DFRWS
DIGITAL FORENSIC RESEARCH CONFERENCE

BlackFeather: A framework for Background Noise Forensics

By:
Qi Li (University of Guelph), Giuliano Sovernigo (University of Guelph), and Xiaodong Lin (University of Guelph)

From the proceedings of
The Digital Forensic Research Conference
DFRWS USA 2022
July 11-14, 2022

Forensic Science International: Digital Investigation 42 (2022) 301396

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS 2022 USA - Proceedings of the Twenty-Second Annual DFRWS USA

BlackFeather: A framework for background noise forensics

Qi Li, Giuliano Sovernigo, Xiaodong Lin*

School of Computer Science, University of Guelph, 50 Stone Road East, Guelph, Ontario, Canada

ARTICLE INFO

Article history:

Keywords:
Background noise
Forensics
Deep learning
Environment

ABSTRACT

Historically, criminal investigations hinging on recorded audio data required manual application of forensic techniques to extract relevant information. These methods usually focus mainly on voices and speaker identification, but rarely focus on the wealth of forensic information available in the background noises present in the recording. Our paper introduces methods of automatically extracting, separating, and classifying background noises, allowing for the difficult, time-consuming process of audio analysis to be handled by software. Once the audio has been classified and examined by our proposed tools, the results can be used by investigators and forensic experts to aid in traditional investigative methods. Using environment information as an example, we propose a fully automated environment inference process based on background noise. Detailed experimental results show that our framework is effective and fast. Our proposed framework intends to provide a neat, automated, and accurate analysis of the information present in background audio, and to provide a new source of forensic information for investigators to leverage. In contrast to existing similar work, our scheme not only realistically considers mixed human voice speech, but also considers the case of multiple background noise mixes. To the best of our knowledge, this is the first forensic work that considers background noise in a complex environment.

1. Introduction

In forensic and criminal investigations, evidence is often extracted from every available source. Even small, seemingly insignificant details can often be the piece of evidence that allows the case to be solved. Some sources, such as recorded audio, require a significant amount of effort or expertise to extract helpful evidence from them due to their complexity and the subtlety of the information present in them. Historically, the main source of information available in recorded audio comes from speech present in the recording. Human speech recognition is, while not trivial, relatively easy for humans. However, human speech may only represent a small fraction of the available information in the recording, or may not even be present at all in audio recordings relevant to the case.

In the last decade, machine learning has become increasingly capable of performing the difficult task of isolating and identifying human voices in recordings. These new algorithms, architectures, and techniques make the processing of human speech in recording significantly easier, as well as providing a means of automating a previously human-centered process. However, as mentioned before, the human voice does not represent the total information in a recorded audio segment. Significant amounts of information are present in the background noise of raw audio recordings that have not been cleaned or isolated. Background noise can include sounds such as passing cars, bird calls, or even chimes or tones from specific machines.

This source of information is largely untapped in investigations, possibly due to the computational complexity and the subtlety of the problem. Compounding the complexity is the fact that the process of removing human speech in order to obtain a speech-free recording is not perfect. Often, traces of human voice are still present in a recording once a speech-removing mechanism has been used. Sometimes, these traces are still very prominent, and can completely obscure the audio in the background. The complexity coupled with the difficulty in isolation creates a problem that is either very time consuming, very human-reliant, or very inaccurate. However, if the process of isolation could be improved, and the process automated, background audio could be used as additional information in investigative processes. In our paper, we explore the possibility of using background audio information to provide

* Corresponding author.
E-mail addresses: qli15@uoguelph.ca (Q. Li), gsovernigo@uoguelph.ca (G. Sovernigo), xlin08@uoguelph.ca (X. Lin).

https://doi.org/10.1016/j.fsidi.2022.301396

estimates to the probable environment locations where this audio could have been recorded. Importantly, while background audio could ideally provide a very specific possible venue of the original recording (for instance, it is unlikely to hear a very loud train anywhere other than a train station or railway track), there is a certain amount of variability in the environment where an audio recording may have been made. As an example, the hum of a fan is equally likely to be heard in a home as it is in an appliance store, rendering the information provided by the presence of the fan less specifying and useful than it could be.

The concept of using background audio as a source of information is not a novel concept, but actually has basis in actual investigations that have hinged on the use of specific sounds present in recordings which were used to place certain people in certain environments. In Europe, some police groups have even hired officers with visual impairments due to their increased abilities to recognize specific sounds and noises that are present in recordings. While possibly exaggerated, some of these officers claim to be able to distinguish the different brands of motors in vehicles simply from their idling frequencies and associated noises (Bilefsky, 2007). Depending on the certainty of the investigating officers, these clues gathered from the recordings could be used directly as evidence in trials and investigations due to their high accuracy. However, there is a major downside to this process: human resources are one of the most expensive resources in the world. Rather than having human officers perform these tasks when their critical thinking and reasoning skills could be used elsewhere, the ideal situation would have a machine handling the classification and separation of audio, with the officers acting as oracles and critical thinkers who can leverage the automated nature of the background classifier.

To address the technical challenges posed by such complicated data processing, some researchers have proposed preliminary schemes. Ikram and Malik (2010) proposed a similar concept for digital audio forensics in 2010, focusing on the use of spectral subtraction to extract background noise. It was noted in the paper that this process leaves significant human voice artifacts in the recording. In addition, Ikram and Malik's work does not explore any automatic method for classifying the extracted background noise. Singh and Joshi proposed a convolutional neural network-based background noise classification technique (Singh and Joshi, 2019). This work also focuses on extracting background noise from audio containing a mixture of human speech and background noise, but can only identify a single source of background noise. In other words, the proposed solution does not have the ability to distinguish, separate and classify multiple overlapping background noises. In addition, the classification model used in this paper tests only 10 classification categories, which is too limited for our application. Thorogood et al. builds a background/foreground classification task in a musicology and production-related context and proposes an automatic segmentation method for soundscape recordings based on this task (Thorogood et al., 2015). However, the paper does not classify specific background noises, but rather classifies noise into six broad categories.

To overcome these challenges and shortcomings, we propose a novel background noise forensic framework, called BlackFeather.[1] Our tool's framework is a multi-network pipeline consisting of several specialized neural network architectures designed for each processing step. Essentially, our proposed solution has three main components for the first step, which handle the respective tasks of extracting, separating, and classifying the background noise present in a recording. Our decision to split the model into three distinct components arises because of the distinction between the three main tasks that the model must accomplish, and how different the network architectures for each part of the solution must be. Audio cannot be reliably classified into categories without first being separated, and cannot reasonably be separated when human speech is present, obscuring potentially important subtle changes. As such, our model first extracts the background audio from the human speech (which we call the “foreground audio”) before passing it to the separator module, which divides the background audio into the most likely distinct sets. Next, our classification module attempts to classify these separated sounds into their most likely categories. Finally, we propose a community framework in the second step to determine the environment of the background noise according to the output of the first step. Our contributions presented in this work are summarized as follows:

- We present the first comprehensive scheme for background noise forensics. Our proposed scheme is divided into two steps. In the first step, the background noises are extracted, separated and identified. In the second step, the identification information is modeled or used by experts to get the environment information.
- This paper is the first forensic work for audio containing multiple, overlapping background noise sources as well as speakers' voices. We introduce the noise extractor and the multi-speaker separation in speaker recognition to the background noise and achieve good results.
- We propose new datasets and mechanisms for the analysis of audio and training of networks designed for these tasks. Among these proposed improvements are the top-K selection for classification, as well as the proposed combined datasets such as MixEsc50.
- Detailed experiments show that our scheme is effective, fast and accurate.

In this paper, we will briefly introduce some preliminaries in Section 2. Additionally, Section 3 will give a detailed breakdown of our model, down to the individual components of the model. In Section 4 and Section 5, we will examine the results of our experiment, and the testing results and scoring metrics of each of our components separately, as well as its performance as a whole. In Section 6, we summarize some applications in digital forensics. Section 7 will discuss works related to our objectives and techniques, and examine the differences and contributions of each. Section 8 will present our conclusion.

2. Preliminaries

In this section, we briefly recollect the basis concepts of sound event classification and speech separation. Finally, we give some remarks about our work.

2.1. Sound event classification

Sound event classification is the process in which sound events are analyzed and assigned a predefined label (e.g., “dogbark”, “siren”) within an audio signal. It has numerous applications, including audio surveillance systems, hearing aids, smart room monitoring, and pornographic content detection. Several studies have been conducted to develop and propose solutions to the problem of identifying and labeling sounds in recordings. Additionally, datasets have been developed with pre-labeled sounds that can be used to train new classifiers with non-standard

[1] Our inspiration for the name BlackFeather is multifaceted: background noise forensics and black feather have the same acronym, B.F.; crows, often alluded to by a black feather, represent death in many cultures, which is highly relevant to a forensic pathologist.

architectures. Examples of these datasets include ESC-50 (Piczak et al., 2015), AudioSet (Gemmeke et al., 2017) and UrbanSound8k (Salamon et al., 2014).

2.2. Speech separation

Fundamentally, speech separation is a problem where separation of signals from other signals must be achieved solely from the audio and without context. Formally, this problem is known as the "cocktail party problem", where a machine must focus on a specific message of interest, and discard the interference of noise in the environment. It is named the "cocktail party problem" because of the human ability to listen to one person's voice in a crowded cocktail party while discarding the voices of others. While this is simple for humans, this is an extremely difficult and complex task for a machine to accomplish.

Speech separation is a signal processing technique which calculates the individual signal from the received mixed speech signal by applying certain methods. It can be roughly divided into two classes of problems: the first is the separation of multiple sources in the speech signal, and the second is the separation of a single source from noise and other interference in the speech signal.

Many studies have been proposed for the separation of human voice (Luo et al., 2020), but no studies have been conducted for the separation of background noise. Compared to the human voice, the sound in background noise is often shorter, and it is very difficult to characterize sufficiently in order to achieve the effect of human voice separation.

2.3. Remarks

In summary, there are several aspects that distinguish BlackFeather from existing background noise classification solutions. First, previous work usually focuses on pure background noise classification without considering multiple noise mixed together. In contrast, we systematically study the full life cycle of background noise with extraction of background from speech and separation of mixed background noise. To the best of our knowledge, this is the first systematic research on the whole background noise forensics.

3. Our proposed framework

In this section, we briefly introduce the details of our proposed background noise forensics framework.

3.1. Problem formulation

Given an audio segment x which contains human speech h and background noise b, a forensic toolkit/framework f aims to recognize the detail of b. b may be a mixed background noise, as in, there may be multiple, distinct background noises present. After extracting, separating, and evaluating the background noises that are present, the framework will attempt to predict likely environmental locations, d, according to signals present in b. To accomplish this, our toolkit, divides the task to 2 subtasks: noise information extraction and noise profile analysis. Noise information extraction can be divided to three functions: $Extract(\cdot)$, $Separate(\cdot)$, $Classify(\cdot)$ as shown as follows:

$$b = Extract(x), \tag{1}$$

$$[b_1, \ldots, b_k] = Separate(b), \tag{2}$$

$$[n_1, \ldots, n_k] = Classify([b_1, \ldots, b_k]), \tag{3}$$

$$d = Analyze([n_1, \ldots, n_k]). \tag{4}$$

First, our framework extracts background noise b from an audio sample x, which is under investigation. Secondly, the framework separates the extracted background noise b to a background noise array $[b_1, \ldots, b_k]$. Third, the array is classified and labeled by their most likely cause for each respective noise that was identified. Finally, the framework examines the array of labelled noises, and attempts to deduce the environments of likely origin where the original audio could have been recorded.

3.2. Overview

The framework we propose uses multiple machine learning modules to create a series of sequential processing steps corresponding to the steps identified before. Each component handles a distinct step in our process, and uses part or all of the output from the previous module as its input. In this section, we will examine each component of our network pipeline in depth, and discuss the parameters and processes in more detail.

A diagram of the models used in our proposed scheme is shown in Fig. 1. As we analyzed in Section 3.1, it consists of four modules: the background noise extraction module, background noise separation module, background noise recognition module, and environment analysis module. The background noise extraction module handles extraction of background noise signals from foreground noise (human speech) that may be present. The noise separation module separates the previously extracted background noise into the most likely individual noises that comprise the background noise. The noise recognition module recognizes and labels the separated noises, outputting those labels. Once the pipeline has operated on a segment of audio, a combination of manual, machine learning, and community analysis is used to predict likely environments where the original audio segment could have been recorded based on the separated and labeled audio.

3.3. Extractor module

As mentioned before, the extractor module handles the separation of background and foreground audio. To accomplish this task, the module processes the sampled noise with a machine learning network designed to separate the background noise from human speech in the foreground. This first module uses the multihead attention neural network model as described by A. Vaswani et al. (2017). For reference, this original track is referred to in this paper as the *original audio*. Next, we use two methods that whether the original audio is passed as input to a denoiser module. In our experiments, the Facebook Denoiser pretrained Demucs model was used as the denoising component (Defossez et al., 2020), though any denoiser with sufficient performance in extracting clean voice signals could theoretically be used. The denoising process produces a secondary audio track which is an extracted, isolated voice track ideally with minimal loss or distortion. This track is referred to as the *extracted voice*. The purpose of this isolation is to provide a "noise profile" that can be used by the extractor to correctly remove the human voice, ideally resulting in less artifacts of improperly extracted speech left in the extracted audio. Both of these tracks were then converted to a Mel spectrogram representation, concatenated, and then provided to the background isolating network. The output of this network is an isolated background audio estimate, which ideally is a lossless representation of the background audio without human voices present.

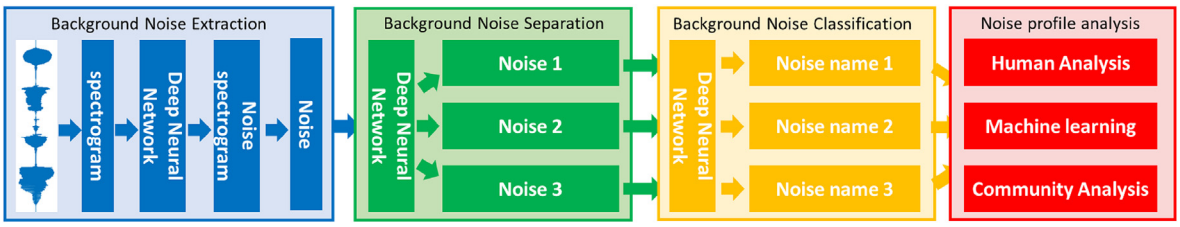

Fig. 1. The workflow of our proposed BlackFeather framework.

For training, the process of this component is slightly different. The original audio is created from a deliberately isolated voice and background track. The separated background component is kept as the training label, and referred to as the target background. During training, the background estimate is compared to this original target background using MSE (mean squared error) (Lohrasbipeydeh and Gulliver, 2021) as an error metric. A diagram of both of these paths can be visualized as shown in Fig. 2.

It should be noted that the output of the neural network model is a Mel spectrogram, similar to the processed inputs. Next, we use a sophisticated vocoder to recover the mel spectrum to audio (Jia et al., 2018). Once processed, the recovered background voice is used as an input to the next component of the pipeline.

An essential step here is the first isolation of the human voice. Previously, our work had focused on extraction training using only the original audio and the desired output as the training data and label, respectively. This process leaves audio artifacts of the human voice behind; the artifacts are barely audible to the human ear, but nevertheless represent a decrease in the extracted audio's quality. In contrast, using the denoised human voice as a secondary input to the background extraction network in combination with the original audio provides audio with a significantly reduced artifact presence.

3.4. Separator module

The next component of the model is the Separator, which attempts to separate distinct sounds in the extracted background noise from the previous component. Because the classification process will be less accurate if multiple noises are overlapped, the Separator module is used to separate the distinct background noise sources before the classification occurs.

Our background noise separation is inspired directly by existing research in the field of speech separation. In our work, we use a Dual-Path-RNN as was proposed in (Luo et al., 2020); this work achieves excellent results in speech separation. Specifically, we feed the extracted background noise into a dual path neural network. The neural network consists of the following three main phases:

- Segmentation phase: The sequential input (extracted background noise sample) is split into overlapping blocks. Then, all blocks are connected into a 3D tensor.
- DPRNN block operation phase: The tensor is passed to the stacked DPRNN blocks and local (intra-block) and global (inter-block) modeling is applied iteratively in an alternating manner.
- Recovery phase: This phase focuses on recovering the generated speech sequences. The output of the last layer is converted back to sequential output by overlap summation. It is the inverse process of the segmentation phase.

Last but not least, although BlackFeather takes two-source separation as an example, the model can be easily extended to the separation of an unknown amount of background noise. Specifically, the separation module can be adapted to recursive separation, where one background noise is separated at a time (Takahashi et al., 2019).

3.5. Classifier module

To classify the noise separated in the previous section, we use existing works focusing on the labeling and classification of audio signals, but with a slightly different method to address the issue of multiple, overlapping noises. The background noise separation inspired by existing solutions in speech separation does not achieve results which are as good as when applied to the originally intended task of speech separation. To address the more complex audio characteristics of mixed background noise, we propose a new classification computation scheme to improve the classification accuracy. The new computational scheme is shown in Fig. 3.

The classic method in deep learning would be to output the resulting labels one by one. The classification network would first

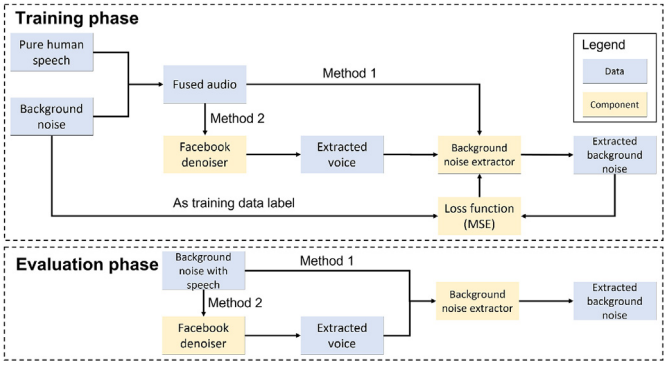

Fig. 2. Extraction pipeline of BlackFeather.

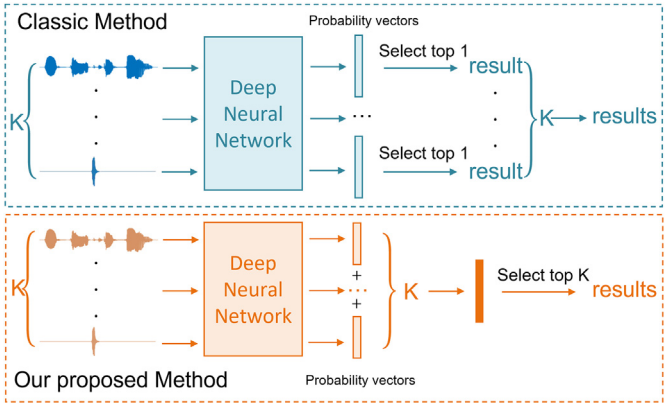

Fig. 3. Computing methods of classification results.

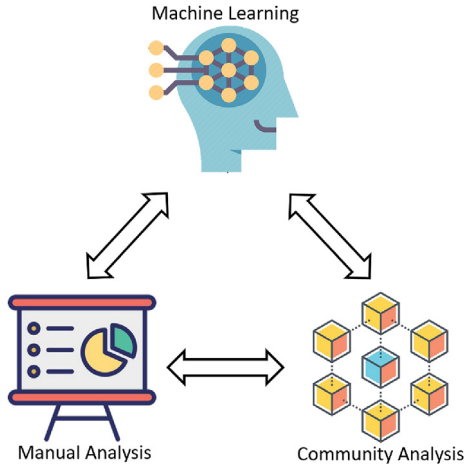

Fig. 4. Noise profile to environment model.

output a feature representation and go through a softmax[2] or similar operational layer to get a probability vector $P = [p_1, p_2, \ldots, p_n]$. Then the largest of these is selected, and the corresponding ordinal number or ID would be the resulting output.

This solution is not ideal for a multi-noise environment however; we need to consider that background noise separation is not as easy as speech separation. When we get the first separated noise, it also contains some part of second separated noise. If we use the classic method of sending the noises to get result one by one and combining these results together $\{ID_k, \ldots, ID_q\}$, it will omit the second separated noise information in the first noise classification process.

To illustrate this issue, let us assume the first and second separated noise are k-th and q-th category noise in all categories. In the classic method, the classification network will first extract the feature representation f_k and f_q of k-th and q-th noise. Through a softmax layer, f_k and f_q are used to calculate the probability vector P_k and P_q, respectively. Due to the lower performance of background noise separation as compared to speech separation, the k-th noise may also contain some traces of the q-th noise. Similarly, the q-th noise may contain traces of the k-th noise. This means that the q-th value of $f_k = [f_{k1}, f_{k2}, \ldots, f_{kn}]$ and $p_k = [p_{k1}, p_{k2}, \ldots, p_{kn}]$ are also high. Regardless of the presence of the q-th audio trace in the k-th audio signal, the classic method would only select the label for the k-th category, omitting the q-th value of f_k and p_k.

This leaked information represents a large amount of useful information for the learning modules. To include this information, we can simply add all the probability vectors together, and select the top K categories. Using this top-K method, previously neglected quantities will be taken into account by subsequent selection methods. Specifically, we propose two top-K classification methods. Method 1 first sums the outputs F of the K separated noises before the softmax layer:

$$F = \sum_{i}^{K} f_i. \tag{5}$$

Finally, the summation result F is fed into the softmax layer to obtain a probability vector P containing the probabilities of each category:

$$P = softmax(F). \tag{6}$$

Method 2, on the other hand, first sends the feature representations f_i, $i \in \{1, 2, \ldots, K\}$ extracted from the K separated noises directly to the softmax layer to calculate the vectors p_i, $i \in \{1, 2, \ldots, K\}$ containing the various probabilities:

$$p_i = softmax(f_i), i \in \{1, 2, \ldots, K\}. \tag{7}$$

Finally, the probability vectors are summed to obtain the total probability vector:

$$P = \sum_{i}^{K} p_i. \tag{8}$$

In addition to our top-k method, we also use AudioCLIP (Guzhov et al., 2021) to solve the problem of recognizing new, user-added background noise. In addition to the existing text-image-similarity loss term, a new loss term is introduced in AudioClip: text-audio. During the training phase, given a set of input text–audio pairs of size N, both the text subnetwork and audio subnetwork produce the corresponding embeddings that are mapped linearly into a multimodal embedding space of size 1024. In this specific setup, AudioCLIP learns to maximize the cosine similarity between matching textual and acoustical representations, while minimizing it between incorrect ones, which is achieved using symmetric cross entropy loss over similarity measures. Based on AudioCLIP, BlackFeather allows users to customize the class of background noise without having to train from scratch. BlackFeather will automatically compute the text-semantic feature vector of all user-defined background noise labels and associate it with the background noise feature vector. By doing so the background noise is classified into self-defined categories.

3.6. Noise profile to environment model

We offer three ways to analyze the noise profile as shown in Fig. 4.

3.6.1. Manual analysis

As the name implies, this approach is a direct analysis by the user or forensic expert. From the output of the previous stage, the user is directly presented information about the type of background noise contained in the original audio, and can make inferences about the environment based on this information and their own intuition and experience. For example, if the user gets the output as a noisy human voice and a bus engine starts, then the user will be able to deduce the environment at a bus stop or a bus station.

3.6.2. Machine learning

Any machine learning inference or multi-classification algorithm can be applied here. In this paper, we use a naive Bayesian inference model as an example to illustrate the process of machine learning inferential forensics. We collect enough training samples (noise–environment pairs) and then use them to train the naive Bayesian inference model.

3.6.3. Community analysis

Community analysis is the final option for noise profile analysis, and would be done by a community of participants (of contextually appropriate trust) who are given the audio and likely labels. The participants' level of trustworthiness is obviously dependent on the context of the use of the model; for non-critical work, any individual can be trusted, whereas for law enforcement, volunteer human classifiers could be given a certain amount of training

[2] Assume x, y are two vectors, x_i, y_i are the i-th value of x, y. If $y = softmax(x)$, then $\forall i$, $y_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$.

appropriate to their task, or else could be experts hired for that purpose. Since officers or machine learning platforms may not be familiar with nuanced and subtle sounds, a community of people who are willing to volunteer their time or hired experts to manually classify a sound can be used as a substitution. When a volunteer user gets a new noise or a new environment location in forensics, he/she can share it with the community, allowing other users of the tool to benefit from other investigative work conducted with it. Other experienced users in the community can update classifications based on their experience if they are able, and can help to answer the environment forensics questions proposed by other users. This approach maintains the ability of our environment forensics analysis tool to evolve and can be combined with the previous two analysis methods of manual analysis and machine learning. The forensic tool can turn the questions answered by the experienced users into new training data and update the accuracy of the machine learning method.

4. Experiment setting

In this section, we describe the datasets that were used, performance evaluation methods, the training process, corresponding hyper-parameters, and finally the experimental results and analysis.

4.1. Dataset

In this work, three public datasets are used: AudioSet (Gemmeke et al., 2017), ESC-50 (Piczak et al., 2015), and Librispeech (Panayotov et al., 2015). In addition to this, we also generate three new datasets, which we will call the "CombinedAudioSet", "CombinedESC50" and "MixESC50", based on the existing AudioSet and ESC-50 datasets, respectively. Here, we describe these datasets and formalize their roles in the training and evaluation processes of our different modules.

4.1.1. AudioSet

The AudioSet dataset is a large-scale collection of auditory data that is non-exclusively grouped into 527 classes (Gemmeke et al., 2017). Each sample is a fragment from a YouTube video that is up to 10 s long and is specified by the accompanying ID and timings. This dataset is used in our research to extract background noise and generate new datasets: CombinedAudioSet.

4.1.2. ESC-50

The ESC-50 dataset provides 2000 single-channel 5s long audio tracks sampled at 44.1 kHz (Piczak et al., 2015). As the name suggests, the dataset consists of 50 classes that can be divided into 5 major groups: animals, natural and water, non-speech human, interior, and exterior sounds. To ensure correctness during the evaluation phase, the raw ESC-50 dataset was divided by the defined major categories when used in our experiment. During training and evaluation, this dataset is used to provide the sources of background noise for the background noise separation module.

4.1.3. Librispeech

This dataset includes both text and audio audiobook datasets (Panayotov et al., 2015). It is a dataset of nearly 500 h worth of audiobook excerpts read clearly by multiple people, and it is structured by book chapters containing speech and the corresponding text. In our experiments, this dataset is used to provide the pure human voices in the background noise extraction phase, and is used for both evaluation as well as training.

4.1.4. CombinedAudioSet

To the best of our knowledge, there is no dataset that can be used for background noise extraction. Instead, we generated a dataset based on AudioSet and Librispeech, which we called CombinedAudioSet. We randomly select a sufficient number of human speech segments from Librispeech, and combine them with the background noise from AudioSet. This yields a labeled dataset containing a combined track, an isolated voice track, and an isolated background track. This dataset is used to train the background noise extraction module, and its labeled isolated tracks are used for training and evaluation of that model.

4.1.5. CombinedESC50

In the same way as CombinedAudioSet, we generated a dataset based on ESC-50 and we call it Combined-ESC50. Its details are the same as CombinedAudioSet, so we do not illustrate it here.

4.1.6. MixESC50

To the best of our knowledge, there is no existing dataset that can be used for background noise separation. Therefore, we created a dataset based on ESC50, which can be used for background noise separation, which we called MixESC50. To generate this new dataset, we select k background audio samples from the 50 categories in ESC-50 and subdivide each sample into 10s lengths to get $50k$ tracks. Finally, we combine these individual 10s tracks into overlaid pairs, yielding $1225k^2$ final samples. In our experiments, we set $k = 5$, and used 80% of the generated samples for training, and 20% for testing.

4.2. Evaluation metric

The performance of our model is assessed based on the performance of each of the four tasks: extraction, separation, classification, and environment deduction. In this section, we will describe the evaluation metrics and methods used in each task, as well as provide a metric to evaluate the overall performance of the entire system.

4.2.1. Extractor module

In this module, we mainly use Euclidean Distance to measure the accuracy of the extracted background noise compared to the label. In mathematical and machine learning terms, this metric is often called the L2-norm or Mean Squared Error (MSE). During evaluation, we use the original, pure background audio that was used to construct the merged track as the label for that merged track, and calculate the Euclidean distance between the extracted background noise and this label to measure the degree of "closeness" of the extracted audio to the original background audio.

4.2.2. Separator module

For background noise separation, we use the SI-SDR (Scale Invariant Signal-to-Distortion Ratio) metric (Vincent et al., 2006), which is most commonly used in speaker voice separation, to evaluate the accuracy of the model's separated noises.

We assume the correct separated vector output from a perfect separation is $\widehat{X}$ and the output of the trained separation model is $X^{\star}$. The projection of $X^{\star}$ onto $\widehat{X}$ yields X_T and the remainder of $X^{\star} - X_T$ yields X_E. These vectors satisfy the following relation:

$$X^{\star} = X_T + X_E.$$

From this relationship, we can know that if the SI-SDR score is larger, $X^{\star}$ and $\widehat{X}$ are more parallel. A higher SI-SDR score therefore also means that the separation model is more effective, since the

vectors $X^{\star}$ and X_T are more similar.

4.2.3. Classifier module

For the classifier module, we use the rate of correct classifications to measure the classification accuracy. We count the classification results for three different cases: top 1, top 3, and top 5. Top k means that we select the top k possible classification results in the classification output, and if the correct answer is present in the top k selected results, then the classification is successful.

4.2.4. Overall performance

Here, we also use the rate of correct classifications to measure the performance of the whole process from extraction, separation to classification. And we output the classification results for three different cases: top 1, top 3, top 5. In other words, we use the cumulative entire process, from beginning input to final classification output as the result, and measure accuracy based on the rate of correct classification from the final module compared to the original labeled inputs given to the extraction module.

4.2.5. Environment deduction

Environment deduction is also treated as a classification task. The same rate of classification correctness is used to measure the accuracy of environment deduction.

4.3. Training

The entire training process was divided into a set of subsequent steps, which made the realization of the final BlackFeather processing scheme reliable and assured its high performance. As described in Section 3.3, the extraction module uses a Transformer-based extraction model and was trained using our generated dataset (Section 4.1). In the separation module, we use an unmodified DualRNN network (Luo et al., 2020) with our generated dataset. In the classification module, we use a pre-trained AudioClip (Guzhov et al., 2021) model directly to handle the functionality of our classification module.

Our experiments were all run on a Windows PC with an NVIDIA p5000 GPU and a Intel(R) Xeon(R) Silver 4215 CPU @ 2.50 GHz. In extraction, separation and classification fine-tuning training, we set the learning rate of the network to 10^{-4}. In the noise extraction network training, we set the number of epochs for the iterations to 200 and use the adam optimizer to optimize. In the separation network training, we set the number of epochs for the iterations to 70 and use the adam optimizer. In the classification network training, we set the number of epochs for the iterations to 60 and use the SGD optimizer with 0.9 momentum. Prior to being processed by the first network, the input is loaded from the filesystem at a sample rate of 16 kHZ using librosa. We set the sample rate of separator and classifier module are both 44.1 kHz.

5. Experimental performance

In this section, we will analyze the evaluation results for each part of our solution. This includes background noise extraction, multiple background noise separation, classification module, full noise process analysis, and environment inference.

5.1. Extractor module

As we stated in Section 4.2, we use the Euclidean distance to measure the accuracy of our extraction results. Our results, which use our generated dataset CombinedAudioSet, are shown in Table 1. We tested our two proposed background noise extraction methods using two datasets, CombinedAudioSet and CombinedESC50. Considering the large dimensionality of the extracted mel spectgram, this results of MSE are extremely small, indicating that our method is successful enough.

Table 1
Background noise extraction results.

MSE	CombinedAudioSet	CombinedESC50
Without Denoiser	2.28	2.19
With Denoiser	1.80	1.68
Original Distance	3062.74	2944.50

5.2. Separation

We trained our separation network using the MixESC50 dataset and the results are shown in the Table 2. In our generated dataset, we use two background noises mixed into one. Therefore, our separation task is also to separate a sound that mixes two background noises into two.

As we can see from the Table 2, background noise separation is more difficult than speech separation. The SI-SDR value of background noise separation is lower than speech separation but still is acceptable and works in BlackFeather.

5.3. Classification

We conducted experiments for the widely-used existing classification calculation method, our proposed new Method 1 (summing in front of the softmax layer), and new Method 2 (summing after the softmax layer), and counted the rate of correct classifications for each method in the case of selecting top 1, top 3, and top 5, as shown in the Table 3. In more detail, if there are 2 background noise sources in an original audio segment, and both sources are correctly identified in the top $k(k \in \{1, 3, 5\})$ results, then the accuracy of the model can be considered 100%. If only one of them is in the classified top k results, the accuracy would be considered 50%. If none of them is in the classified top k results, the accuracy is 0%.

It can be seen that both of our proposed methods improve the classification of the separated noises compared to the existing method. As we stated in Section 3.5, separating the background noise does not achieve the same accuracy compared to the separation of the speech. Additionally, the background noise fed to the classification network may be mixed with another type of noise. Our methods take full advantage of these features and are therefore more suitable for the task of separating background noise and then classifying it.

5.4. Full pipeline noise process

We conducted experiments for the three processing modules of the entire scheme for noise analysis in tandem. We tested the correctness of individual pure background noise, the accuracy of the separated background noise, the accuracy after fine-tuning. And compared the accuracy of the new and old methods. The results are shown in Table 4.

Table 2
Background noise separation results.

	SI-SDR
Human speech (Luo et al., 2020)	19.10
Background noise with training	7.64
Background noise without training	−25.4

Table 3
Classification results of Method 1 and Method 2 in MixESC50.

	Top 1	Top 3	Top 5
Old method	82.67%	93.45%	97.53%
Method 1	84.91%	94.39%	97.84%
Method 2	86.74%	95.16%	98.22%

Table 4
Full pipeline classification results of Method 1 and Method 2.

	Old method	Method 1	Method 2
Pure single noise	94.85%	–	–
Separated noise	82.67%	84.91%	86.74%
Fine-tuned model	87.18%	90.15%	91.32%

When we use the pre-trained background noise classification model directly and evaluate it using the test set, the accuracy is 94.85% as seen in Table 4. However, when we use the separated background noise as input, we see a decay in the accuracy, which drops to 82.67%. This is because the separation of background noise is not as consistent and reliable as the separation of human speech. At the same time, we use the proposed methods to improve the accuracy to 84.91% and 86.74% respectively. However, when we generated the training dataset, we used two single background noises mixed together, which means that the inputs are correctly labeled with the correct classification, allowing us to use these labels to fine-tune the pre-trained classification network so that it is suitable for separating the noise. Using this method of fine-tuning the pre-trained model, the accuracy is can be improved to 87.18%. When we use the proposed new method, we can see from the Table 4 that the accuracy rises to 90.15% and 91.32%, respectively. This proves that our proposed methods make full use of the forgotten information of the original method, thus improving the accuracy.

At the same time, we also tested AudioClip for mixed noise classification, and the top two categories with the highest probability of recognition results were regarded as the classification results. The results of BlackFeather and AudioClip on the MixEsc50 dataset are shown in Table 5. Clearly, BlackFeather separation-and-reclassification approach achieves a qualitative improvement over AudioClip and is more suitable for complex background noise classification.

Since BlackFeather is intended to function as a deployable toolkit, we also tested its time overhead. We focused on testing the time overhead of the noise information processing steps (including extraction, separation, and classification modules, without noise profile analysis), since these would likely be the most time-consuming components in the tool. The experimental results of the time overhead tests are shown in Fig. 5. The processing time of the whole system is less than 0.5s when performed on the GPU used in these experiments. When using a CPU exclusively, the time is still less than 10s. The results show that our solution is very fast and can meet user requirements whether running on GPU or CPU. It is fully capable of real-time analysis and application in forensic work.

5.5. Environment analysis

For the machine learning component of the environment analysis, we employed a naïve Bayesian model as an example to prove the viability of our solution. We generated 1721 background noise and environment location pairs in a small test dataset, using 1376 pairs of them to train and 345 pairs for testing. These pairs including 37 unique classification (environments) and 112 kinds of noise. Using this test dataset, we achieve 95.8% accuracy in environment classification, though this method is not necessarily ideal for a production version of the toolkit.

Table 5
Classification results of BlackFeather and Original AudioClip (Guzhov et al., 2021).

	BlackFeather	AudioClip (Guzhov et al., 2021)
Accuracy	91.32%	46.3%

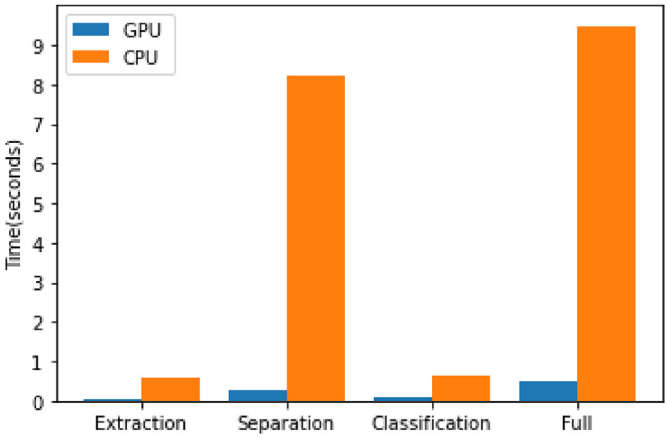

Fig. 5. Time overhead of BlackFeather.

6. Applications in digital forensics

Audio forensics is complicated and time consuming, and the benefits of automation in the analysis of captured audio will ideally result in more effective use of audio as evidence during investigation. In this section, we summarize some possible applications of BlackFeather in the field of digital forensics in this section.

6.1. Environmental inference

BlackFeather can analyze the background noise in a recorded audio clip and extract sensitive information about the sources of background noise present in the recording; these can be used to provide context and clues to investigators which can assist in placing a recording at a specific environment where the audio sources detected could have occurred at. Similarly, this could be used to suggest environments where the sound could *not* have been recorded. As an example, audio recordings which were purported to take place in a deep forest away from other people should not contain the sounds of a busy highway or chatter in a crowd. This could be used both to place a suspect in an audio recording at a specific environment or else lend credibility to a suspect's alibi should the audio clip contain sounds that could not have occurred in that locale. To determine guilt, the categorical trinity, means, motive, and opportunity, must be satisfactorily addressed; in the proving opportunity, location is often key to establishing whether a party could have committed the crime.

6.2. Temporal inference

Some sources for background noises are time sensitive: certain birds may only be known to make their calls at night; many bell towers have specific bells for specific times, or else may chime only at noon and midnight. Busses and trains follow set schedules, and

their presence can be used to narrow recording time down to minutes. The presence of certain sounds may completely disprove a false timeline, disproving or confirming alibis. Seasonally, certain bugs and birds make their calls only during their mating or hatching seasons, such as the cicada; a time of year could be determined from the presence of geese in a northern climate, for instance. Determining time from sound has been proven possible in the U.K. using the hum of the mains power supply, which has been used in courtroom cases before (Morelle, 2062). In a forensic investigation, timelines are critical, and establishing a timeline for a series of events where the order of events can determine guilt or innocence is of the utmost importance.

7. Related work

Ikram and Malik (2010) proposed a similar concept of digital audio forensics focusing on the extraction of background noise in 2010. Ikram and Malik's method of isolating background noise involved using spectral subtraction to remove the human voice; a process which was noted in the paper to leave significant human voice artifacts in the recording. Because isolation almost never produces a perfect representation of the target audio, spectral subtraction, lacking any interpolation techniques, will often miss large parts of the speech that were not correctly isolated by the isolation method. These remaining artifacts are referred to in the paper as “speech leakage”.

Singh and Joshi propose a background noise classifying technique based on a convolutional neural network that operates on a dataset, called YBSS-200, which was sourced from YouTube (Singh and Joshi, 2019). This paper also focuses on the extraction of background noise from audio containing mixed human speech and background noise, but can only identify single-source background noise. In other words, the proposed solution does not have the ability to differentiate, separate, and classify multiple overlapping background noises. In addition, the classification model used in this paper was only tested with 10 classification categories, which is far too limited for our application. Nevertheless, this paper represents foundational work that we hope to expand upon.

Thorogood et al. established a background/foreground classification task in a musicological and production-related context and proposed an automatic segmentation method for soundscape recordings based on this task (Thorogood et al., 2015). However, the paper does not classify background noise into many varied classifications, but classifies noise into 6 general categories.

Numerous researchers have presented many articles on background noise classification (Guzhov et al., 2021; Saki and Kehtarnavaz, 2014; Hornauer et al., 2021), but all of them consider only pure sound events. Not only is the background noise mixed with speech is not considered, but also the case of mixing multiple background noises is not considered.

There has been some recent work on acoustic scene detection. Bisot et al. proposed an acoustic scene classification method based on supervised feature learning (Bisot et al., 2016). They argued that acoustic scenes can be expressed by time-frequency image features, which can effectively use the structural features of time-frequency images and do not need to focus on the specificity of time-frequency images. After pre-processing the time-frequency images, the time-frequency images are decomposed using the Non-negative Matrix Factorization (NMF) technique, followed by Principal Component Analysis (PCA) for feature dimension reduction, and finally a classifier is used to determine the acoustic scene class. Abidin et al. proposed a local binary pattern (LBP) based Binary Pattern and Random Forest (RF) approach for acoustic scene classification (Abidin et al., 2018). They subjected the audio signal to Constant-Q Transform (CQT). To obtain the local features related to the spectral information, they divided the spectrum after Constant-Q Transform into several sub-bands and then extracted the time-frequency features using LBP. Zhao et al. concluded that although the speech spectrogram can achieve better results in acoustic scene classification, it fails to take into account the continuous spectral features of the audio signal (Ren et al., 2018). They investigated a depth-scale transformed spectrum based on the speech spectrogram by first converting the acoustic scene spectrum into a Morse-scale spectrum and then feeding it into a trained convolutional neural network. The acoustic features are then extracted using a fully connected network and a recurrent neural network. However, all these methods work from the collected scene sound; not only is the number of recognizable scenes small, but the methods cannot be easily extended. Additionally, the methods do not take into account the presence of human speech in the audio recordings.

As far as we are aware, no other research explores the use of background sound as a means of identifying the environment of a recording. Other research has explored the extraction and classification of background audio, but usually not for the purpose of a forensic investigation, and none attempt to do so in a multi-sound-source environment or with such a high number of classifications.

8. Conclusion

Background audio represents a wealth of untapped digital information which could be used in investigations, but may be too time-consuming or difficult to extract manually. To combat these shortcomings, automation can be used to augment or even replace the manual analysis in audio forensics, leaving the investigators to focus on the larger facts in the investigation. In this paper, we propose a systematic, multi-faceted framework to address the challenge of automated background noise forensics. Our aim is to produce a framework for the analysis of background audio to expose information such as the potential classifications of the environment of the recording. In our experiments, we focus primarily on the environment information as a target for analysis, and achieve good results. We have carefully designed, constructed and experimented for each module of the proposed framework, and propose new datasets and mechanisms for the analysis of audio and training of networks designed for these tasks. Among these proposed improvements are the top-K selection for classification, as well as the proposed combined datasets such as MixEsc50. Our aim for this research is to propose a framework capable of replacing or augmenting existing manual or assisted manual forensic work. While our work focuses mainly on environment inference, it can also be extended and fine-tuned to accommodate forensic tasks beyond environment.

In future work, the extension of our proposed background audio separation techniques would likely be most significant as well as challenging. In particular, the major gap between speech separation accuracy and background audio separation accuracy requires more investigation and more specialized techniques to close. However, at present our work represents a new direction for digital forensic work, and will hopefully lead to more research into the automation of forensics. Ideally, our work will pioneer a new field of forensic research, and will provide a fundamental baseline for improvement in the field of automated audio analysis. Additionally, future work on this topic will hopefully explore the significance of end-to-end models as compared to our modular approach.

Acknowledgement

This work is supported in part by NSERC (Natural Sciences and Engineering Research Council of Canada), Canada.

References

Abidin, S., Xia, X., Togneri, R., Sohel, F., 2018. Local binary pattern with random forest for acoustic scene classification. In: 2018 IEEE International Conference on Multimedia and Expo, ICME 2018. IEEE Computer Society, San Diego, CA, USA, pp. 1–6. https://doi.org/10.1109/ICME.2018.8486578. July 23-27, 2018.

Bilefsky, D.. A blind sherlock holmes: fighting crime with acute listening. https://www.nytimes.com/2007/10/29/world/europe/29iht-blind.4.8100944.html.

Bisot, V., Serizel, R., Essid, S., Richard, G., 2016. Acoustic scene classification with matrix factorization for unsupervised feature learning. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016. IEEE, Shanghai, China, pp. 6445–6449. https://doi.org/10.1109/ICASSP.2016.7472918. March 20-25, 2016.

Defossez, A., Synnaeve, G., Adi, Y., 2020. Real time speech enhancement in the waveform domain. In: Interspeech.

Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M., 2017. Audio set: an ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017. IEEE, pp. 776–780. https://doi.org/10.1109/ICASSP.2017.7952261.

Guzhov, A., Raue, F., Hees, J., Dengel, A., 2021. Audioclip: Extending Clip to Image, Text and Audio arXiv:2106.13043.

Hornauer, S., Li, K., Yu, S.X., Ghaffarzadegan, S., Ren, L., 2021. Unsupervised discriminative learning of sounds for audio event classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021. IEEE, Toronto, ON, Canada, pp. 3035–3039. https://doi.org/10.1109/ICASSP39728.2021.9413482. June 6-11, 2021, L.

Ikram, S., Malik, H., 2010. Digital audio forensics using background noise. In: Proceedings of the 2010 IEEE International Conference on Multimedia and Expo, ICME 2010, 19-23 July 2010. IEEE Computer Society, Singapore, pp. 106–110. https://doi.org/10.1109/ICME.2010.5582981.

Jia, Y., Zhang, Y., Weiss, R.J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Lopez-Moreno, I., Wu, Y., 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. Adv. Neural Inf. Proc. Syst. 31: Ann. Conf. Neural Inf. Proc. Syst. 4485–4495. NeurIPS 2018, December 3–8, 2018, Montréal, Canada.

Lohrasbipeydeh, H., Gulliver, T.A., 2021. Rssd-based mse-sdp source localization with unknown position estimation bias. IEEE Trans. Commun. 69 (12), 8416–8428. https://doi.org/10.1109/TCOMM.2021.3112583.

Luo, Y., Chen, Z., Yoshioka, T., 2020. Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020. IEEE, Barcelona, Spain, pp. 46–50. https://doi.org/10.1109/ICASSP40776.2020.9054266. May 4-8, 2020.

Morelle, R.. The Hum that Helps to Fight Crime. https://www.bbc.com/news/science-environment-20629671 (Dec 2012.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: an ASR corpus based on public domain audio books. ICASSP 2015. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, South Brisbane, Queensland, Australia, pp. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964. April 19-24, 2015.

Piczak, K.J., 2015. ESC: dataset for environmental sound classification. MM '15. In: Zhou, X., Smeaton, A.F., Tian, Q., Bulterman, D.C.A., Shen, H.T., Mayer-Patel, K., Yan, S. (Eds.), Proceedings of the 23rd Annual ACM Conference on Multimedia Conference. ACM, Brisbane, Australia, pp. 1015–1018. https://doi.org/10.1145/2733373.2806390. October 26 - 30, 2015.

Ren, Z., Qian, K., Wang, Y., Zhang, Z., Pandit, V., Baird, A., Schuller, B.W., 2018. Deep scalogram representations for acoustic scene classification. IEEE CAA J. Autom. Sinica 5 (3), 662–669. https://doi.org/10.1109/JAS.2018.7511066.

Saki, F., Kehtarnavaz, N., 2014. Background noise classification using random forest tree classifier for cochlear implant applications. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014. IEEE, Florence, Italy, pp. 3591–3595. https://doi.org/10.1109/ICASSP.2014.6854270. May 4-9, 2014.

Salamon, J., Jacoby, C., Bello, J.P., 2014. A dataset and taxonomy for urban sound research. In: 22nd ACM International Conference on Multimedia (ACM-MM'14), pp. 1041–1044. Orlando, FL, USA.

Singh, J., Joshi, R., 2019. Background sound classification in speech audio segments. In: 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pp. 1–6. https://doi.org/10.1109/SPED.2019.8906597.

Takahashi, N., Parthasaarathy, S., Goswami, N., Mitsufuji, Y., 2019. Recursive speech separation for unknown number of speakers. In: Kubin, G., Kacic, Z. (Eds.), Interspeech 2019, 20th Annual Conference of the International Speech Communication Association. ISCA, Graz, Austria, pp. 1348–1352. https://doi.org/10.21437/Interspeech.2019-1550, 15-19 September 2019.

Thorogood, M., Fan, J., Pasquier, P., 2015. Bf-classifier: background/foreground classification and segmentation of soundscape recordings. In: Proceedings of the Audio Mostly 2015 on Interaction with Sound. AM '15, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2814895.2814926.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, pp. 5998–6008. Long Beach, CA, USA.

Vincent, E., Gribonval, R., Fevotte, C., 2006. Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. 14 (4), 1462–1469. https://doi.org/10.1109/TSA.2005.858005.