018; Zhou et al. 2023). This section provides a concise introduction to ESN and outlines the process of mapping sequence data to a higher-dimensional dynamic feature space using ESN.

ESN is a simplified Recurrent Neural Network (RNN) that uses a reservoir to replace traditional recurrent connections. The reservoir, shown in Figure 2, is a high-dimensional and sparsely connected structure, with its architecture and inter-neuron weights randomly determined based on the Echo State Property (ESP)[3] (Buehner and Young 2006). The output layer captures the target sequence by linearly combining the nonlinear dynamic features, and the output layer weights of the reservoir are obtained by solving the regression problem.

Assuming the input vector at time t is $s(t) \in \mathbb{R}^n$, the formula through the reservoir and the output layer can be expressed as follows:

$$\begin{cases} r(t) = f(W^{res} r(t-1) + W^{in} s(t)), \\ g(t) = W^{out} r(t) + d = A(r(t)), \end{cases} \tag{1}$$

where the state vector $r(t) \in \mathbb{R}^m$ of the reservoir at time t depends on both the current input vector $s(t)$ and the state vector $r(t-1)$ from the previous time step. The output $g(t) \in \mathbb{R}^l$ at time t is obtained by linearly transforming the state vector $r(t)$. The input weight between the input layer and the reservoir is denoted by $W^{in} \in \mathbb{R}^{m \times n}$, the reservoir weight matrix is denoted by $W^{res} \in \mathbb{R}^{m \times m}$, and the output weight between the reservoir and the output layer is denoted by $W^{out} \in \mathbb{R}^{l \times m}$. The activation function used (sigmoid or tanh function) is denoted as f, and the bias vector is denoted as d. W^{in} and W^{res} are randomly initialized and remain

[3]ESP requires the spectral radius of the internal connection weight matrix of the reservoir to be less than 1, which is determined by the maximum absolute value of its eigenvalues.

fixed, while W^{out} is the only parameter that needs to be determined through ridge regression.

Using ESN to fit the feature sequence captures its internal dynamic information, specifically the waveform fluctuation pattern, by predicting the sequence value at the next time step (Chen et al. 2013b). In Figure 2, the MFCC sequence from the frequency-domain module is used as an example. The cepstral coefficients $s(t)$ of the MFCC sequence at time t is input into the reservoir to obtain the predicted output $g(t)$. This generates the prediction sequence of the output layer. The predicted sequence g is combined with the MFCC sequence shifted one-time step later to perform ridge regression, resulting in the calculation of the readout model that characterizes the fluctuation pattern as follows:

$$W^{out} = (R^T R + \theta I)^{-1} R^T G, \tag{2}$$

where R is a matrix of the reservoir states, I is the identity matrix, and G is a matrix of the target value.

Based on the above, the dynamic information within the feature sequence is captured into the fitted readout model, a more stable and parsimonious representation of the original sequence for further process (Chen et al. 2013a).

Correlation between Frequency Domain and Time Domain Waveform Fluctuation Patterns

Building Kernel Matrix: An intuitive approach to transform the original signal space into the dynamic feature space is by using the trained model's parameters to characterize the data (Chen et al. 2015). This involves directly calculating the distance (e.g., Euclidean distance) between the parameters of the readout model. However, this method depends on specific model parameters and only reflects the distance of those parameters. In this paper, we will directly calculate the distance between the readout models.

The readout $A(r)$ that fits the sequence s_i is denoted as $A_i(r)$. When there are two sequences s_1 and s_2 passing through the same reservoir, their respective readouts $A_1(r)$ and $A_2(r)$ are obtained as follows:

$$\begin{aligned} A_1(r) &= W^{out_1} r + d_1, \\ A_2(r) &= W^{out_2} r + d_2, \end{aligned} \tag{3}$$

where $r \in \mathbb{R}^m$ represents the state vector, $d \in \mathbb{R}^l$ is the bias vector, and W^{out_1} and W^{out_2} are the respective output weight matrices. The L_2 distance between these weight matrices in the dynamic feature space can be measured as:

$$L_2(A_1, A_2) = \left(\int_{\mathbb{C}} \|A_1(r) - A_2(r)\|^2 d\mu(r) \right)^{\frac{1}{2}}, \tag{4}$$

$\mu(r)$ represents the probability density function of r, and $\mathbb{C}$ is the integral domain, with $\mathbb{C} = [-1, +1]^m$ when f indicates the tanh function. Assuming that r is uniformly distributed, a scaling process is provided (Chen et al. 2013b) to obtain the square of the distance between A_1 and A_2:

$$L_2(A_1, A_2) = \frac{1}{3} \sum_{j=1}^{m} \sum_{i=1}^{l} w_{i,j}^2 + \|d\|^2,$$

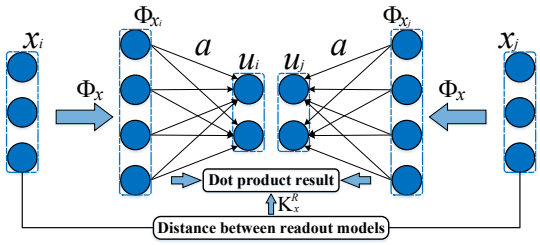


Figure 3: Illustration of RKCCA. Assuming x_i and x_j are two readout models in dynamic feature space, they are mapped to Hilbert space:$x_i \rightarrow \Phi_{x_i}, x_j \rightarrow \Phi_{x_j}$. Calculate the distance between x_i and x_j according to Equation (4) and construct the reservoir kernel of Equation (5): $\mathbf{K}^R = \Phi'_{x_i} \Phi_{x_j}$.

where $w_{i,j}$ is the (i, j)-th element of $W = W^{out_1} - W^{out_2}$, and $d = d_1 - d_2$. This approach yields the distance between the readout models in the dynamic feature space. Based on this distance, the reservoir kernel matrix can be constructed:

$$\mathbf{K}^R_{ij} = \exp\{-\gamma \cdot L_2^2(A_i, A_j)\}, \tag{5}$$

which is further adopted in our RKCCA.

Reservoir Kernel Canonical Correlation Analysis: Typical Canonical Correlation Analysis (CCA) investigates the linear correlation between two variable sets, X and Y, which aims to find linear combinations of variables within each set that maximize the correlation coefficient between the two sets. For the non-linear relationship between variables, KCCA further maps two variable sets into the Hilbert space (Akaho 2006) through nonlinear mapping, given as:

$$X \rightarrow \Phi_X, Y \rightarrow \Phi_Y, \tag{6}$$

where $\Phi_X \in \mathbb{R}^{k \times p}$, $\Phi_Y \in \mathbb{R}^{k \times q}$, p and q represent the feature dimensions of the mapped X and Y respectively, and k is the number of examples. KCCA then performs linear transformations on the mapped two variable sets:

$$\begin{aligned} U &= a' \Phi'_X, \\ V &= b' \Phi'_Y, \end{aligned} \tag{7}$$

where U and V are canonical variables. The canonical correlation coefficient between U and V is defined as λ. The concatenation or summation of U and V represents the fused outcome obtained from the correlation analysis of Φ_X and Φ_Y. Given the mapping as Equation (6), $\Phi\Phi'$ is constructed for the application of kernel methods to solve U and V.

In this paper, we propose RKCCA by introducing the reservoir kernel from Equation (5) into KCCA. Figure 3 illustrates a schematic diagram of RKCCA, depicting the construction of reservoir kernels between two samples in the dynamic feature space. After transforming Equation (5) into matrix form, it could be obtained : $\mathbf{K}^R_X = \Phi_X \Phi'_X, \mathbf{K}^R_Y = \Phi_Y \Phi'_Y$. Consequently, the process for maximizing the canonical correlation coefficient λ could be formulated as:

$$\begin{aligned} \max \quad & \alpha' \mathbf{K}^R_X \mathbf{K}^R_Y \beta \\ s.t. \quad & \alpha' \mathbf{K}^R_X \mathbf{K}^R_X \alpha = 1 \\ & \beta' \mathbf{K}^R_Y \mathbf{K}^R_Y \beta = 1, \end{aligned} \tag{8}$$

where α and β are scalars. Equation (8) can be solved by the Lagrangian multiplier method, given as:

$$\begin{aligned}\mathbf{K}_X^R\mathbf{K}_Y^R\beta &= \lambda(\frac{\lambda_1}{\lambda}\mathbf{K}_\mathbf{X}^{\mathbf{R}\,2} - \frac{\eta}{\lambda}\mathbf{I})\alpha,\\ \mathbf{K}_Y^R\mathbf{K}_X^R\alpha &= \lambda(\frac{\lambda_2}{\lambda}\mathbf{K}_\mathbf{Y}^{\mathbf{R}\,2} - \frac{\eta}{\lambda}\mathbf{I})\beta.\end{aligned} \tag{9}$$

It could be defined that:

$$\begin{aligned}\mathbf{Z}_1 &= \begin{bmatrix}0 & \mathbf{K}_X^R\mathbf{K}_Y^R\\ \mathbf{K}_Y^R\mathbf{K}_X^R & 0\end{bmatrix},\\ \mathbf{Z}_2 &= \begin{bmatrix}\frac{\lambda_1}{\lambda}\mathbf{K}_X^{R\,2} - \frac{\eta}{\lambda}\mathbf{I} & 0\\ 0 & \frac{\lambda_2}{\lambda}\mathbf{K}_Y^{R\,2} - \frac{\eta}{\lambda}\mathbf{I}\end{bmatrix},\\ \delta &= \begin{bmatrix}\alpha\\ \beta\end{bmatrix},\end{aligned} \tag{10}$$

thus Equation (9) could be transformed as:

$$\mathbf{Z}_1\delta = \lambda\mathbf{Z}_2\delta \Rightarrow \mathbf{Z}_2^{-1}\mathbf{Z}_1\delta = \lambda\delta, \tag{11}$$

where $\lambda_1 = \lambda_2 = \alpha'\mathbf{K}_X\mathbf{K}_Y\beta \triangleq \lambda$ is deduced through the Lagrange multiplier formulation without adding a regular term, that is, λ_1, λ_2 and λ are equal to the correlation coefficients of U and V. Thus U and V could be solved by finding eigenvalues and eigenvectors in Equation (11).

The above calculation explains the main process of RKCCA. By solving Equation (11), U and V could be obtained. RKCCA introduces an additional mapping from the original data space to the dynamic feature space compared to KCCA. This provides a more stable representation of dynamic information for subsequent correlation analysis.

Analyzing the Correlation of Fluctuation Patterns: This section explains how to use RKCCA to explore the correlation between waveform fluctuations and integrate this information into a feature vector.

In Figure 4, there are two RKCCA modules. The first one operates on the readout model A_{rms} for the RMS sequence and the readout model A_{zcr} for the ZCR sequence. By substituting A_{rms} and A_{zcr} into the variables X and Y in Equation (6), we obtain the time-domain model representation A_{time} through the first RKCCA analysis. Similarly, by substituting A_{time} and A_{freq} into Equation (6), the integrated audio dynamic features are obtained through the second RKCCA module. Here, A_{freq} is derived from capturing the dynamic information of the MFCC sequence.

The first RKCCA module explores the correlation between temporal waveform fluctuations and fluctuations in waveform rate of change, and utilizes A_{time} to represent the fusion of information from both, namely the temporal waveform fluctuation patterns. In the second RKCCA module, the correlation between temporal waveform fluctuation patterns and frequency-domain waveform fluctuation patterns is analyzed, resulting in integrated dynamic features. Algorithm 1 outlines the aforementioned process. The acquired integrated dynamic features are combined with static features (including MFCC, RMS, and ZCR) for audio classification.

In ASNet, it should be noted that MFCC sequence contains the dynamic information of the waveform, thus A_{freq} that captures its dynamic information is adopted to represent the waveform fluctuation pattern in the frequency domain, without the need for extra RKCCA.

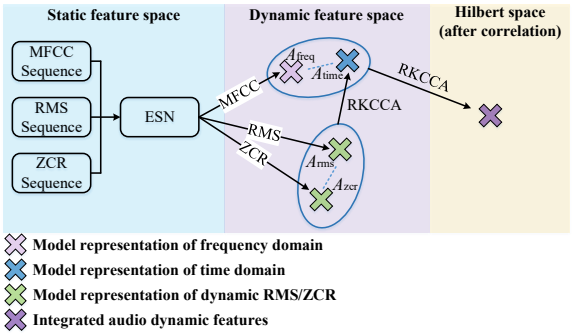


Figure 4: The Reservoir Computing model in ASNet transfers data from the static feature space to the dynamic feature space, while RKCCA integrates dynamic features and maps them to the Hilbert space after correlation analysis.

Algorithm 1: Exploring the Correlation of Waveform Fluctuation Pattern

Input: Set of audio $s_1, s_2, \ldots, s_k$; parameters of ASNet.
Output: Integrated audio dynamic features.

1: **for** each audio $s_i, i = 1, 2, \ldots, k$ **do**
2: Construct audio feature sequences: MFCC sequence, RMS sequence and ZCR sequence.
3: Reservoir state evolution is driven by input feature sequences, respectively.
4: The readout models corresponding to the feature sequences were obtained by ridge regression: A^i_{rms}, A^i_{zcr} and A^i_{freq}.
5: **end for**
6: Put all the obtained $A^i_{\text{rms}}, A^i_{\text{zcr}}, A^i_{\text{freq}}$ together to get $A_{\text{rms}}, A_{\text{zcr}}, A_{\text{freq}}, i = 1, 2, \ldots, k$.
7: First RKCCA module fuses A_{rms} and A_{zcr} into A_{time}.
8: Second RKCCA module analyzes the correlation of A_{time} and A_{freq} and fuses them into a representation of the integrated audio dynamics.

Experimental Studies

We used ASNet for environmental sound and music genre classifications, with results in Table 1. We first assessed how different audio feature combinations affected classification, validating the importance of waveform fluctuation patterns in audio differentiation. We then compared ASNet with advanced audio classification techniques, emphasizing its consistent performance across varied datasets. Importantly, ASNet effectively handles datasets with varied audio lengths.

The Utilized Datasets

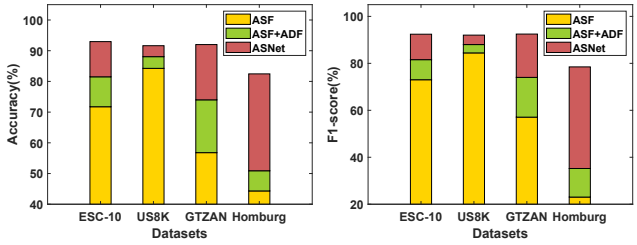
ESC-10 Dataset: The ESC-10 dataset (Piczak 2015) consists of 400 5-second environmental soundtracks with 10 classes, each containing 40 samples. The categories include dog, rooster, rain, sea waves, crackling fire, crying baby, sneezing, clock alarm, helicopter, and chainsaw.

Method	ESC-10				US8K				GTZAN				Homburg			
	SVM		RF		SVM		RF		SVM		RF		SVM		RF	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
RMS(S)	41.75	43.17	53.50	53.91	23.85	23.93	42.48	44.57	24.30	25.43	34.10	33.52	19.89	14.30	39.98	19.93
RMS(S+D)	44.25	46.04	53.75	54.07	27.92	27.96	43.12	44.80	33.20	33.58	39.50	39.61	24.50	16.83	41.68	20.26
ZCR(S)	33.50	34.87	43.50	45.39	23.71	24.86	40.30	41.65	23.80	23.21	30.20	30.48	20.78	14.69	40.99	20.27
ZCR(S+D)	41.25	43.32	48.75	50.29	28.64	27.91	44.92	45.26	30.30	30.11	39.10	39.19	28.47	19.05	41.63	20.70
MFCC(S)	77.75	77.95	78.00	79.34	72.97	72.79	84.45	84.19	66.60	67.17	66.40	66.61	49.63	33.89	47.67	28.27
MFCC(S+D)	81.50	82.71	80.00	80.35	79.98	79.85	88.42	88.10	73.00	73.54	72.50	72.93	51.59	38.05	49.95	30.64
ASF	69.00	70.33	71.75	73.02	72.62	74.15	84.25	84.45	46.10	48.57	56.80	57.09	34.52	23.05	44.33	20.32
ASF+ADF	81.50	81.57	81.25	81.53	78.15	78.72	88.06	88.00	72.40	72.61	74.00	74.02	50.10	35.22	50.90	30.14
ASNet(CCA)	78.50	77.23	79.25	80.42	65.26	62.36	74.13	69.76	70.60	71.50	73.10	72.92	48.52	35.05	50.43	30.35
ASNet(KCCA-P)	76.75	77.78	79.50	80.29	66.79	64.09	73.23	70.12	74.30	74.86	72.80	72.81	50.85	36.51	50.27	31.64
ASNet(KCCA-G)	77.50	77.77	79.25	80.47	77.11	78.01	80.81	82.06	82.30	82.43	79.20	79.51	**93.00**	**90.38**	76.56	67.68
ASNet	90.00	90.80	**92.96**	**92.41**	86.83	87.77	**91.63**	**92.02**	**92.00**	**92.47**	89.70	90.15	82.45	78.49	70.37	62.69

Table 1: Accuracy and F1-score are employed to assess the classification performance of various feature combinations. The classifiers utilized are Support Vector Machine (SVM) and Random Forest (RF).

US8K		Homburg	
Category	**Number**	**Genre**	**Number**
Air conditioner	1000	Alternative	145
Car horn	429	Blues	120
Children playing	1000	Electronic	113
Dog bark	1000	Folk/Country	222
Drilling	1000	Funk/Soul	47
Engine idling	1000	Jazz	319
Gun shot	374	Pop	116
Jackhammer	1000	Rap/Hip-hop	300
Siren	929	Rock	504
Street music	1000	-	-

Table 2: Information on imbalanced datasets



(a) Accuracy (b) F1-score

Figure 5: The impact of analyzing the correlation between waveform fluctuation patterns on classification performance.

Urbansound8k Dataset: The Urbansound8k (US8K) dataset is collected from various real-world environments (Salamon, Jacoby, and Bello 2014), comprising a total of 8732 samples, each with a duration of up to 4 seconds. It consists of 10 sound categories, and the specific categories and their respective quantities are shown in Table 2.

GTZAN Dataset: The GTZAN dataset was proposed in 2002 (Tzanetakis and Cook 2002), which contains 1000 songs and is divided into 10 genres. Each genre contains 100 tracks and each track is 30 seconds long. GTZAN is a dataset with balanced data, and the style gap between each genre is relatively obvious.

Homburg Dataset: The Homburg dataset (Homburg et al. 2005) contains a total of 1886 10-second audios. The amount of audio in each genre is unbalanced and some genres are classified with similar styles into one category, so it could be more challenging to classify the genres. The specific genres and quantities are given in Table 2.

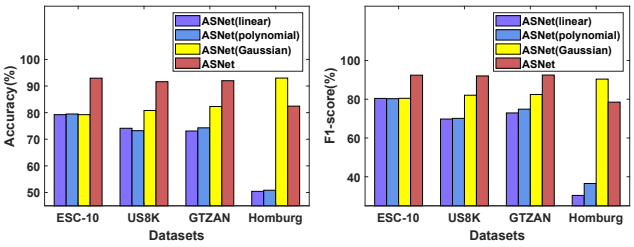
Experimental Configuration

Deployment Details: The experiments in this paper use MATLAB R2020b with a reservoir size of 50. Kernel scale parameter $\gamma \in \{10^{-6}, 10^{-5}, \ldots, 10^{-1}\}$. Ridge regression parameter $\theta \in \{10^{-5}, 10^{-4}, \ldots, 10^{1}\}$. The ESC-10 dataset is evaluated using 5-fold cross-validation, while all other datasets are evaluated using 10-fold cross-validation.

Baseline Methods: The performance of ASNet is evaluated from three aspects: 1) Comparing with methods that directly concatenate static and dynamic audio features without correlation analysis for classification. 2) Using KCCA combined with other kernel functions to explore the correlation of waveform patterns as a reference. 3) Comparison with some of the current state-of-the-art audio classification methods, which have been described in related works.

Experimental Results and Specific Analysis

Table 1 presents the impact of waveform fluctuation patterns and their correlations on audio classification. "S" denotes Audio Static Features (ASF), and "D" represents Audio Dynamic Features (ADF) without correlation analysis, which refer to waveform fluctuations. For instance, RMS(S) represents the untreated static RMS sequence, while RMS(S+D) signifies the concatenation of the static RMS sequence with the dynamic information of the RMS sequence acquired through reservoir computing. CCA stands for replacing the RKCCA module in ASNet with CCA, while KCCA-P and KCCA-G respectively replace the reservoir kernel with



(a) Accuracy (b) F1-score

Figure 6: The impact of accurately analyzing the correlation between waveform fluctuation patterns on classification.

polynomial kernel and Gaussian kernel. We use accuracy and F1-score to measure classification performance.

Effect of Waveform Patterns and Their Correlations: Table 1 shows that introducing dynamic features improves classification performance to varying extents across both audio tasks, regardless of whether single or multiple audio features are used. The degree of this improvement is task-specific, reflecting the unique sensitivity to different dynamic information. This highlights the critical role of both frequency-domain and time-domain waveform fluctuations in enhancing audio discrimination. As evidenced in Figures 5(a) and 5(b), ASNet, utilizing RKCCA to analyze correlations between these waveform fluctuations, surpasses the straightforward merging of static and dynamic features without correlation analysis in classification efficacy.

The Homburg dataset is characterized by a notable class imbalance, leading to significant differences between F1-score and accuracy when classifying directly from audio features. In Figures 5(b) and 6(b), methods lacking proper waveform fluctuation correlation analysis exhibit a larger disparity in F1-score with ASNet on the Homburg dataset compared to other datasets. ASNet's incorporation of correlation analysis not only bolsters accuracy but also narrows the gap between F1-score and accuracy. This accentuates ASNet's prowess in handling imbalanced classification.

Effectiveness of Reservoir Kernel: Figures 6(a) and 6(b) show that, in terms of classification performance on the ESC-10, US8K, and GTZAN datasets, ASNet using the reservoir kernel surpasses other kernels in the figure. Though the Gaussian kernel in ASNet excels on the Homburg dataset, the reservoir kernel consistently proves more stable. Table 1 indicates that linear (CCA without kernel trick) or polynomial kernels don't notably boost classification and may even underperform compared to merely merging static and dynamic features without correlation analysis. These results underline the importance of precise correlation analysis between frequency and time-domain waveform fluctuations in audio classification, with the reservoir kernel standing out for its stability and efficiency.

Comparison with State-of-the-Art Methods: Table 3 and Table 4 compare the classification performance of ASNet and some advanced methods on ESC and MGC tasks, respectively. For the ESC-10, GTZAN, and Homburg datasets, ASNet consistently achieves superior accuracy compared to the other methods. On the US8K dataset, The accuracy of ASNet is also close to that of AMM, which has a much larger parameter size. While much of the audio classification research heavily focuses on deep learning techniques, our model serves as a viable alternative, especially when computational resources or data are limited.

Method	ESC-10	US8K
ISHMMs (Chandrakala et al. 2019)	74.00	85.47
MCLNN (Medhat et al. 2020)	85.50	74.22
Toffa-FS (Toffa et al. 2020)	88.50	-
AST (Gong et al. 2021)	90.03	90.86
AMM (Tripathi et al. 2023)	92.93	**93.77**
ASNet	**92.96**	91.63

Table 3: Comparison with state-of-the-art methods on ESC.

Method	GTZAN	Homburg
JSLRR (Panagakis et al. 2014)	89.40	63.46
MCLNN (Medhat et al. 2020)	85.10	61.45
BN (Heakl et al. 2022)	90.00	64.00
Cai-FS (Cai et al. 2022)	91.80	65.20
ASNet(KCCA-G)	82.30	**93.00**
ASNet	**92.00**	82.45

Table 4: Comparison with state-of-the-art methods on MGC.

Regarding the datasets used in these experiments, there's a noticeable variance in audio lengths. This inconsistency is especially prominent in the US8K dataset, which features a mix of audio durations. However, ASNet maintains stable classification performance across all these datasets, emphasizing its proficiency in managing audio classification tasks with varied audio lengths.

Conclusion

The proposed Audio Scanning Network (ASNet) proficiently captures waveform fluctuations in both the time and frequency domains. Additionally, it employs the Reservoir Kernel Canonical Correlation Analysis (RKCCA) module to analyze their correlations and fuse them into an integrated audio dynamic feature vector, significantly enhancing the model's classification capacity. ASNet focuses on the inherent changing characteristics within both the time and frequency domains of the data itself. Besides, the introduction of the reservoir computing model in ASNet avoids iterative training. These factors make it more suitable for data-expensive tasks. Looking ahead, ASNet combines dynamic information from both the time and frequency domains in a high-dimensional space, thereby offering audio classification models more valuable features to enhance accuracy and stability. This opens up new possibilities in audio analysis and classification. Additionally, its ability to handle different audio lengths provides a practical solution, especially in situations where computational resources and data are limited. We expect that future research will expand and customize ASNet to address specific challenges in audio classification, contributing to the ongoing development of this field.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (No. 2021ZD0111700), the National Nature Science Foundation of China (No. 62206261, 62137002, 62176245), in part by the Key Research and Development Program of Anhui Province (No. 202104a05020011), in part by the Key Science and Technology Special Project of Anhui Province (No. 202103a07020002), and the Special Foundation for Science and Technology Innovation and Entrepreneurship of CCTEG (No. 2020-2-TD-CXY006).

References

Akaho, S. 2006. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*.

Buehner, M.; and Young, P. 2006. A tighter bound for the echo state property. *IEEE transactions on neural networks*, 17(3): 820–824.

Cai, X.; and Zhang, H. 2022. Music genre classification based on auditory image, spectral and acoustic features. *Multimedia Systems*, 28(3): 779–791.

Chandrakala, S.; and Jayalakshmi, S. 2019. Generative model driven representation learning in a hybrid framework for environmental audio scene and sound event recognition. *IEEE Transactions on Multimedia*, 22(1): 3–14.

Chen, H.; Tang, F.; Tino, P.; Cohn, A. G.; and Yao, X. 2015. Model Metric Co-Learning for Time Series Classification. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 3387–3394. AAAI Press.

Chen, H.; Tang, F.; Tino, P.; and Yao, X. 2013a. Model-based kernel for efficient time series analysis. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 392–400.

Chen, H.; Tiňo, P.; Rodan, A.; and Yao, X. 2013b. Learning in the model space for cognitive fault diagnosis. *IEEE transactions on neural networks and learning systems*, 25(1): 124–136.

Chen, L.; Wang, K.; Li, M.; Wu, M.; Pedrycz, W.; and Hirota, K. 2022. K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human–robot interaction. *IEEE Transactions on Industrial Electronics*, 70(1): 1016–1024.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Gong, Y.; Lai, C.-I.; Chung, Y.-A.; and Glass, J. 2022. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10699–10709.

Gong, Z.; Chen, H.; Yuan, B.; and Yao, X. 2018. Multiobjective learning in the model space for time series classification. *IEEE transactions on cybernetics*, 49(3): 918–932.

Haghighat, M.; Abdel-Mottaleb, M.; and Alhalabi, W. 2016. Fully automatic face normalization and single sample face recognition in unconstrained environments. *Expert Systems with Applications*, 47: 23–34.

Hardoon, D. R.; Szedmak, S.; and Shawe-Taylor, J. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12): 2639–2664.

Heakl, A.; Abdelgawad, A.; and Parque, V. 2022. A Study on Broadcast Networks for Music Genre Classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Homburg, H.; Mierswa, I.; Möller, B.; Morik, K.; and Wurst, M. 2005. A Benchmark Dataset for Audio Classification and Clustering. In *ISMIR*, volume 2005, 528–31.

Jaeger, H. 2001. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148(34): 13.

Luo, Z.; Xu, H.; and Chen, F. 2019. Audio Sentiment Analysis by Heterogeneous Signal Features Learned from Utterance-Based Parallel Neural Network. In *AffCon@ AAAI*, 80–87.

Ma, Q.; Li, S.; Zhuang, W.; Wang, J.; and Zeng, D. 2020. Self-supervised time series clustering with model-based dynamics. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9): 3942–3955.

Marafioti, A.; Perraudin, N.; Holighaus, N.; and Majdak, P. 2019. Adversarial generation of time-frequency features with application in audio synthesis. In *International conference on machine learning*, 4352–4362. PMLR.

Medhat, F.; Chesmore, D.; and Robinson, J. 2020. Masked conditional neural networks for sound classification. *Applied Soft Computing*, 90: 106073.

Panagakis, Y.; Kotropoulos, C. L.; and Arce, G. R. 2014. Music genre classification via joint sparse low-rank representation of audio features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12): 1905–1917.

Panda, R.; Malheiro, R. M.; and Paiva, R. P. 2020. Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing*.

Piczak, K. J. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, 1015–1018.

Salamon, J.; Jacoby, C.; and Bello, J. P. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, 1041–1044.

Sharma, G.; Umapathy, K.; and Krishnan, S. 2020. Trends in audio signal feature extraction methods. *Applied Acoustics*, 158: 107020.

Stan, A.; Yamagishi, J.; King, S.; and Aylett, M. 2011. The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication*, 53(3): 442–450.

Toffa, O. K.; and Mignotte, M. 2020. Environmental sound classification using local binary pattern and audio features collaboration. *IEEE Transactions on Multimedia*, 23: 3978–3985.

Tripathi, A. M.; and Pandey, O. J. 2023. Divide and distill: new outlooks on knowledge distillation for environmental sound classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1100–1113.

Tzanetakis, G.; and Cook, P. 2002. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5): 293–302.

Wang, J.; Xue, M.; Culhane, R.; Diao, E.; Ding, J.; and Tarokh, V. 2020. Speech emotion recognition with dual-sequence LSTM architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6474–6478. IEEE.

Wu, J.; Zhou, X.; and Chen, Q. 2022. A Characteristic of Speaker's Audio in the Model Space Based on Adaptive Frequency Scaling. In *2022 8th International Conference on Big Data and Information Analytics (BigDIA)*, 99–103. IEEE.

Zhang, H.; Wu, Q.; Yan, J.; Wipf, D.; and Yu, P. S. 2021. From canonical correlation analysis to self-supervised graph neural networks. *Advances in Neural Information Processing Systems*, 34: 76–89.

Zhou, X.; Liu, S.; Chen, A.; Chen, Q.; Xiong, F.; Wang, Y.; and Chen, H. 2023. Underground Anomaly Detection in GPR Data by Learning in the C3 Model Space. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–11.