Abeßer *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2025) 2025:1
https://doi.org/10.1186/s13636-024-00390-2

EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

Sound recurrence analysis for acoustic scene classification

Jakob Abeßer[1*†], Zhiwei Liang[1,2†] and Bernhard Seeber[2]

Abstract

In everyday life, people experience different soundscapes in which natural sounds, animal noises, and man-made sounds blend together. Although there have been several studies on the importance of recurring sound patterns in music and language, the relevance of this phenomenon in natural soundscapes is still largely unexplored. In this article, we study the repetition patterns of harmonic and transient sound events as potential cues for acoustic scene classification (ASC). In the first part of our study, our aim is to identify acoustic scene classes that exhibit characteristic sound repetition patterns concerning harmonic and transient sounds. We propose three metrics to measure the overall prevalence of sound repetitions as well as their repetition periods and temporal stability. In the second part, we evaluate three strategies to incorporate self-similarity matrices as an additional input feature to a convolutional neural network architecture for ASC. We observe the characteristic repetition of transient sounds in recordings of "park" and "street traffic" as well as harmonic sound repetitions in acoustic scene classes related to public transportation. In the ASC experiments, hybrid network architectures, which combine spectrogram features and features from sound recurrence analysis, show increased accuracy for those classes with prominent sound repetition patterns. Our findings provide additional perspective on the distinctions among acoustic scenes previously primarily ascribed in the literature to their spectral features.

Keywords Acoustic scene classification, Sound recurrence analysis, Sound repetition patterns, Self-similarity matrix, Harmonic-percussive source separation, Result fusion, Ensemble models

1 Introduction

During the course of their daily lives, humans are surrounded by a wide variety of different soundscapes. A soundscape defines the "acoustic environment as perceived or experienced and/or understood by a person or people, in context" [1]. These sounds come from various sources and include, among others, geophonic sounds such as running water and wind, biophonic sounds coming from animals such as birds or insects, and anthrophonic sounds generated from human activities such as industrial or traffic noise [2, 3]. Even complex acoustic environments with multiple concurrent and consecutive sound events can be decomposed by the human auditory system, allowing us to locate, detect, and classify individual sound events. In the research field of machine listening, sound event detection (SED) [4] aims to detect and classify individual sound events as local temporal-spectral patterns, while acoustic scene classification (ASC) [5] aims at a high-level categorization of the entire soundscape. Common ASC taxonomies cover between 10 and 15 classes of indoor and outdoor scenes [6].

Audio recordings from different domains, such as speech, music, and everyday sounds, differ in their temporal structure and especially in the occurrence of sound repetitions. For example, the temporal structure

†Jakob Abeßer and Zhiwei Liang contributed equally to this work.

*Correspondence:
Jakob Abeßer
jakob.abesser@idmt.fraunhofer.de
[1] Semantic Music Technologies, Fraunhofer IDMT, Ehrenbergstr. 31, 98693 Ilmenau, Germany
[2] Audio Information Processing, TU München, Theresienstr. 90, 80333 München, Germany

of speech signals is influenced by prosody, i. e., rhythmic and intonational aspects of spoken language. Music, on the other hand, exhibits structure and repetitions on multiple levels, from individual notes to entire segments. At the most basic level, notes are sometimes repeated or arranged in patterns to create melodies and rhythms. These notes are further grouped into phrases that are repeated or varied to create larger sections of a song. At the highest level, the structure of a musical piece is defined, among others, by repeating segments, such as verses and choruses. This hierarchical organization of repetition and variation is a fundamental aspect of musical compositions [7]. Despite the large number of independent sound sources, studies have shown that natural soundscapes also exhibit repeating sound structures on multiple time scales [8, 9]. Examples range from short-term tone repetitions, such as those found in bird songs in natural soundscapes, to long-term repetition, such as the daily ebb and flow of traffic noise in urban soundscapes.

In this work, we address two main research questions. First, we systematically investigate the temporal structure in different types of acoustic scenes. More concretely, we study the self-similarity matrix (SSM) as a structural representation for the overall task of identifying repeating sound patterns in audio signals, i. e., sound recurrence analysis (SRA). Second, we compare different strategies to integrate SRA into a deep learning-based ASC model to improve its performance. The main contributions of this work are as follows. First, we conduct a qualitative data analysis and illustrate in which acoustic scene classes sound repetitions occur, which repetition rates are typical, and whether these sound repetitions relate to harmonic or transient signal components. Second, we propose and evaluate several fusion strategies to integrate information about the self-similarity of an audio signal in an ASC model based on a convolutional neural network (CNN) architecture.

1.1 Acoustic scene classification

For the last decade, the Detection and Classification of Acoustic Scenes and Events (DCASE) research community has been an important driver of research in the field of ASC [6]. Historically, ASC has emerged from Computational Auditory Scene Analysis (CASA) research, which studies how human listeners process complex environments by segregating and fusing auditory features into auditory objects [10]. Early research on ASC focused on combining audio signal processing algorithms such as Mel frequency cepstral coefficients (MFCC) with traditional machine learning algorithms such as hidden Markov models (HMMs) and Gaussian mixture models (GMMs) [5]. Similar to other fields of machine listening, ASC underwent a paradigm shift towards data-driven deep learning-based algorithms, in which both feature processing and classification are part of the model to be trained.

In a more recent review [11] written by the first author, common ASC algorithms are compared according to different stages of data preparation and data modeling. To this day, spectrogram-based audio signal representations, such as the Mel spectrogram, are most often applied and combined with logarithmic magnitude compression similar to the human auditory system [12–14]. Other feature representations applied in recent ASC works include MFCC [15], Wavelet transformation [16], amplitude modulation bank features [17], scalogram [18], as well as raw audio signals [19] for end-to-end learning.

ASC algorithms may include a source separation stage, where a given acoustic scene recording is decomposed into multiple audio streams, or signal enhancement, where some signal parts are enhanced while others are suppressed. For example, harmonic-percussive source separation (HPSS) [20] separates harmonic and transient signal components [13, 21] while the repeating pattern extraction technique (REPET) [22] separates foreground sounds from repetitive background sounds [23]. The per-channel energy normalization (PCEN) technique [24, 25] suppresses stationary background noise. Another core element of deep learning-based ASC algorithms is data augmentation to increase the amount and variability of training data. Common methods include mixup [14, 26], temporal crop [12], and SpecAugment [27].

In the last 3 years, different research trends can be identified in ASC research, driven in part by designated tasks in the annual DCASE challenges. Acoustic scenes are often defined by a set of characteristic sound events. Consequently, approaches for joint SED and ASC modeling have been proposed, e.g., using binaural audio features [28] or relational graph representations [29]. In multichannel recording setups, spatial features based on cross-correlation analysis have been shown to improve ASC performance [30]. Another main research focus is on device mismatch scenarios, where the different recording characteristics between the ASC model training and the inference stage cause a domain shift. Several domain adaptation methods have been proposed as countermeasures, ranging from normalization techniques [31] to data augmentation techniques such as impulse response augmentation [32]. Finally, another recent trend is to design efficient deep neural network architectures and model compression techniques [33, 34], which allow the deployment of ASC models on resource-constrained mobile devices and hearables.

1.2 Sound recurrence analysis

In several audio domains, methods for SRA have been developed. As shown in [35], both repetitive and stationary sound components can be identified using a binarized self-similarity matrix (SSMs) computed over acoustic features such as MFCCs. This has been shown to be beneficial for different environmental sound classification tasks such as ASC [35] and urban sound event classification [36] as well as crowd sound classification during basketball games [37]. As an alternative to self-similarity matrices, the real and imaginary parts of the complex modulation spectrogram have been used as a multichannel feature representation for an ASC model based on a convolutional recurrent neural network (CRNN) [38]. In music information retrieval (MIR), SSMs are a widely used feature representation, as they can be computed on various audio feature representations to unravel repetitions and homogeneous segments in different domains such as rhythm, timbre, or harmony [7]. Extending the concept of self-similarity, similarity between different music recordings has been analyzed using cross recurrence plots (CRP) to identify cover songs [39].

2 Dataset

In this study, we use the development set of the TAU Urban Acoustic Scenes 2019 dataset used in the DCASE 2019 Task 1A[1]. This dataset includes 40 hours of audio data recorded in 10 European cities at a sample rate of 48 kHz. The audio recordings were divided into 14,000 ten-second-long long stereo audio segments, which are evenly distributed throughout the 10 acoustic scene classes *airport*, *shopping mall*, *metro station*, *pedestrian street*, *public square*, *street with medium level of traffic*, *traveling by tram*, *bus*, and *metro* as well as *urban park*. We use the pre-defined dataset partitioning into training and test sets based on separated recording locations. Furthermore, we manually split the official training set into a training set ($\approx 70\%$) and a validation set ($\approx 30\%$) following the same criterion. In total, the ratio between training, validation, and the test set is approximately 5:2:3.

3 Audio processing

3.1 Audio feature representations

We process 10 s long audio signals at a sample rate of $f_s = 22.05\,\text{kHz}$. We use both channels of the stereo recordings in the baseline system (see Section 5.1.1) and a mono-down-mix thereof in all SRA-based models (see Section 5.1.2). The Mel spectrogram $X \in \mathbb{R}^{K \times N}$ is calculated using an FFT size of 1024 samples (46.4 ms), a hop size of 512 samples (23.2 ms), $K = 128$ Mel bands, and $N \in \mathbb{N}$ time frames. Logarithmic magnitude scaling is applied. In our experiments, we use the *librosa* Python library [40][2].

3.2 Self-similarity matrix

Before computing the self-similarity matrix, we perform background subtraction to enhance salient foreground sounds and suppress stationary background noise. We estimate the background spectrogram X_b by applying a median filter on X using a kernel of size 81 along the time axis (corresponding to 1.88 s) and 5 along the frequency axis and compute a spectrogram with enhanced foreground events as

$$X_\text{f} = X - \min(X, X_\text{b}). \tag{1}$$

The self-similarity matrix (SSM) $S \in \mathbb{R}^{N \times N}$ is calculated based on the cosine similarity between pairs of spectral frames, i.e., columns in X_f. We first create a modified spectrogram $\tilde{X}_\text{f}$ by normalizing all columns in X_f to unit L_2 norm and compute the SSM as

$$S = \tilde{X}_\text{f}^\top \tilde{X}_\text{f}. \tag{2}$$

As a final step, we enhance the path structures in S as proposed in [41]. Here, we first use diagonal smoothing with a filter length of 5 to enhance the path structures. In a second step, we apply a tempo difference adaptation to compensate for tempo variations of roughly -50 to 50% as some similar sound patterns may appear at different tempos. As an example, Fig. 1 illustrates (from top to bottom) the Mel spectrogram X, the Mel spectrogram after applying the background subtraction X_f, and the final self-similarity matrix S for an audio recording of the DCASE2019 Task1A development set, which is associated with the acoustic scene class *tram*.

3.3 Harmonic-percussive source separation (HPSS)

In this section, we briefly introduce harmonic-percussive source separation (HPSS). This method outputs two alternative signal representations for SRA, from which we compute SSM representations as introduced in Section 3.2. HPSS separates an audio signal into two streams with orthogonal spectral characteristics. On the one hand, the harmonic stream includes harmonic sounds with stable fundamental frequency that manifest themselves as horizontal line structures in a spectrogram. On the other hand, the percussive stream is characterized by transient sounds, which appear as vertical line structures in a spectrogram. We use the method proposed in [20] and further discussed in [7], which combines two median filtering operations in horizontal and

[1] https://dcase.community/challenge2019/task-acoustic-scene-classification

[2] Version 0.9.2

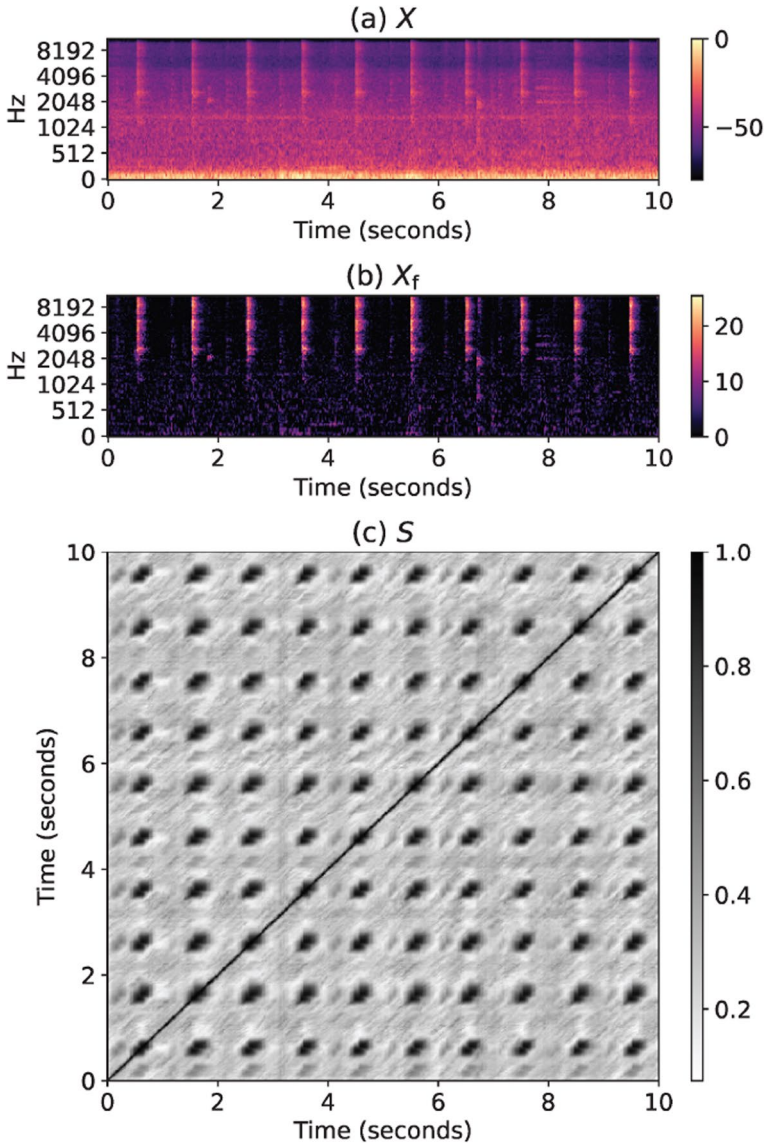
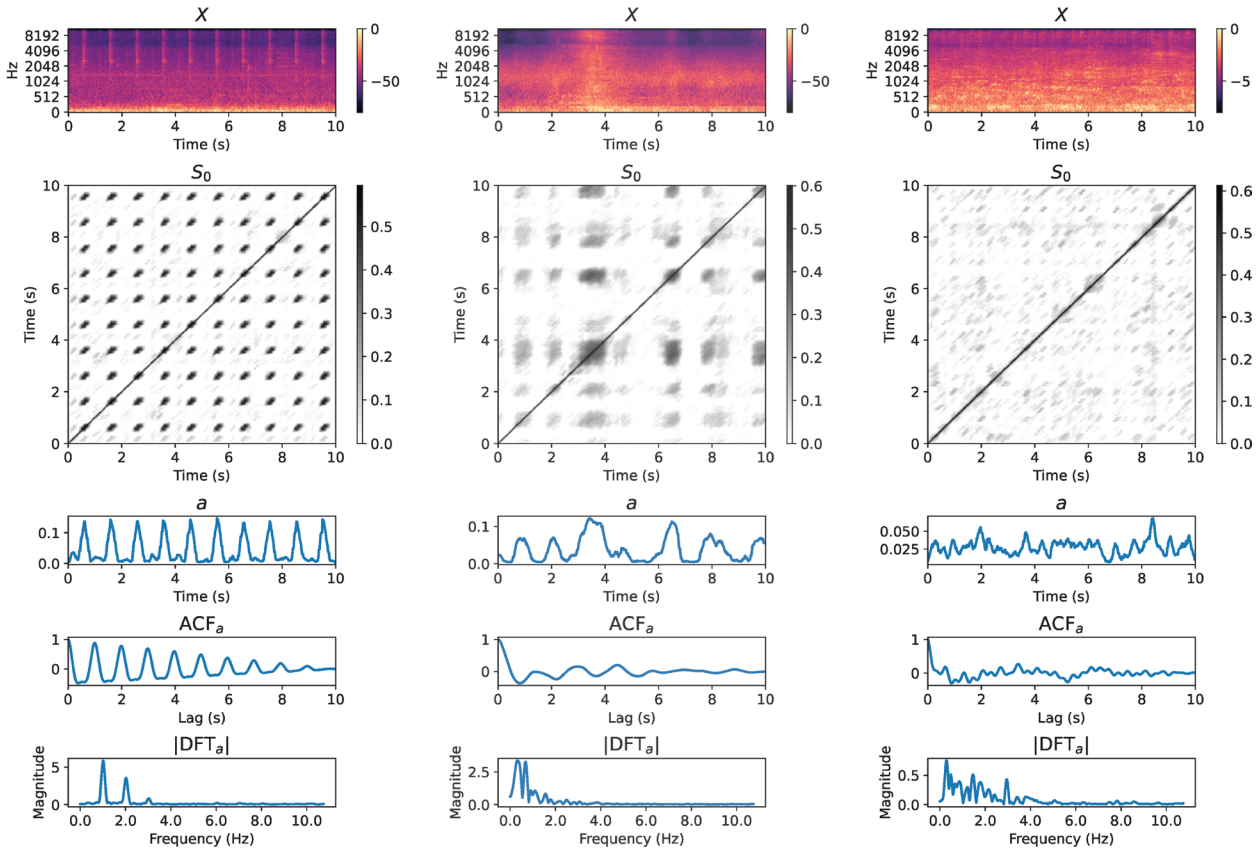


Fig. 1 Example audio recording from *tram* class shown as **a** Mel spectrogram X, **b** Mel spectrogram after applied background subtraction X_f, and **c** self-similarity matrix S

vertical directions, i.e., along time and frequency. From the filtered spectrograms, time-frequency masks are derived, which allow one to filter the original signal into a harmonic and percussive stream. In preliminary experiments, we found that suitable median filter sizes are 81 along time and 15 along frequency, respectively.

4 Sound recurrence analysis in acoustic scenes

In this section, we introduce three metrics to describe sound repetition patterns in acoustic scenes. First, the prevalence metric introduced in Section 4.1 describes how clearly a sound repetition pattern is pronounced. Second, the sound repetition rate introduced in Section 4.2 characterizes the number of sound repetitions in a given time interval. Finally, the temporal irregularity metric introduced in Section 4.3 describes how regularly the sound repetitions occur. Figure 2 illustrates the concepts discussed for three acoustic scenes with regular, irregular, or no sound repetitions. In our data analysis, we focus on two aspects. First, we initially apply a signal decomposition using HPSS (see Section 3.3) to investigate whether these repetitions manifest themselves in harmonic or transient sound components. Second, we study to what extent these repetitions are characteristic of particular acoustic scenes and therefore could provide important cues for an ASC algorithm (see Section 5).



(a) Regular Repetition (b) Irregular Repetition (c) No Repetition

Fig. 2 Three examples of acoustic scenes with **a** regular sound repetitions (beeping sounds in tram), **b** irregular sound repetitions (passing cars in traffic environment), and **c** no sound repetition (shopping mall ambience). For each acoustic scene, the figures (from top to bottom) show the Mel spectrogram X, the modified SSM S_0, the column-wise average a, its autocorrelation function ACF_a, and its magnitude spectrum $|\text{DFT}_a|$

4.1 Prevalence of sound repetition patterns

We define two metrics based on the SSM S to measure the prevalence of sound repetitions. In order to emphasize repetitive patterns, we first subtract the global mean of S and apply clipping to a lower bound of 0 as

$$S_0 := \max\left(0, S - \frac{1}{N^2}\sum_{n_1=1}^{N}\sum_{n_2=1}^{N} S(n_1, n_2)\right). \quad (3)$$

We observed that the global average over S_0 increases with a higher number and a longer duration of repeated sound events. Consequently, we derive a prevalence metric γ_{P} as

$$\gamma_{\text{P}} = \frac{1}{N^2}\sum_{n_1=1}^{N}\sum_{n_2=1}^{N} S_0(n_1, n_2). \quad (4)$$

For the three examples shown in Fig. 2, we observe a higher prevalence value of $\gamma_{\text{P}} = 0.041$ for the example with regular repetitions in Fig. 2a compared to $\gamma_{\text{P}} = 0.037$ and $\gamma_{\text{P}} = 0.026$ for the examples with irregular repetitions in Fig. 2b and no repetitions in Fig. 2c, respectively.

4.2 Sound repetition rate

The second characteristic we investigate is the repetition rate of a sound sequence, which can vary greatly depending on the type of sound source. For example, after listening to several *traffic* class examples in our dataset, we observed that vehicles pass at a rate of 2 s to 5 s, which corresponds to repetition rates of 0.2 Hz to 0.5 Hz. In contrast, noises from running machines in a factory scenario often exhibit higher sound repetition rates in the range of 1 Hz. Since acoustic scenes are characterized by a unique set of common sound sources, we expect the

sound repetition rate to be a meaningful cue to inform ASC algorithms.

As a basis for a periodicity analysis, we calculate the column-wise average $a \in \mathbb{R}^N$ over the modified SSM S_0 as

$$a(n_1) = \frac{1}{N} \sum_{n_2=0}^{N-1} S_0(n_1, n_2) \tag{5}$$

for $n_1 \in [0, N-1]$. We estimate the predominant repetition period T_R using two common methods: the autocorrelation function (ACF) and the discrete Fourier transform (DFT). Based on the applied feature extraction parameters (see Section 3.1), the temporal resolution of a is $\Delta t \approx 23.2$ ms, which corresponds to a sampling frequency of $f_\mathrm{s} = 1/\Delta t = 43.1$ Hz.

The ACF over a is defined at time lag l as

$$\mathrm{ACF_a(l)} := \sum_{\mathrm{m}=0}^{\mathrm{N}-1} \mathrm{a(m)} \cdot \mathrm{a(m-l)}, \tag{6}$$

All peaks at non-zero lag positions indicate high local self-correlation and hence a possible signal repetition. Therefore, we estimate T_R in x as the lag of the largest local peak for $l > 0$.

While the ACF characterizes the self-similarity properties of a signal in the time domain, the DFT decomposes a signal into multiple periodic signals defined by different frequency, magnitude, and phase values. Similarly to the ACF, we estimate the repetition rate from the frequency position of the highest peak in the DFT magnitude spectrum. Before computing the DFT over a, we apply zero-padding to increase the signal length by factor 5 and Hann window to reduce spectral leakage.

Finally, when choosing whether to use $\mathrm{ACF_a}$ or $\mathrm{DFT_a}$, we generally prioritize $\mathrm{DFT_a}$ to estimate the sound repetition rate T_R unless 1) the difference between two highest detected peaks of $\mathrm{ACF_a}$ exceeds an empirical threshold of 0.1, or 2) T_R estimated with $\mathrm{DFT_a}$ is larger than 10 s, which occasionally happens due to zero padding.

In the first acoustic scene illustrated in Fig. 2a, the $\mathrm{ACF_a}$ shows equidistant peaks, which gradually decay with increasing lag values. The lag position of the highest peak for $l > 0$ is around 1 s, leading to a repetition rate of around 1 Hz. This result is confirmed by the largest peak in the $\mathrm{DFT_a}$ magnitude spectrum around 1 Hz. The remaining peaks can be interpreted as sub-harmonics at multiples of the actual repetition period. For acoustic scenes without sound repetitions, as shown in Fig. 2c, both $\mathrm{ACF_a}$ and $\mathrm{DFT_a}$ do not exhibit a clear peak structure and therefore do not allow for reliably estimating T_R. In the case of irregular repetitions (Fig. 2b), the $\mathrm{ACF_a}$ function has a more ambiguous peak structure, since multiple repetition rates can be observed.

4.3 Temporal irregularity

As a third characteristic, we measure the temporal irregularity of the sound repetitions. If sounds are emitted, for instance, from evenly running machines, the sound repetition rate is approximately constant. If, in contrast, sounds are emitted from multiple independent sound sources, sound repetitions are often irregular. Examples of such irregular sound repetitions are vehicle sounds from passing cars in a traffic environment or bird calls in nature environments. The temporal irregularity of the sound repetitions is well reflected in the non-sparsity in $\mathrm{DFT_a}$. Regular sound repetitions result in sparse peaks in $\mathrm{DFT_a}$, whereas irregular repetitions lead to a more noisy and less sparse structure in $\mathrm{DFT_a}$.

Following [42], we use a non-sparseness measure to quantify temporal irregularity in $\mathrm{DFT_a}$ as

$$\gamma_\mathrm{TI} = 1 - \frac{\sqrt{N} - \frac{\|\mathrm{DFT_a}\|_1}{\|\mathrm{DFT_a}\|_2}}{\sqrt{N} - 1}, \tag{7}$$

where $\|\mathrm{DFT_a}\|_1$ and $\|\mathrm{DFT_a}\|_2$ denote the ℓ_1 norm and ℓ_2 norm over $\mathrm{DFT_a}$. With $\gamma_\mathrm{TI} \in [0, 1]$, lower values of γ_TI indicate more regular sound repetitions, whereas higher values of γ_TI indicate more random sound repetitions. This intuition is confirmed for the three examples of acoustic scenes illustrated in Fig. 2. For the two examples of regular and irregular repetitions in Figs. 2a and b, we observe $\gamma_\mathrm{TI} = 0.251$ and $\gamma_\mathrm{TI} = 0.269$, respectively, while the third example without sound repetitions shown in Fig. 2c leads to a higher value of $\gamma_\mathrm{TI} = 0.454$.

4.4 Selecting acoustic scenes with sound repetitions

Naturally, the sound repetition rate introduced in Section 4.2 is only a meaningful concept in acoustic scenes with actual sound repetitions. We therefore combine three criteria to select those audio recordings for SRA analysis, which either (i) simultaneously show a high prevalence value γ_P and a low irregularity γ_TI, (ii) show a very high prevalence value, or (iii) show a very low irregularity value. We empirically found the following threshold values by manually observing around 300 example recordings from our dataset:

$$(\gamma_\mathrm{P} > 0.03 \wedge \gamma_\mathrm{TI} < 0.4) \vee (\gamma_\mathrm{P} > 0.04) \vee (\gamma_\mathrm{TI} < 0.3). \tag{8}$$

These selection criteria are applied only for the sound recurrence analyses discussed in Section 4. For the ASC experiments that will be discussed in Section 5, the entire dataset is used. Considering again the three

examples from Fig. 2, we would include for the sound recurrence analyses the first example with regular repetitions, which fulfills $\gamma_{\mathrm{P}} > 0.04$ as well as the second example with irregular repetitions, which fulfills $\gamma_{\mathrm{P}} > 0.03 \wedge \gamma_{\mathrm{TI}} < 0.4$. Figure 3 illustrates the joint distribution between the metrics γ_{P} and γ_{TI} for each acoustic scene class in our dataset. Samples shown in blue fulfill the combined selection criterion given in (8).

4.5 Results and discussions

Sections 4.5.1 and 4.5.2 summarize the main findings from a statistical analysis of the class-wise distributions of the sound recurrence metrics γ_{P}, γ_{TI}, and T_{R}, computed over the harmonic and percussive signal streams, respectively. After confirming the non-normality of the class-wise distributions using the Shapiro-Wilk test, the Kruskal-Wallis H-test and Dunn's post-hoc test using a Bonferonni correction were used to identify groups of acoustic scene classes with significantly different group medians.

4.5.1 *Transient sound repetitions*

Figure 4 summarizes the characteristics of repetitive transient sound patterns for different acoustic scenes classes. We observe similar distributions in γ_{P} for the classes *bus* and *metro*. Similarly, the classes *shopping mall* and *street pedestrian* form a second group with similar group medians. Looking at γ_{TI}, we observe that the classes *street traffic* and *park* form the group with the lowest prevalence of transient sound repetitions while the class *tram* has the highest prevalence, which is presumably caused by rattling driving noises. Despite the low prevalence, transient sound repetitions in *street traffic* recordings have the highest temporal

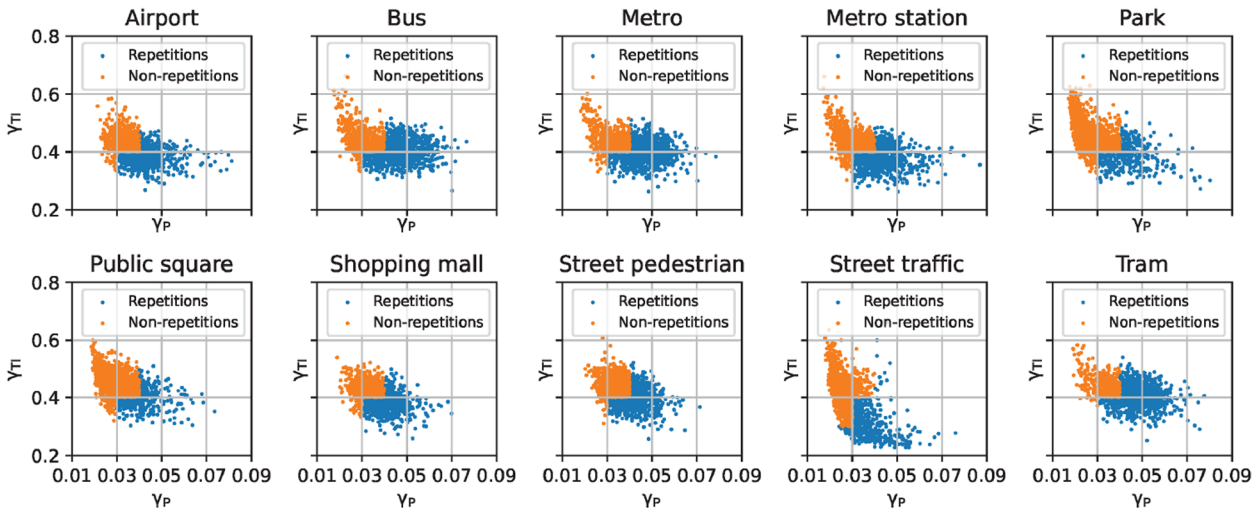
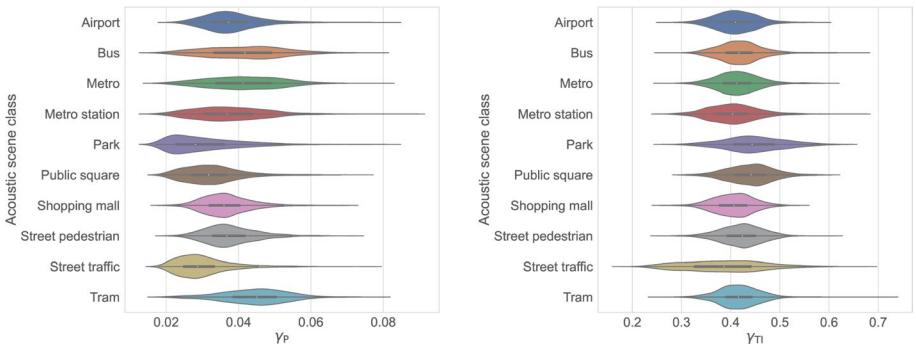


Fig. 3 Joint distribution between the prevalence of sound repetition patterns γ_{P} and temporal irregularity γ_{TI} illustrated for each acoustic scene class. Audio clips with sound repetitions, which fulfill the selection criterion given in (8), are shown in blue



(a) Prevalence (b) Temporal Irregularity

Fig. 4 Statistical summary over **a** the prevalence of transient sound repetition patterns (γ_{P}) and **b** their temporal irregularity (γ_{TI}). The violin plots are a combined illustration of the distribution median (white dot), interquartile range (black box), and overall kernel density plot (colored area)

regularity (lowest median γ_P values). Our interpretation is that the sounds emitted from passing vehicles are quite different among themselves, but occur at a constant rate.

As a third attribute, Fig. 5 provides histograms of the sound repetition period for individual acoustic scene classes. Audio recordings without clear sound repetitions are excluded according to criterion (8). The number of acoustic scenes considered for the SRA is given in parentheses. We make the following observations. First, the acoustic scenes *park* and *public square* show fewer sound repetitions compared to the other classes. Second, histograms generally indicate similar distributions between classes, with most of the observed repetition periods being around 2 s. As an exception, the sound repetition periods of passing vehicles in *street traffic* scenes are more evenly distributed and exhibit values up to 5 s. This can be explained by different distances between consecutive vehicles.

4.5.2 Harmonic sound repetitions

In this section, we focus on repetitions of harmonic sounds, which are characterized by stable frequency components over time. In our dataset, examples of such sounds are car horns in the *traffic* scene, beeping sounds in public transportation announcements, or bird calls in the *park* scene. Figure 6 illustrates the class-wise distributions of the prevalence metric γ_P and the temporal irregularity metric γ_{TI}. Figure 6a shows that in general, repetitions of harmonic sounds are less frequent than

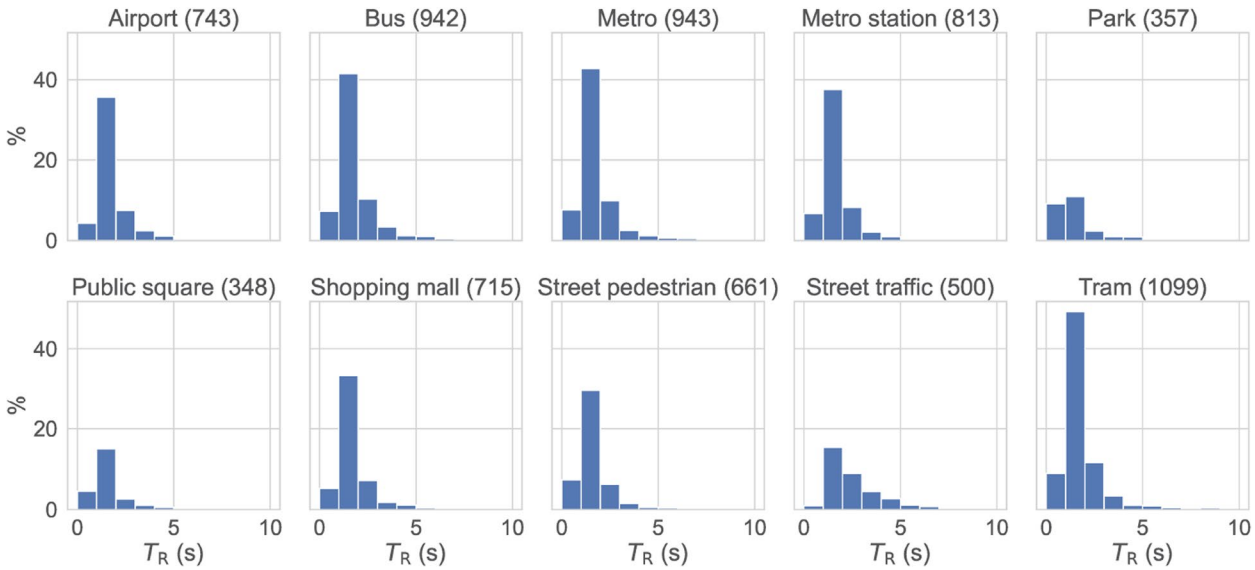
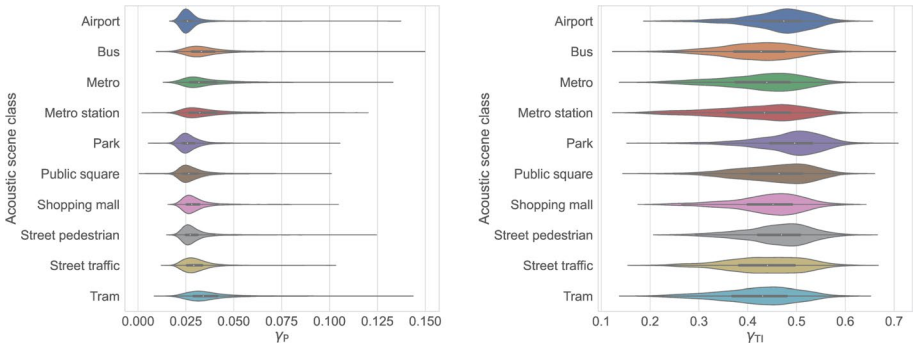


Fig. 5 Histograms of the estimated repetition periods T_R of transient sound repetition patterns for different acoustic scene classes. Absolute number of recordings per class considered for SRA is given in brackets



(a) Prevalence (b) Temporal Irregularity

Fig. 6 Statistical summary over **a** the prevalence of harmonic-like repetition patterns (γ_P) and **b** their temporal irregularity (γ_{TI})

repetitions of transient sounds, which is indicated by average values of γ_{p} around 0.025 and 0.04, respectively. Furthermore, we observe a higher number of repetitive harmonic sounds in acoustic scene classes related to public transportation (*bus*, *metro*, *metro station*, and *tram*). At the same time, Fig. 6b confirms that in particular in nature environments (*park*), harmonic sound repetitions, such as from bird calls, tend to be more irregular. As can be seen in Fig. 4b, a similar observation can be made for transient sound repetitions.

5 Integration of sound recurrence analysis for acoustic scene classification

Qualitative data analysis presented in Section 4 indicates that sound recurrences are a characteristic property in several types of acoustic scenes. Based on this observation, we investigate different approaches to integrate SRA features into a deep neural network-based ASC model. This approach complements qualitative data analysis and allows us to investigate to what extent the recognition of class-specific sound repetition patterns can contribute to improve acoustic scene classification. In the experiments described in this section, our focus is not on surpassing state-of-the-art performance on the ASC data set used but on a quantitative performance comparison between different input feature combinations.

5.1 Model architectures

5.1.1 Baseline model

Different frequency bands contain discriminatory information for ASC [43]. We use a ResNet architecture proposed in [44] (denoted as `M-SPEC`), which uses two input branches with multiple residual layers that separately process the low- and high-frequency parts of a Mel spectrogram. The authors hypothesize that a separate processing of low and high frequencies allows to learn separate feature representation for improved ASC. While this hypothesis has not been validated directly, the proposed model architecture achieved second place in the DCASE 2019 task 1B "Acoustic Scene Classification with mismatched recording devices." The feature maps at the output of both input branches are concatenated. Then, convolutional layers of shape 1×1 are used such that the resulting feature map has 10 output channels. A final global average pooling operation yields classification scores for all 10 classes. The architecture of this baseline model is illustrated in Fig. 7.

In the residual layers, strided convolutions with a kernel size of 3×3 are performed in such a way that feature maps are only down-sampled across time, not across frequency. The model uses a multichannel input feature that includes the Mel spectrogram and its first two derivatives. The input branches of the network process the upper and lower 64 Mel bands of the input feature, which corresponds to a cutoff frequency of around 2150 Hz. For training the `M-SPEC` model, we use Mixup data augmentation [45] with $\alpha = 0.6$ and stochastic gradient descent (SGD) optimization with a momentum of 0.9 and a batch size of 32. Following [12, 46], we apply a cosine annealing-based learning rate scheduler with a maximum learning rate of 0.001 and use early stopping on the validation set.

5.1.2 Model based on sound recurrence analysis

The `M-SPEC` model introduced in Section 5.1.1 processes spectrogram features and therefore can learn to recognize acoustic scenes based on their timbral properties. Given the scope of this work, we investigate another model architecture (denoted as `M-SRA`), which processes only SRA features such as the self-similarity matrix obtained from the Mel spectrogram (compare Section 3.2). By excluding spectrogram features, this model allows us to investigate to what extent different acoustic scenes can be classified only based on class-specific sound repetition patterns. Unlike the time and frequency axes of a spectrogram, an SSM has two time axes with a similar meaning. Since images similarly have two semantically related dimensions, we adapt the ResNet-101 [47]

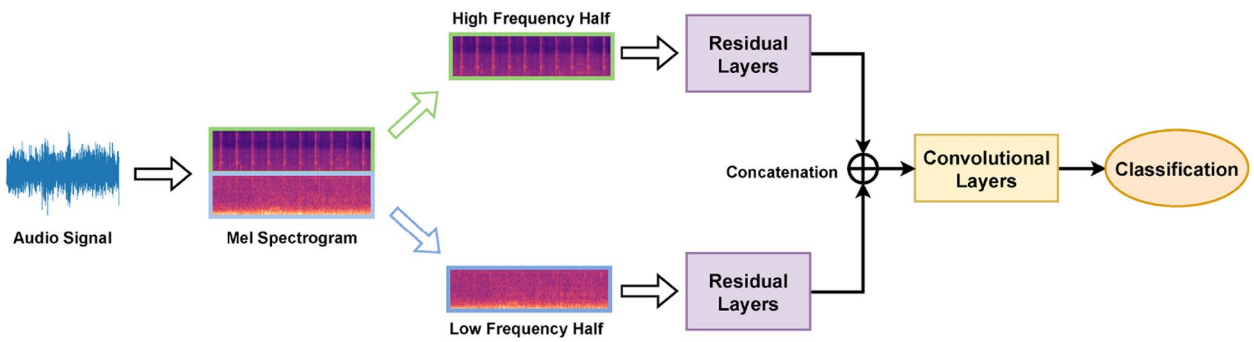


Fig. 7 Architecture of the baseline model [44] (`M-SPEC`)

architecture from the computer vision domain and apply it to SSM based input features. The general architecture of the model is illustrated in Fig. 8. Before computing the SSM, we down-mix the stereo audio recordings to single-channel audio recordings. Then, we obtain a two-channel feature with separate SSMs for the harmonic and percussive stream if HPSS is applied prior to the SSM computation and a single-channel feature if not. We use the same training hyperparameters as for the M-SPEC model introduced in Section 5.1.1.

5.1.3 Hybrid models

In the hybrid network architectures introduced in this section, we compare different fusion strategies to combine spectrogram features and SRA features for ASC. All models are trained using the same methodology as described in Section 5.1.1.

As a first strategy, we apply an intermediate fusion approach as illustrated in Fig. 9. In this model, both the spectrogram features and the SRA features are processed by individual network branches before the resulting feature maps are fused and processed by the final layers to yield a classification result. We compare two fusion strategies: a concatenation of feature maps from both input branches based on combining 1×1 convolution and global average pooling (denoted as M-HYB-CON) and a weighted element-wise summation of both feature maps (denoted as M-HYB-SUM). For the latter, the weighted summation is defined as

$$z = \lambda z_1 + (1 - \lambda) z_2 \tag{9}$$

given two flattened feature maps $z_1 \in \mathbb{R}^Z$ and $z_2 \in \mathbb{R}^Z$ as well as a weighting factor $\lambda \in \mathbb{R}$, which is implemented as a trainable parameter in the network.

As a second strategy, we combine the two models introduced Sections 5.1.1 and 5.1.2 into a classifier ensemble, which benefits from the individual strengths of each model. As long as the individual models are diverse and independent, we expect the prediction error to decrease using the ensemble model. During inference, we average the probability estimates for each acoustic scene class across both classifiers. This model is denoted as M-HYB-ENS.

As a third strategy, we implement an additional ensemble model (denoted as M-ENS-GLOB) that combines the predictions of the single-branch models M-SPEC and M-SRA as well as the dual-branch model M-HYB-CON. This model can be considered as an upper performance limit, as it combines the individual strengths of these three models. However, realistically, this model would

Fig. 8 Architecture of the M-SRA model

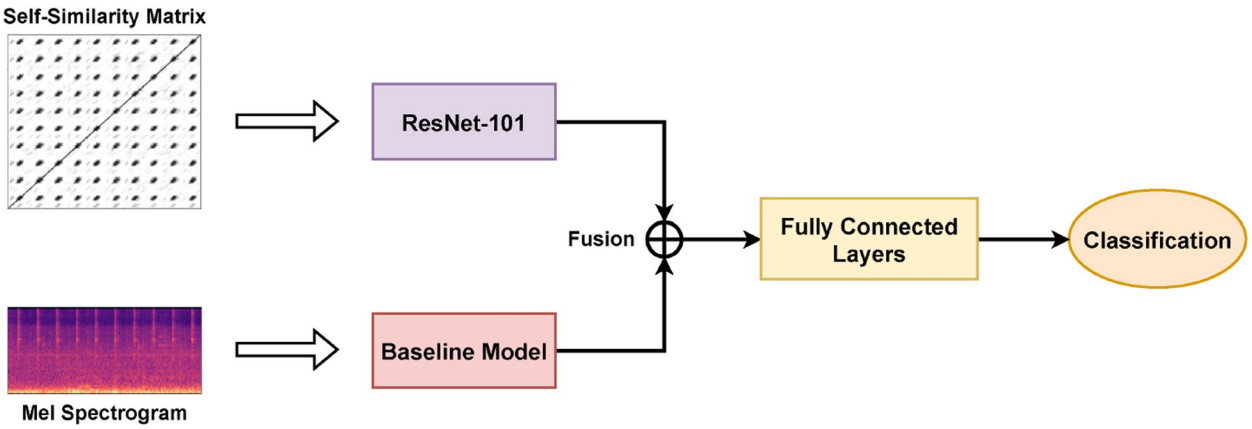


Fig. 9 Model architecture overview of M-HYB-CON and M-HYB-SUM models, which combine spectrogram features and SRA features using different intermediate fusion approaches

Table 1 Summary of ASC models categorized by model name, section reference, input feature(s) (MS: Mel spectrogram, SSM: self-similarity matrix), and fusion strategy

Model	Section	Input feature	Fusion strategy
M-SPEC	5.1.1	MS (frequency split)	-
M-SRA	5.1.2	SSM	-
M-HYB-CON	5.1.3	MS + SSM	IF (concate)
M-HYB-SUM	5.1.3	MS + SSM	IF (sum)
M-HYB-ENS	5.1.3		ENS(M-SPEC, M-SRA)
M-ENS-GLOB	5.1.3		ENS(M-SPEC, M-SRA, M-HYB-CON)

not be used in practice, as the spectrogram features and the SRA features are processed redundantly in different models. We summarize all models introduced in Section 5.1 in Table 1.

5.2 Results and discussions

Table 2 provides both the class-wise and average accuracy scores for ASC obtained by the single-branch models M-SPEC and M-SRA, the dual-branch models M-HYB-CON, M-HYB-SUM, and M-HYB-ENS as well as the global ensemble model M-ENS-GLOB.

5.2.1 Spectrogram features

The M-SPEC model reaches 67.5% average accuracy on the test set. The confusion matrix illustrated in Fig. 10a shows that *park* and *street traffic* are classified most reliably with accuracy values above 80%. At the same time, we observe several misclassifications due to explainable similarities between acoustic scene classes. Around 16% of the test instances are confused between *public square* and *street pedestrian*, which both share sounds from human actions and conversations. As another example, we observe several confusions of around 20% between recordings of the *metro* and *tram* classes, which were recorded in public transportation with similar indoor acoustic characteristics.

5.2.2 SRA features

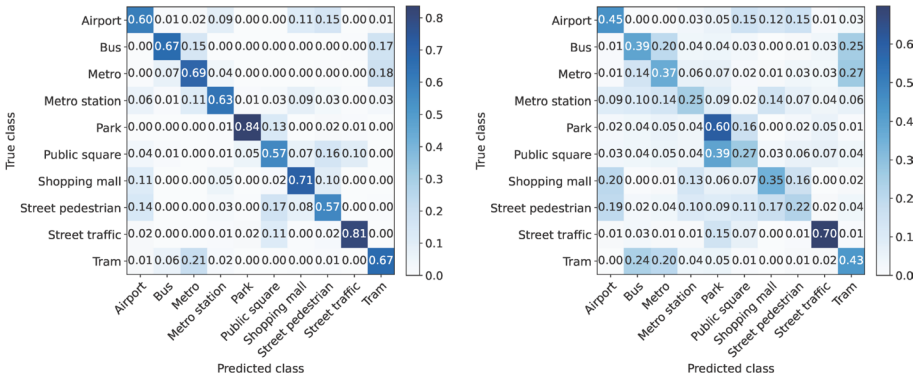
During an initial experiment, we observe a slight improvement in the classification accuracy from 39.3 to 39.9% when both background subtraction (BS, see Section 3.2) and HPSS (see Section 4) are used as preprocessing steps. Figure 10b shows the confusion matrix obtained for this M-SRA model configuration. Therefore, for the remainder of this work, we used HPSS consistently for all model configurations that involve SRA features.

Given that this model only takes SRA features and no spectrogram features as input, it is notable that it still shows a good classification performance of 60% for *park* and 70% for *street traffic*. These results confirm that both acoustic scene classes exhibit class-specific sound repetition patterns. The results of the qualitative analysis discussed in Section 4.5 offer a possible explanation: Here, we find that *street traffic* recordings show transient sound repetitions from passing vehicles, which are regular but have a lower repetition rate. On the other hand, the *park* recordings show fewer regular repetition structures for harmonic and transient sounds. On the contrary, *metro*

Table 2 Class-wise and averaged accuracy values for single-branch, dual-branch, and global ASC models listed in Table 1

Scene class	Model					
	Single-branch		Dual-branch			Global
	M-SPEC	M-SRA	M-HYB-CON	M-HYB-SUM	M-HYB-ENS	M-ENS-GLOB
Airport	0.60	0.45	**0.75**$^{\diamond}$	0.58	0.61	**0.75**
Bus	0.67	0.39	0.70	**0.76**$^{\diamond}$	0.69	0.73
Metro	**0.69**	0.37	0.61	0.67	0.67	0.67
Metro station	0.63	0.25	0.69$^{\diamond}$	0.67	0.64	**0.72***
Park	0.84	0.60	0.82	0.77	0.84	**0.86***
Public square	0.57	0.27	0.57	0.42	0.53	**0.59***
Shopping mall	**0.71**	0.35	0.70	0.70	**0.71**	0.70
Street pedestrian	0.57	0.22	0.55	0.56	0.52	**0.58***
Street traffic	0.81	0.70	0.78	0.79	**0.85**$^{\diamond}$	0.83
Tram	0.67	0.43	0.75	**0.77**$^{\diamond}$	0.67	0.75
Average accuracy	0.675	0.399	0.690	0.671	0.669	**0.715**

Bold font indicates highest class-wise score across all models. Classes, where the dual-branch models outperform single-branch models, are marked with $^{\diamond}$. Classes, where the global ensemble model M-ENS-GLOB achieved a higher accuracy than all other models, are marked with *



(a) M-SPEC (b) M-SRA

Fig. 10 Confusion matrices of the M-SPEC model (**a**) and the M-SRA model (**b**)

station, *public square*, and *street pedestrian* show poor classification results below 25%, which indicates that these classes do not have distinctive sound repetition patterns or that typical long-term patterns are not captured due to the limited audio clip duration of 10 s.

In addition to the class-wise performance, we observe three groups of acoustic scenes with prominent confusions: (1) *airport*, *shopping mall*, and *street pedestrian*, (2) *tram*, *bus*, and *metro* as well as (3) *public square* and *park*. The acoustic scene classes in each group share very similar distributions of the three sound recurrence metrics introduced in Section 4, namely the prevalence γ_{P}, the temporal irregularity γ_{TI}, as well as the repetition period T_{R}, across the transient and harmonic sounds as shown in Figs. 4, 5, and 6.

5.2.3 Hybrid ASC models

After discussing the performance of the ASC models, which exclusively use spectrogram or SRA features, we now focus on the performance of the hybrid models introduced in Section 5.1.3, which combine both types of features. The accuracy scores of the single-branch models M-SPEC ($A = 0.675$) and M-SRA ($A = 0.4$) serve as a reference to compare the performance of the hybrid models.

We first focus on the performance of the dual-branch models. For each of the classes *airport*, *bus*, *street traffic*, and *tram*, one of the dual-branch models outperforms the M-SPEC model (marked by $^{\diamond}$), which demonstrates the potential to complement spectrogram-based features with SRA-based features for ASC. However, none of the models shows a significant global improvement in all classes compared to the M-SPEC model. We suspect that sound repetition patterns provide only discriminative cues for a subset of the investigated acoustic scene classes, as also confirmed by the results discussed in Section 4.5. Among the aggregation strategies, the intermediate fusion approach in the M-HYB-CON model, where intermediate feature maps are concatenated, leads to the highest accuracy value $A = 0.69$, which represents an improvement of 1.5% compared to the M-SPEC model. As expected, the global model M-ENS-GLOB achieves the highest average accuracy of $A = 0.715$ and outperforms the other models for the classes *metro station*, *park*, *public square*, and *street pedestrian* as indicated by *. However, the differences in accuracy between the M-SPEC model and the M-HYB-CON as well as M-ENS-GLOB models are not statistically significant as confirmed by one-way ANOVAs (distribution normality was test using the Shapiro-Wilk test). Furthermore, for a practical application, such a global ensemble model is not very efficient, since it requires the individual results of three different classification models to arrive at a decision.

6 Conclusion

In this paper, we examined sound repetitions in various types of acoustic scenes. Specifically, we used the self-similarity matrix as a suitable signal representation for sound recurrence analysis (SRA) in audio recordings. Through an initial qualitative data analysis, we identified characteristic repetition patterns for harmonic and transient sound elements for different acoustic scene classes. In the second part of our work, we introduced and evaluated multiple fusion strategies designed to integrate information about the self-similarity of an audio signal into an acoustic scene classification model based

on a convolutional neural network architecture. Our results demonstrate that hybrid network architectures, which combine spectrogram features and SRA features, can lead to an increase in ASC accuracy for those acoustic scene classes which exhibit characteristic sound repetition patterns. However, when looking at the overall classification performance averaged across all acoustic scene classes, we found that the accuracy gains of 0.015 (hybrid model `M-HYB-CON`) and 0.04 (global ensemble model `M-ENS-GLOB`) are not statistically significant. This stands in contrast to previous work [48], where the performance of an ASC algorithm based on support vector machine (SVM) classifier and MFCC timbre features improved when features derived from a recurrent quantification analysis were integrated. However, similar to our findings, features derived from a spectrogram filtering, which was informed by self-similarity, only improved classification accuracy when being used within classifier ensembles in the ASC system presented in [23].

Our analyses further provided a deeper insight into the characteristics of sound repetition patterns in natural acoustic scenes. In particular, we found that sound repetitions are particularly present in vehicle-related scene classes such as *bus*, *metro*, and *tram* and generally show up in transient rather than harmonic sounds. The classes with the most regular sound repetitions are *street traffic* for transient sound repetitions as well as *bus*, *metro*, and *tram* for harmonic sound repetitions. In general, most repetition periods are around 2 s. Therefore, while we have only examined signals with a duration of 10s in this work, characteristic sound repetitions should therefore also be recognizable in shorter signals with a duration of at least 4s.

It should be noted that the highest average accuracy of $A = 0.715$ (see Table 2) is clearly below the best accuracy reported in [44] for the baseline model of $A = 0.81$. We see two possible reasons for this. On the one hand, we did not use the random crop data augmentation as proposed in [44], since it significantly increases the computational cost for computing self-similarity matrices on the fly during training. On the other hand, we only used 70% of the development set as training data as described in Section 2. Although we did not achieve state-of-the-art results for the applied acoustic scene classification (ASC) dataset, we believe that our results nevertheless provided valuable insights into the importance of sound repetition patterns for different acoustic scenes. As a potential limitation of the SSM-based SRA approach, many perceptually relevant sound repetitions in natural acoustic scenes may not manifest as exact repetition patterns and might require a more fuzzy detection approach. Furthermore, we believe that the modulation spectrogram could be a possible alternative signal representation for SRA.

Abbreviations

AFC	Autocorrelation function
ASC	Acoustic scene classification
CASA	Computational auditory scene analysis
CNN	Convolutional neural network
CRNN	Convolutional recurrent neural network
CRP	Cross recurrence plot
DCASE	Detection and Classification of Acoustic Scenes and Events
DFT	Discrete Fourier transform
FFT	Fast Fourier transform
GMM	Gaussian mixture model
HPSS	Harmonic-percussive source separation
HMM	Hidden Markov model
MFCC	Mel frequency cepstral coefficient
MIR	Music information retrieval
PCEN	Per-channel energy normalization
REPET	Repeating pattern extraction technique
SED	Sound event detection
SRA	Sound recurrence analysis
SSM	Self-similarity matrix
SVM	Support vector machine

Acknowledgements
Not applicable.

Authors' contributions
BS and JA supervised the research that led to this publication. Data analysis and classification experiments were conceptualized by JA and ZL and implemented by ZL. The final manuscript was written by JA with additional contributions from ZL and BS. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. JA was supported by the German Research Foundation (DFG AB 675/2-2, Grant No. 350953655) and the Federal Ministry of Education and Research in Germany (BMBF) within the project news-polygraph (funding code 03RU2U151D).

Data availability
In this study, we used the publicly available development set of the TAU Urban Acoustic Scenes 2019 dataset used in the DCASE 2019 Task 1A (https://dcase.community/challenge2019/task-acoustic-scene-classification#subtask-a). The code used for this study is available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 26 April 2024 Accepted: 18 December 2024
Published online: 14 January 2025

References

1. ISO 12913-1:2014, Acoustics – Soundscape – Part 1: Definition and conceptual framework. Standard, International Organization for Standardization (2014). https://www.iso.org/standard/52161.html
2. B.C. Pijanowski, L.J. Villanueva-Rivera, S.L. Dumyahn, A. Farina, B.L. Krause, B.M. Napoletano, S.H. Gage, N. Pieretti, Soundscape ecology: The science of sound in the landscape. BioScience **61**(3), 203–216 (2011). https://doi.org/10.1525/bio.2011.61.3.6
3. B. Schulte-Fortkamp, A. Fiebig, J. Sisneros, A. Popper, R. Fay (eds.), *Soundscapes: Humans and Their Acoustic Environment*, 1st edn. (Springer International Publishing, Berlin/Heidelberg, 2023)
4. A. Mesaros, T. Heittola, T. Virtanen, M.D. Plumbley, Sound event detection: A tutorial. IEEE Signal Process. Mag. **38**(1), 67–83 (2021)
5. D. Barchiesi, D. Giannoulis, D. Stowell, M.D. Plumbley, Acoustic scene classification: Classifying environments from the sounds they produce. IEEE

Signal Process. Mag. **32**(3), 16–34 (2015). https://doi.org/10.1109/MSP.2014.2326181
6. T. Virtanen, M.D. Plumbley, D.E. Ellis, *Computational Analysis of Sound Scenes and Events*, 1st edn. (Springer International Publishing, Cham, 2018)
7. M. Müller, *Fundamentals of Music Processing Using Python and Jupyter Notebooks*, 2nd edn. (Springer, Cham, 2021). https://doi.org/10.1007/978-3-030-69808-9
8. M.N. Geffen, J. Gervain, J.F. Werker, M.O. Magnasco, Auditory perception of self-similarity in water sounds. Front. Integr. Neurosci. **5**, 1–11 (2011). https://doi.org/10.3389/fnint.2011.00015
9. A. Jati, A. Nadarajan, R. Peri, K. Mundnich, T. Feng, B. Girault, S. Narayanan, Temporal dynamics of workplace acoustic scenes: Egocentric analysis and prediction. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 756–769 (2021). https://doi.org/10.1109/TASLP.2021.3050265
10. D. Wang, G.L. Brown (eds.), *Computational auditory scene analysis: Principles, algorithms, and applications*, 1st edn. (Wiley, Hoboken, 2006)
11. J. Abeßer, A review of deep learning based methods for acoustic scene classification. Appl. Sci. **10**(6) (2020). https://doi.org/10.3390/app10062020
12. M.D. McDonnell, W. Gao, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths (Institute for Electrical and Electronics Engineers (IEEE), New York City, United States, 2020), pp. 141–145
13. Y. Han, J. Park, K. Lee, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*. Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification (Munich, 2017), pp. 46–50
14. K. Koutini, H. Eghbal-zadeh, G. Widmer, in Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE). Receptive-field-regularized cnn variants for acoustic scene classification (New York, NY, USA, 2019), pp. 124–128
15. Y. Li, X. Li, Y. Zhang, W. Wang, M. Liu, X. Feng, in *Proceedings of the 2018 International Conference on Audio, Language and Image Processing (ICALIP)*. Acoustic scene classification using deep audio feature and BLSTM network (Shanghai, 2018), pp. 371–374
16. K. Qian, Z. Ren, V. Pandit, Z. Yang, Z. Zhang, B. Schuller, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*. Wavelets revisited for the classification of acoustic scenes (Munich, 2017), pp. 108–112
17. N. Moritz, J. Schröder, S. Goetze, J. Anemüller, B. Kollmeier, Acoustic scene classification using time-delay neural networks and amplitude modulation filter bank features. Complexity **12**, 13 (2016)
18. H. Chen, P. Zhang, H. Bai, Q. Yuan, X. Bao, Y. Yan, in *Proceedings of the Interspeech Conference*. Deep convolutional neural network with scalogram for audio scene modeling (Hyderabad, 2018), pp. 3304–3308
19. D. Fedorishin, N. Sankaran, D.D. Mohan, J. Birgiolas, P. Schneider, S. Setlur, V. Govindaraju, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*. Waveforms and spectrograms: Enhancing acoustic scene classification using multimodal feature fusion (2021), pp. 216–220
20. D. FitzGerald, in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*. Harmonic/percussive separation using median filtering (Graz, 2010), pp. 246–253
21. H. Seo, J. Park, Y. Park, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA*. Acoustic scene classification using various pre-processed features and convolutional neural networks (New York, 2019), pp. 25–26
22. Z. Rafii, B. Pardo, Repeating pattern extraction technique (repet): A simple method for music/voice separation. IEEE Trans. Audio Speech Lang. Process. **21**(1), 73–84 (2013). https://doi.org/10.1109/TASL.2012.2213249
23. T. Nguyen, F. Pernkopf, in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*. Acoustic scene classification using a convolutional neural network ensemble and nearest neighbor filters (Surrey, 2018). https://dcase.community/workshop2022/proceedings
24. Y. Wang, P. Getreuer, T. Hughes, R.F. Lyon, R.A. Saurous, in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Trainable frontend for robust and far-field keyword spotting (Institute for Electrical and Electronics Engineers (IEEE), New York City, United States, 2017), pp. 5670–5674
25. V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, J.P. Bello, Per-channel energy normalization: Why and how. IEEE Signal Process. Lett. **26**(1), 39–43 (2019). https://doi.org/10.1109/LSP.2018.2878620
26. H. Wang, Y. Zou, D. Chong, in Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE). Acoustic scene classification with spectrogram processing strategies (Tokyo, Japan, 2020), pp. 210–214
27. X. Zheng, J. Yan, in IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE). Acoustic scene classification combining log-mel CNN model and end-to-end model (Tokyo, Japan, 2020). https://dcase.community/documents/challenge2019/technical_reports/DCASE2019_Zheng_57.pdf
28. D.A. Krause, A. Mesaros, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*. Binaural signal representations for joint sound event detection and acoustic scene classification (Belgrade, 2022), pp. 399–403. https://doi.org/10.23919/EUSIPCO55093.2022.9909581
29. Y. Hou, S. Song, C. Yu, W. Wang, D. Botteldooren, Audio event-relational graph representation learning for acoustic scene classification. IEEE Signal Process. Lett. **30**, 1382–1386 (2023). https://doi.org/10.1109/LSP.2023.3319233
30. T. Kawamura, Y. Kinoshita, N. Ono, R. Scheibler, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Effectiveness of inter- and intra-subarray spatial features for acoustic scene classification (Rhodes Island, 2023), pp. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096935
31. D. Johnson, S. Grollmisch, in *Proceedings of the 2020 European Signal Processing Conference (EUSIPCO)*. Techniques improving the robustness of deep learning models for Industrial Sound Analysis (Online, 2021), pp. 81–85. https://doi.org/10.23919/Eusipco47968.2020.9287327
32. T. Morocutti, F. Schmid, K. Koutini, G. Widmer, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*. Device-robust acoustic scene classification via impulse response augmentation (Helsinki, 2023), pp. 176–180. https://doi.org/10.23919/EUSIPCO58844.2023.10289983
33. A. Madhu, S. K, Rqnet: Residual quaternion CNN for performance enhancement in low complexity and device robust acoustic scene classification. IEEE Trans. Multimedia 1–13 (2023). https://doi.org/10.1109/TMM.2023.3241553
34. X.Y. Kek, C.S. Chin, Y. Li, An intelligent low-complexity computing interleaving wavelet scattering based mobile shuffling network for acoustic scene classification. IEEE Access **10**, 82185–82201 (2022). https://doi.org/10.1109/ACCESS.2022.3196338
35. G. Roma, P. Herrara, W. Nogueira, Environmental sound recognition using short-time feature aggregation. J. Intell. Inf. Syst. **51**, 457–475 (2018). https://doi.org/10.1007/s10844-017-0481-4
36. J. Ye, T. Kobayashi, M. Murakawa, Urban sound event classification based on local and global features aggregation. Appl. Acoust. **117**, 246–256 (2017). https://doi.org/10.1016/j.apacoust.2016.08.002
37. S. Proksch, M. Reeves, K. Gee, M. Transtrum, C. Kello, R. Balasubramaniam, Recurrence quantification analysis of crowd sound dynamics. Cogn. Sci. **47**(10), e13363 (2023). https://doi.org/10.1111/cogs.13363
38. S. Mirzaei, I.K. Jazani, Acoustic scene classification with multi-temporal complex modulation spectrogram features and a convolutional LSTM network. Multimed. Tools Appl. **82**, 16395–16408 (2023)
39. J. Serrà, X. Serra, R.G. Andrzejak, Cross recurrence quantification for cover song identification. New J. Phys. **11** (2009). https://doi.org/10.1088/1367-2630/11/9/093017
40. B. McFee et al. librosa/librosa: 0.8.0. Zenodo (2020). https://doi.org/10.5281/zenodo.3955228
41. M. Müller, F. Kurth, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Enhancing similarity matrices for music audio analysis (IEEE, Toulouse, 2006), pp. 437–440
42. C. Weiß, M. Müller, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Tonal complexity features for style classification of classical music (2015), pp. 688–692
43. S.S.R. Phaye, E. Benetos, Y. Wang, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Subspectralnet - using sub-spectrogram based convolutional neural networks for acoustic scene classification (2019), pp. 825–829. https://doi.org/10.1109/ICASSP.2019.8683288
44. M.D. McDonnell, W. Gao, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Acoustic scene classification using deep residual networks with late fusion of separated high

and low frequency paths (Institute for Electrical and Electronics Engineers (IEEE), New York City, United States, 2020), pp. 141–145
45. H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, in *Proceedings of the International Conference on Learning Representations (ICLR)*. mixup: Beyond empirical risk minimization (Vancouver, 2018). https://openreview.net/group?id=ICLR.cc
46. I. Loshchilov, F. Hutter, in Proceedings of the International Conference on Learning Representations (ICLR). SGDR: Stochastic gradient descent with warm restarts (Toulon, France, 2017), pp. 1–16
47. K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. Deep residual learning for image recognition (Las Vegas, 2016), pp. 770–778
48. S. Park, W. Choi, H. Ko, Acoustic scene classification using recurrence quantification analysis. J. Acoust. Soc. Korea **35**(1), 42–48 (2016). https://doi.org/10.7776/ASK.2016.35.1.042

Publisher's Note