







Article

Binaural Acoustic Scene Classification Using Wavelet Scattering, Parallel Ensemble Classifiers and Nonlinear Fusion

Vahid Hajhashemi ¹, Abdorreza Alavi Gharahbagh ¹, Pedro Miguel Cruz ², Marta Campos Ferreira ¹, José J. M. Machado ³ and João Manuel R. S. Tavares ^{3,*}

¹ Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal; hajhashemi.vahid@yahoo.com (V.H.); abalavi.gh@gmail.com (A.A.G.); mferreira@fe.up.pt (M.C.F.)

² Bosch Security Systems S.A., EN109-Zona Industrial de Ovar, 3880-080 Ovar, Portugal; pedro.cruz4@pt.bosch.com

³ Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal; jjmm@fe.up.pt

* Correspondence: tavares@fe.up.pt; Tel.: +351-22-041-3472

Abstract: The analysis of ambient sounds can be very useful when developing sound base intelligent systems. Acoustic scene classification (ASC) is defined as identifying the area of a recorded sound or clip among some predefined scenes. ASC has huge potential to be used in urban sound event classification systems. This research presents a hybrid method that includes a novel mathematical fusion step which aims to tackle the challenges of ASC accuracy and adaptability of current state-of-the-art models. The proposed method uses a stereo signal, two ensemble classifiers (random subspace), and a novel mathematical fusion step. In the proposed method, a stable, invariant signal representation of the stereo signal is built using Wavelet Scattering Transform (WST). For each mono, i.e., left and right, channel, a different random subspace classifier is trained using WST. A novel mathematical formula for fusion step was developed, its parameters being found using a Genetic algorithm. The results on the DCASE 2017 dataset showed that the proposed method has higher classification accuracy (about 95%), pushing the boundaries of existing methods.

Keywords: urban sounds classification; stereo signal; sound base intelligent system; machine learning; genetic algorithm



Citation: Hajhashemi, V.; Gharahbagh, A.A.; Cruz, P.M.; Ferreira, M.C.; Machado, J.J.M.; Tavares, J.M.R.S. Binaural Acoustic Scene Classification Using Wavelet Scattering, Parallel Ensemble Classifiers and Nonlinear Fusion. *Sensors* **2022**, *22*, 1535. <https://doi.org/10.3390/s22041535>

Academic Editor: Paolo Mercorelli

Received: 24 December 2021

Accepted: 14 February 2022

Published: 16 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The analysis of ambient sounds can be very useful when developing sound base intelligent systems. In the last few years, sound based intelligent systems have received a lot of attention in indoor and outdoor scenarios. Some possible applications of these systems are smart devices and phones, robotics, data archiving, surveillance, security systems, and hearing aids. Acoustic scene classification (ASC) is defined as identifying the area of a recorded sound or clip among some predefined scenes [1]. ASC is a subset of algorithms and systems for audio understanding by machine learning audio based algorithms, i.e., computer audition (CA).

Computer audition systems attempt to suggest intelligent algorithms to extract meaningful information from audio data [2]. ASC is a preprocessing step in some of these systems that attempt to identify the scene of audio data, e.g., airport, park and subway, just to name a few. In many CA applications, the background audio of real-time speech/music should be discarded in the preprocessing stage or used for environmental noise assessment [3]. ASC helps CA systems to limit the search area, select a denoising strategy and enhance overall accuracy. ASC systems have many challenges, one of them is the different type of inputs, since the quality of the microphones or audio sensors varies, the number of recorded audios from a scene varies, and the sensors can finally be mono (single channel) or stereo (dual channel) [3].

Another challenge is choosing the right inputs for the classifier, which is known as feature extraction and selection. Some of the features used in previous works include Mel frequency cepstral coefficients (MFCC), wavelet, constant-Q transform (CQT), and histograms of oriented gradients (HOG) [4]. An additional problem is to select the best classifier. The Gaussian mixture models (GMMs), hidden Markov models (HMM), support vector machines (SVMs), ensemble learning and deep learning methods, such as convolutional neural network (CNN), are some of the most common used ASC classifiers [5].

The main goal of all ASC systems is to achieve the best classification accuracy with the lowest required quality input, maximum processing speed, and minimum implementation complexity; however, intrinsically, these goals can be contradictory. Low quality input and fast classification typically decrease the classification accuracy. The hybrid and complex systems usually are more accurate, especially in real world applications. Based on the aforementioned, the main goal of this research was to achieve the best classification accuracy and, at the same time, to increase the classifier training speed.

Due to the stereophonic nature of the used audio recordings, two audio channels were used to train different probabilistic classifiers. The outputs of the studied classifiers were then used in a novel nonlinear mathematical formula, which was optimized by a genetic algorithm to achieve the best possible accuracy. The use of two channels to train different classifiers, and the proposed fusion scheme based on a nonlinear transformation to combine the output of the used classifiers in order to obtain the final decision, are the main novelties of this research. The following of this article is organized as follows: Section 2 presents the related background of ASC. Section 3 presents the proposed method and gives details of the used dataset. Sections 4 and 5 present the results and their discussion, respectively. Finally, Section 6 presents the main conclusions of this research.

2. Related Work

ASC research has been focused on two main areas, i.e., feature extraction and classification schemes. The classification schemes were divided into ensemble based methods and other classification schemes, such as those based on deep learning, which can be used to highlight the difference between previous works.

2.1. Feature Extraction and Preprocessing

There are many audio features that can be used in ASC. Short-Term Fourier Transform (STFT) is one of the public features used in acoustic research. This kind of feature was used for scene classification [4,6,7] or in a preprocessing step such as in [8]. Usually, STFT is not used individually and its features are combined with other features, such as Mel Frequency Cepstral Coefficients (MFCC), Mel-frequency cepstrum (MFC) and log-Mel spectrogram instead. MFCCs are coefficients of MFC that independently or in combination with STFT and wavelet can be used for audio processing [9]. Mel-frequency cepstrum is the logarithmic power spectrum of the linear cosine transform of short-term audio signals in a nonlinear scale of frequency, usually known as Mel. In the Mel spectrogram, the frequencies are transformed to a nonlinear scale, similar to the human auditory system response, i.e., the Mel scale. On the other hand, a spectrogram is an image related to the spectrum of signal (audio) frequencies. Mel based features, such as log-Mel spectrogram, Mel-frequency cepstrum, MFCC, log-Mel delta, and delta-delta, are among the most commonly used features in ASC. For example, the Log-Mel spectrogram has been used in [8,10–22], with differences between parameters such as filter banks, STFT and windowing function.

In some research, log-Mel spectrogram is used after some processing or in combination with other features. For example, Alamir [23] used log-Mel spectrogram and wavelet scattering, Wu and Lee [24] used audio framing and log-Mel spectrogram, and Log-Mel spectrogram image after median filtering was used in [25]. Log-Mel spectrogram clustered by k means [26] and Log-Mel spectrogram and Gammatonegram (Gamma) [27,28] are other types of Mel based features that have been used in ASC. Some researchers showed that the first and second temporal derivatives of log-Mel spectrogram are good ASC

descriptors [29–33]. These first and second log-Mel spectrogram derivatives are known as delta and delta–delta features. Log-Mel energies [4], log-Mel filter bank (LMFB) [34,35], log-Mel band energies and Single Frequency Filtering Cepstral Coefficients (SFFCC) [36] and MFCC and log-Mel filter bank [37] are other types of Mel-based features that have been used for ASC.

Lostanlen and Andén used wavelet scattering features for ASC [38]; this family of feature is comparable to MFCC. Raw audio data without any preprocessing have been used for ASC in [39,40].

2.2. Classification Schemes

2.2.1. Ensemble Methods

Nguyen et al. [41] proposed a CNN ensemble ASC method for tasks of the DCASE 2018 challenge. The authors combined the output probabilities of CNNs as ensembles of CNNs to improve ASC accuracy. Jung et al. [42] proposed an ensemble model that extracts some audio features. The authors trained several deep neural networks (DNNs) in parallel, and the scores from DNN classifiers were applied to a score-level ensemble block to make the final result. Singh et al. [43] combined deep convolutional neural network (DCNN) scores in a score-level ensemble step and made the final output. Sakashita and Aono [44] trained nine neural networks and used their outputs as an input of an ensemble learning block in order to increase accuracy. Jiang et al. [11] used an ensemble learning method to make final decisions using the output of CNNs. Mars et al. [45] improved the result of a ASC system by including two distinct and light-weight architectures of CNNs, by using an ensemble of the output of CNNs. Huang et al. [46] used the ensemble of four different CNNs and improved the final result by about 4%. Wang et al. [47] used a CNN ensemble for voting of all scores of CNNs and improved the classification result. Ding et al. [48] applied a composed of two CNN and GMM scores to an ensemble system to enhance accuracy. Xu et al. [49] ensembled deep classifiers that trained using different features with the goal of using complementary information features.

Gao et al. [50] ensembled the output of trained CNN networks on different representations to boost classification performance. Wang et al. [51] trained 5-layer or 9-layer CNNs with average pooling using features conveying complementary information. The authors developed several ensemble methods to integrate the outputs of the CNNs such as random forests and extremely randomized trees. Other research, such as in [52–54], used ensemble learning to integrate different classifiers that trained using different features in order to increase ASC accuracy. Sarman and Sert [55] used two ensemble methods, mainly bagging and random forest, to overcome the imbalance problem of data with minimum computational cost. The authors classified violent scenes based on audio signal. Alamir proposed a hybrid method which includes a CNN and an ensemble classifier in a parallel form [23], the features used by the classifiers being different.

Based on the above review, one can conclude that many researchers used ensemble learning to integrate the result of a set of trained classifiers in a fusion or post-processing step in order to increase ASC accuracy. Only a few researchers used ensemble learning as a primary classification step. Based on our best knowledge, the results of this research are usually equal to or less than similar ASC classification methods, mainly, based on deep learning.

2.2.2. Deep Learning Methods

ASC algorithms mostly use CNN based architectures in the classification step. For example, a Multi-task Conditional Atrous CNN (CAA-Net) is used in [2]. Liu et al. [4] used CNNs as main learners and an extra random forest method for final classification. Visual Geometry Group CNN (VGG net) has been suggested as a good architecture for ASC in some research [5,11,21]. Naranjo-Alcazar suggested a VGG-style CNN where convolutional blocks were replaced with residual squeeze-excitation blocks [5]. Jiang et al. used twelve VGG style CNNs in the first step of their method [11]. Another simplified

VGGNet-InceptionNet architecture was suggested in [21]. Vilouras implemented CNNs and concluded that two modified Resnet, including “shake–Shake” regularization and squeeze–excitation block, had better accuracy [8].

McDonnell selected a residual network pre-activated CNN and rounded the layer values to reduce memory usage [10]. A modified SegNet [12], fine-resolution CNN (FR-CNN) [13] and a multi-scale feature fusion CNN [14] are other types of modified CNNs that have been used for ASC. The generative adversarial neural networks (GAN) [15], CNN with cross-entropy (CE) as loss function [16], CNN including a semantic neighbors over time (SeNoT) module [17], optimized CNNs [18–20] and conditional autoencoders [22] are among the deep learning methods used for audio scene classification. All of the above research has a similar feature: the use of log-Mel spectrogram. A one-dimensional, CNN, comprised of multiple convolutional/pooling layers followed by fully-connected layers, was suggested in [24]. A modified 2D CNN, long short-term memory (LSTM) and VGG16 CNN with a processed log-Mel spectrogram were suggested in [25–27], respectively.

An encoder–decoder network [27] and Multi-kernel CNN-DNN architecture [28] with log-Mel spectrogram, gamma and CQT are other suggested CNN based classification schemes. Three modified CNN [30,32,33], Resnet based CNN [29] and Four-pathway residual CNNs [31], which use log Mel spectrograms, delta and delta–delta features, are other proposed classification methods. Optimized CNN [4], Fully CNN [34], Light CNN (LCNN) [35], DNN [36,56], VGG16 based CNN [37], SoundNet [39] and Front-end DNN + SVM [40] are among different classification schemes that have been used with hybrid features or RAW data in ASC. Based on the above review, one can conclude that many CNNs, DNNs and other deep learning classification methods have been suggested for ASC. Only a few recent research works have used ensemble learning, SVM, Fuzzy C-means clustering, Adaptive Neuro-Fuzzy Inference System and Naïve–Bayes, as classification schemes. Their need for many training samples, sensitivity to learning parameters, execution time and memory usage are among the main deep learning problems. According to these challenges, a hybrid ensemble learning based method is suggested that overcomes the state-of-the-art deep learning methods. The suggested scheme is trained with fewer samples than the required by the traditional deep learning methods. In the meantime, the sensitivity of the suggested method to training parameters, training time and memory usage is lower than that of the common deep learning based methods.

3. Wavelet Scattering

In MFC, high-frequency spectrogram coefficients are not stable to time-warping, which means that two signals have a similar form but vary in speed. The MFCC scheme stabilizes these coefficients by averaging them along with Mel frequency. The averaging process dictates some loss of information. A scattering transform recovers the MFCC lost information in averaging with a cascade of wavelet decompositions and modulus operators [57]. It is stable to time warping deformation and locally translation invariant. WST, introduced in [57,58], is an accurate representation based on the iterative wavelet transform modulus. It has been applied to different signal classification tasks, such as synthetic aperture radar [59], speaker identification [60] and ASC [38]. The three main steps of WST are: wavelet filter bank, modulus, and averaging, as depicted in Figure 1.

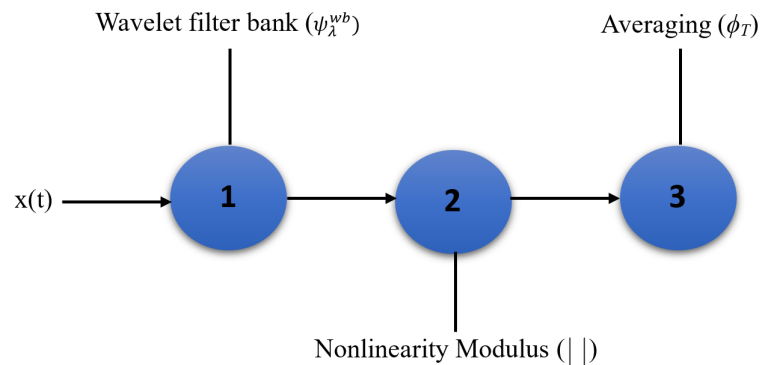


Figure 1. Main steps of WST, where $x(t)$ is the input signal, ψ_λ^{wb} the wavelet filter bank in λ centre frequency, and ϕ_T is the averaging in window with width T .

3.1. Wavelet Filter Bank

In wavelet scattering, filter banks are used to assure important properties that are essential to the implementation of WST. These properties include reconstruction and orthogonality as normal wavelet properties and special passband to satisfy the WST concept. First, for signal $x(t)$, the Fourier transform $X(\omega)$ is defined as:

$$X(\omega) = \int_R x(t)e^{-i\omega t} dt, \quad (1)$$

where R is the total time duration of x , i the imaginary unit, and ω the angular frequency, respectively. An analytic mother wavelet, i.e., Gammatone function, was chosen as a bandpass complex filter ($\psi(t)$). The Gammatone function is a sinusoid function modulated by a gamma distribution function [61]:

$$\psi(t) = t^{(N-1)}e^{-st}u(t), \quad s = \alpha - i\omega_c, \quad (2)$$

where t is the time, α the effective duration of ψ , i.e., the filter bandwidth, ω_c the filter centre frequency, $u(t)$ denotes the unit step function, and N is the filter order that determines the transition bands of the filter. In acoustic applications, the Gammatone filter order, typically settled in the range of [3..5], provides a good approximation to the human ear. In this research, N in the first wavelet bank was set to 4 based on [38,62]. It can be easily proved that the Fourier transform of $\psi(t)$ is [63,64]:

$$\Psi(\omega) = \frac{(N-1)!}{(\alpha + j(\omega - \omega_c))^N}. \quad (3)$$

The suggested filter bank should remove all the signal DC components because these components have no data. In (3), the signal does not have this property, so the Gammatone function's derivative that introduces a zero at $\omega = 0$ is suggested as filter bank functions. The Fourier transform of the time derivative of Gammatone function is:

$$\Psi'(\omega) = j\omega \frac{(N-1)!}{(\alpha + j(\omega - \omega_c))^N}. \quad (4)$$

It can be seen that $\Psi'(0) = 0$. Hence, the corresponding form of the filter in time domain is:

$$\begin{aligned} \psi'(t) &= (N-1)t^{(N-2)}e^{-st}u(t) - st^{(N-1)}e^{-st}u(t) \\ &= ((N-1) - st)t^{(N-2)}e^{-st}u(t), \\ s &= \alpha - i\omega_c. \end{aligned} \quad (5)$$

This form can be used to define a bank of bandpass filters. For simplicity, it is used the following notations:

$$\begin{aligned}\Psi_{\lambda}^{wb}(\omega) &= \Psi'(\omega), \\ \psi_{\lambda}^{wb}(t) &= \psi'(t),\end{aligned}\quad (6)$$

where $\psi_{\lambda}^{wb}(t)$ and $\Psi_{\lambda}^{wb}(\omega)$ are the time and Fourier transform of the wavelet filter bank with $\lambda = \omega_c$ centre frequency.

Slaney [65] described each ω_c filter center frequency as:

$$\omega_c(k) = 2\pi \left(-C + e^{\frac{k}{K} \log\left(\frac{f_{\min}+c}{f_{\max}+c}\right)} \cdot (f_{\max} + C) \right), \quad (7)$$

subject to:

$$\log_2\left(\frac{f_{\min}}{f_{\max}}\right) > \frac{1}{NV},$$

where $1 \leq k \leq K$, K is the total number of bank filters, C is a constant, f_{\min} and f_{\max} are lowest and highest pass band frequencies of the filterbank, and NV is the Number of Voices per octave. The value of f_{\max} must be less than or equal to half of the sampling frequency. The lowest-frequency interval is covered with about equally-spaced filters with constant frequency bandwidth to guarantee that the filter bank covers all frequencies. For simplicity, these filters are still identified as wavelets. The frequency response of filters is given in Figure 2. The narrow band of filters in lower Cycles/Sample shows the inherent property of WST to extract more details in low bands.

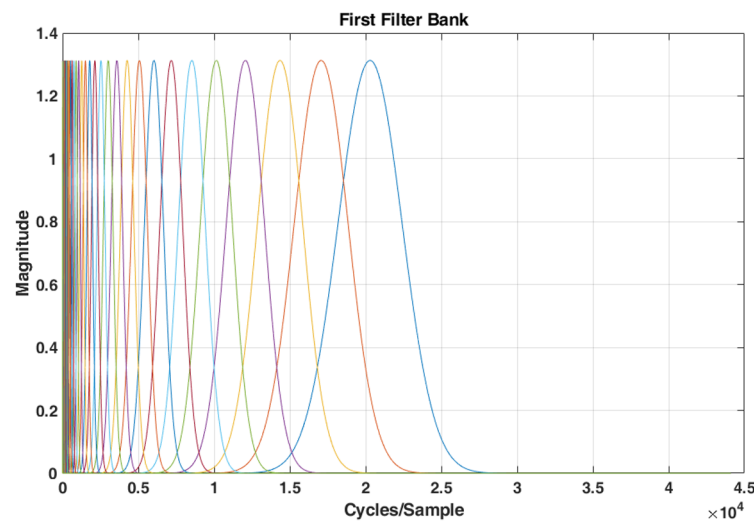


Figure 2. The frequency response of a typical wavelet filter bank with $N = 47$.

3.2. Nonlinearity Modulus

The output of wavelet filter banks can be calculated by

$$y(t, \log \lambda) = x(t) * \psi_{\lambda}^{wb}(t), \quad \text{for all } \lambda \in \Lambda, \quad (8)$$

where $x(t)$ is the input signal, $\psi_{\lambda}^{wb}(t)$ is defined as in Equation (6), $*$ represents the convolution operator, and the set of all filter bank centre frequencies is denoted by Λ . The modulus of filters is defined by:

$$x_1(t, \log \lambda) = |y(t, \log \lambda)|, \quad \text{for all } \lambda \in \Lambda, \quad (9)$$

where $||$ is the modulus, i.e., amplitude, of a complex number, which is the wavelet scalogram. $x_1(t, \log \lambda)$ is a 2D time-frequency representation of the filter bank output, and the wavelet scalogram is a visual representation of this output, which depicts the intensity

of $x(t)$ at time t and λ log frequency. The scalogram of an acoustic scene for wavelet filter bank of Figure 2 is shown in Figure 3.

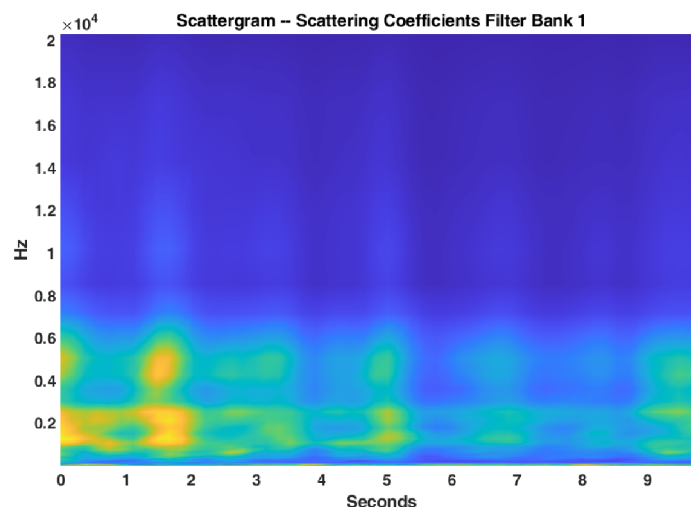


Figure 3. Scalogram of an acoustic scene taking into account the wavelet filter bank of Figure 2.

The scalogram is not time shift-invariant and does not have stability properties. To achieve this, the last step of Figure 1, i.e., averaging step, is necessary.

3.3. Averaging

Three operators have been used for making a time shift invariant system in the last step of Figure 1. $x_1(t, \log \lambda)$ is passed from a ϕ_T low-pass filter, which applies a time averaging to the lowest-frequency interval filters, as:

$$S_1 x_1(t, \log \lambda) = \langle x_1(t, \log \lambda) \rangle_t = x_1(t, \log \lambda) * \phi_T(t), \quad (10)$$

where $\langle \rangle_t$ is a notation for time averaging using ϕ_T and $*$ denotes the convolution integral. The S_1 output is approximately equal to MFCC, and is commonly known as first order scattering coefficients [58]. This averaging eliminates some high frequencies details. To recover high frequencies details, $x_1(t, \log \lambda)$ is applied to the second wavelet filter bank $\psi_\gamma^{wb}(t)$. Similarly to Equations (8)–(10), making some assumptions, one can have:

$$x_2(t, \log \lambda, \log \gamma) = \left| x_1(t, \log \lambda) * \psi_\gamma^{wb}(t) \right|, \quad (11)$$

$$S_2 x_2(t, \log \lambda, \log \gamma) = x_2(t, \log \lambda, \log \gamma) * \phi_T(t). \quad (12)$$

The second filter bank centre frequencies must be different from the first filter bank. The ϕ_T low-pass filter is similar in Equations (10) and (12). S_2 is known as the second-order time scattering coefficients. The final feature vector is a combination of S_1 and S_2 :

$$Sx = [S_1 x_1(t, \log \lambda), \quad S_2 x_2(t, \log \lambda, \log \gamma)]. \quad (13)$$

Although the above formulas are continuous in time and frequency, the t , λ , and γ parameters can be discretized without considerable loss.

4. Proposed Method

4.1. Training Scheme

Figure 4 shows the block diagram of the training scheme of the proposed method. The two channels of input stereo signal are first decomposed to left (A) and right (B) channels. Feature extraction is then performed by applying wavelet scattering to A and B channels separately. Two ensemble classifiers are trained for the channels. The differences, such as

delays and noises, between channels lead to different ensemble classifiers that are important in the proposed step. The output of each classifier (A and B) is sent to the fusion step. In the training phase of the fusion step, a genetic algorithm (GA) finds the coefficients of a nonlinear transform to maximize the accuracy of the final result. In fact, GA creates a combination of the two classifier outputs (after their training process) nonlinearly in order to increase the final accuracy. The proposed training scheme was implemented in the Matlab R2020b software package.

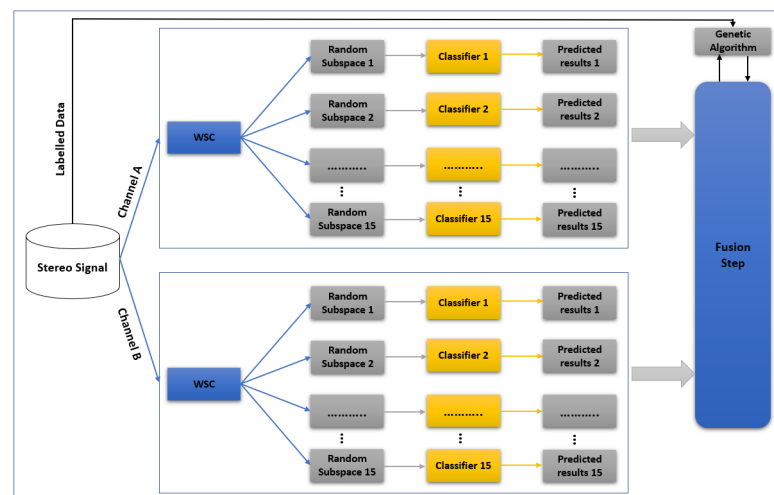


Figure 4. Block diagram of the training scheme of the proposed method.

4.1.1. Feature Extraction

WST has been used for feature extraction as a good representation of acoustic scenes (Figure 5). The N term, i.e., the order of the wavelet filter or quality factor, presented in Equations (4) and (5), in the first wavelet bank ($M = 1$ in Figure 5) was set to 4, and in the second wavelet filter bank ($M = 2$ in Figure 5) was set to 1 (one). The invariance scale was chosen equal to 0.75 s. Using these parameters, the number of filter banks in the first step ($N = 4$) was 47, and in the second step ($N = 1$) was 13. The highest and lowest filter bank centre frequencies of the first step were 20296 and 3.7 Hz, and for second filter bank were 16,537.5 and 4 Hz, respectively. In a 10 s sound signal with $F_s = 44,100$ Hz, the size of the extracted feature is 290×54 , where 290 is the number of resolutions across all orders of the scattering transform, and 54 is the resolution of the scattering coefficients. Quality factors and invariance scale values were chosen based on [23,38].

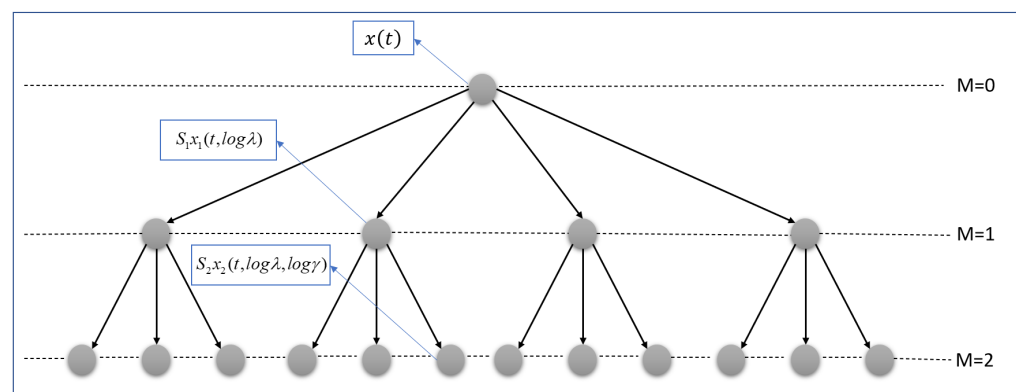


Figure 5. Hierarchical form of WST at first and second filter banks of the proposed method.

4.1.2. Ensemble Classifiers

For the classification step, the random subspace method was selected. The random subspace method also called attribute bagging, i.e., feature bagging, is a modified ensemble

classifier. Ensemble learning methods typically combine several weak learners in order to make a classifier that works better than the original learners. The random subspace method is an ensemble learning method except that it only uses some training features, which are randomly selected from the training feature set, and are changed for each learner. The random subspace method scheme trains individual learners without over focusing on highly predictive features. For this reason, random subspaces are known as a good choice for acoustic scene classification problems, mainly where the number of features is larger than the number of training data. The random subspace method can be implemented via parallel learning, so it is suitable for fast learning, which is desirable in ASC. The suggested classifier is a systematic construction of a decision forest that relies on a pseudorandom procedure to select each weak learner's training features. Each weak learner, i.e., decision-tree, generalizes classification by invariances in the excluded features [66]. Finally, the results of the weak learners are combined by averaging the posterior probabilities.

4.1.3. Fusion Step

The proposed fusion step uses two ensemble classifiers output and decides about the scene class. For the fusion step, a nonlinear transform is proposed that must satisfy the following conditions:

- If the result of the classifiers are similar, the value of Fusion result leads to the maximum in this result;
- If the result of the classifiers are different, Fusion result should be maximized in a true class.

To fulfil the above conditions, the following structure was adopted:

- The summation of each class probabilities in two classifiers;
- The summation of class probabilities square in two classifiers;
- The multiply of class probabilities in two classifiers;
- The absolute value of the difference of each class probabilities in two classifiers.

The mathematical formulation of this weighted nonlinear function is:

$$\text{Fusion Result}_i = \alpha(x_{i,A} + x_{i,B}) + \beta(x_{i,A}^2 + x_{i,B}^2) + \gamma(x_{i,A} \times x_{i,B}) + \lambda|x_{i,A} - x_{i,B}| \quad (14)$$

where $x_{i,A}$ and $x_{i,B}$ are scores assigned using ensemble classifiers to class i , belong to A and B channels, respectively. Obviously, if the result of the classifiers are similar, the value of Equation (14) leads to the maximum and α , β , γ , λ are the parameters that should be found to maximize the accuracy in cases where the classifiers' results are different.

A genetic algorithm was used to find the best values of α , β , γ , and λ parameters in order to maximize the accuracy of the train set. To avoid unreal answers, the GA search space for α , β , and γ were limited between $[0... 3]$. The true value of λ should be negative because the difference between classifier scores negatively affects the true class. In other words, the classifier scores in true scene should be approximately similar, and the absolute scores' difference should be minimized. Therefore, the search space for λ was limited between $[-3... 0]$. The optimization toolbox from Matlab software was used for implementing GA, and the final values of the parameters were obtained as: $\alpha = 1.6406$, $\beta = 2.8792$, $\gamma = 1.8059$, and $\lambda = -1.5274$.

4.2. Test Scheme

In the test scheme of the proposed method, Figure 6, the extracted WST features from A and B channels, are applied to train the classifiers. The output of each ensemble classifier, which includes a vector of scores assigned to each class, is used as input of the fusion step. The fusion step calculates (Equation (14)) and chooses the maximum value as final output.

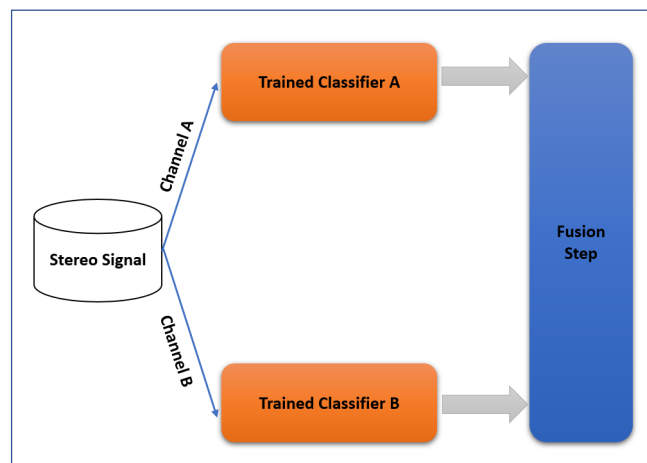


Figure 6. Block diagram of the test scheme of the proposed method.

4.3. Evaluation Methodology

A dataset from the TUT Acoustic Scenes 2017 (DCASE 2017 challenge) [67] was chosen for the evaluation of the proposed method. This database is a free publicly accessible database, and has been used in most ASC recent research [1,23,27,68–82]. The TUT Acoustic Scenes 2017 dataset consists of two subsets: development set and evaluation set, Table 1. The development dataset that is used for the training phase consists of the complete TUT Acoustic Scenes 2016 dataset [83]. The recorded data were divided into subsets based on the original recordings’ location, so the DCASE 2017 evaluation set contains recordings of similar scenes, but from different locations. For each scene, there are 312 audio segments with equal length (10 s). This dataset includes 15 different well balanced classes.

Table 1. Details of the TUT Acoustic Scenes 2017 (DCASE 2017 challenge) ASC dataset.

	Development	Evaluation
Number of Files	4680	1620
Number of Classes	15	
Duration per audio signal	10 s	
Data Format	44.1 kHz, 24 bit resolution, Binaural stereo wave files	
Location	Dataset was recorded in different cities, including London and Paris.	
Device	Roland Edirol R-09 wave recorder	
Task	Acoustic Scene Classification	

As aforementioned, the TUT Acoustic Scenes 2017 dataset contains 15 different, well balanced, classes:

- Bus-travelling by bus in the city (vehicle);
- Cafe/Restaurant—small cafe/restaurant (indoor);
- Car-driving or travelling as a passenger, in the city (vehicle);
- City centre (outdoor);
- Forest path (outdoor);
- Grocery store -medium size grocery store (indoor);
- Home (indoor);
- Lakeside beach (outdoor);
- Library (indoor);
- Metro station (indoor);
- Office-multiple persons, typical workday (indoor);
- Residential area (outdoor);

Train (travelling, vehicle);
Tram (travelling, vehicle);
Urban park (outdoor).

The proposed method was trained using the development dataset and tested on the evaluation dataset. As in many other works, the confusion matrix was considered an interesting criterion for comparing the proposed method with the existing state-of-the-art methods. As another evaluation criterion, the total accuracy of the proposed method was compared to the ones achieved by some related methods. The total accuracy was calculated by dividing the number of true classifications to the number of total test data.

5. Results and Discussion

5.1. Results of Ensemble Classifiers and of the Proposed Method

The confusion matrices of the ensemble classifiers separately are given in Figures 7 and 8. The total system confusion matrix, i.e., as to the result after the fusion step, is given in Figure 9. All studied methods showed good results as to the development data. A good model has good accuracy in all classes at the evaluation phase, so these matrices were reported here for the evaluation dataset. The suggested mono ensemble classifiers are fully aligned with Alamir [23] results, which confirms its correct implementation. The differences between mono channels cause differences between ensemble classifiers, Figures 7 and 8, which are relevant in the fusion step. Figure 9 presents the result after the fusion step for the evaluation dataset that shows the correctness of the fusion step of the proposed method. The values in the main diagonals of the matrices in Figures 7–9 are the number of correct classifications in the class that belong to that row or column. The other elements are false classifications. In Figures 7–9, the main diagonals (correct classifications) are identified by blue, while other elements (false classifications) are in orange. Bolder cells have higher values. (For more details of used colours in Figures 7–9, the reader is suggested to see the web version of [23]). The results in Figure 9 confirm the efficiency of the novel fusion step. The main improvements of the novel fusion step are regarding the “residential area” (about 100 false classifications were corrected), “café restaurant” (about 40), “metro station” (about 40), and “home” (about 30 false classifications were corrected) classes, i.e., scenes. In some cases, that one channel was so poor or noisy and the other was good, the fusion step accuracy was lower than of the one of good channel, but intrinsically was better than the one of the poor channel. For example, in the “beach” scene, the number of correct classification that belongs to the A channel was 26 (Figure 7), and, for B channel, it was 58. This difference shows that, in this case, A channel was noisy, and B channel was good. The result after the fusion step was 38, which indicates that the proposed fusion step increased the accuracy of the poor channel, but because the information of both channels was used simultaneously in the fusion step, the accuracy of the results in this class was lower than the one of the good channel.

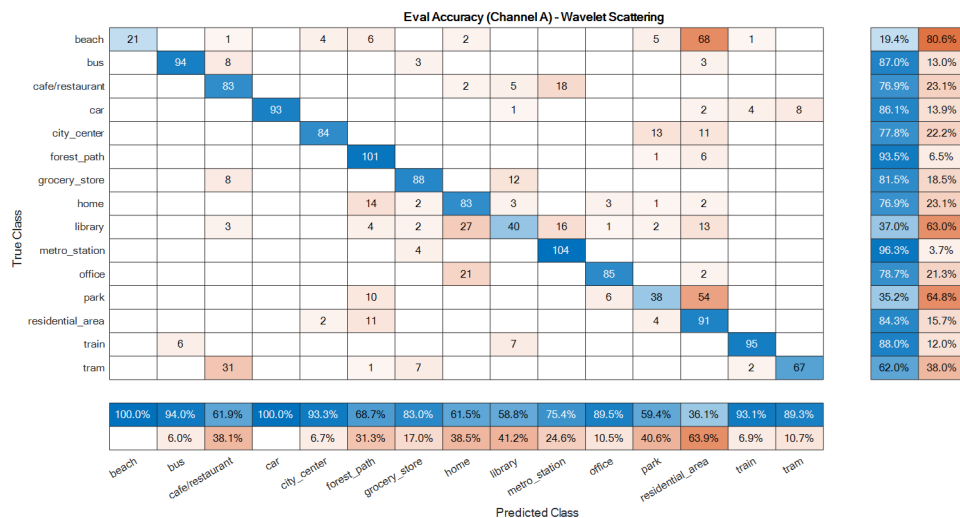


Figure 7. Ensemble classifier confusion matrix for the evaluation data: A channel.

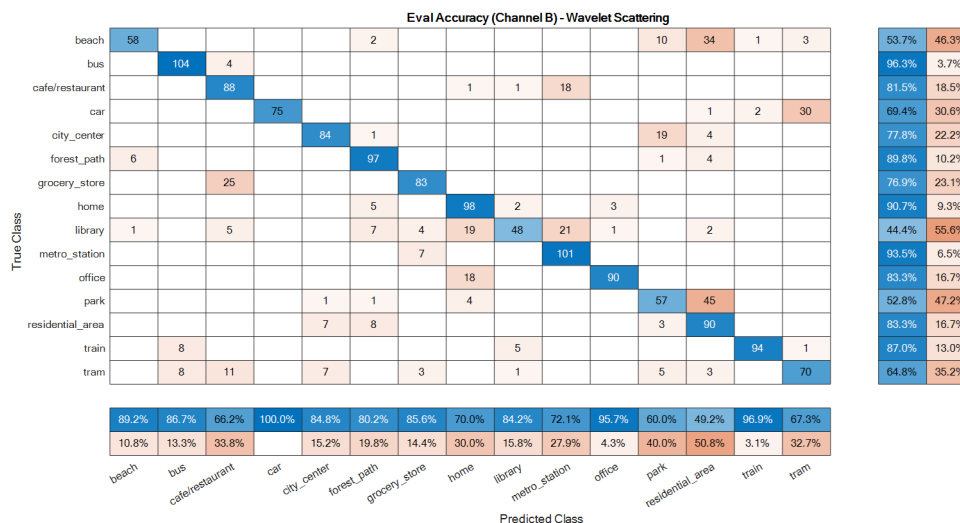


Figure 8. Ensemble classifier confusion matrix for the evaluation data: B channel.

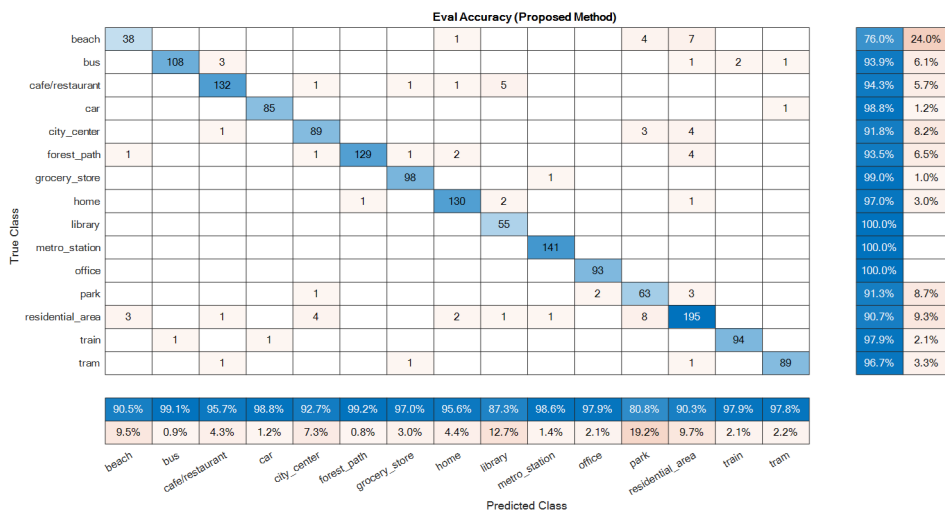


Figure 9. Confusion matrix built for the proposed method as to the evaluation data: stereophonic.

5.2. Sensitivity Analysis

In the sensitivity analysis of the fusion step, the effect of uncertainty, i.e., of α , β , γ and λ parameters (Equation (14)), on the accuracy of the proposed method was analysed. Because of correlation or mutual effect of these parameters on each other, the 1D and 2D uncertainties were studied. In the 1D sensitivity analysis, one of α , β , γ and λ parameters changed at a time, and the others remain constant. In the 2D analysis, two of α , β , γ and λ parameters change at a time, and the other two remain constant. Therefore, four results for the 1D analysis and six results for the 2D analysis were obtained. Figures 10 and 11 show the results for the 1D analysis. In this analysis, the initial values of the parameters were $\alpha = 1.6406$, $\beta = 2.8792$, $\gamma = 1.8059$, and $\lambda = -1.5274$ and, in each curve, only one parameter was changed around its initial value. Figure 10 shows the behaviour of accuracy in terms of α and β parameters. These two parameters are very important because both, with some values, lead to an accuracy equal to 0 (zero). α and β shown similar behaviours, and after a threshold value (-4 for α and -2.5 for β), accuracy increased rapidly and reached approximately 100%. The initial point was tagged on each curve according to a safe distance to the threshold points. Figure 11 shows the behaviour of accuracy in terms of γ and λ parameters, which is considerably different relatively to the behaviour found as to α and β parameters. Based on Figure 11, one can realize that γ and λ had a lower effect on accuracy in comparison to α and β and, in fact, these parameters tune the fusion formula efficiently. These observations were under the assumption that the other parameters were fixed in their initial value, but the results indicate that all parameters after a threshold value have no effect on accuracy. On the (A) curve, due to accuracy value, the tagged point was placed in a good interval, and, on the (B) curve, the initial point was placed at a safe distance to the threshold point, and before the accuracy decrease interval.

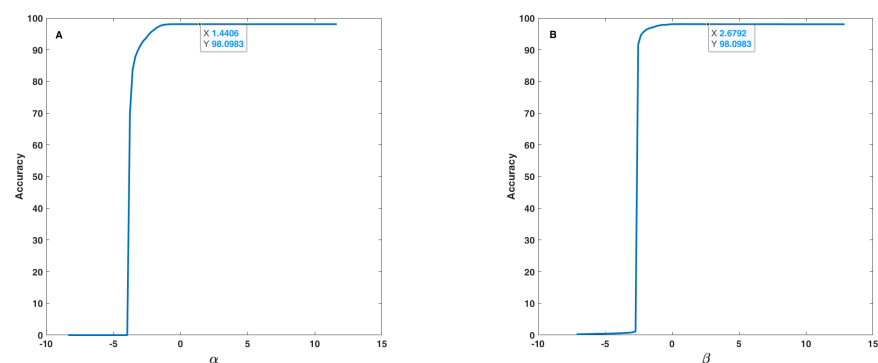


Figure 10. Accuracy in terms of α (A) and β (B) parameters used in the fusion step (Equation (14)).

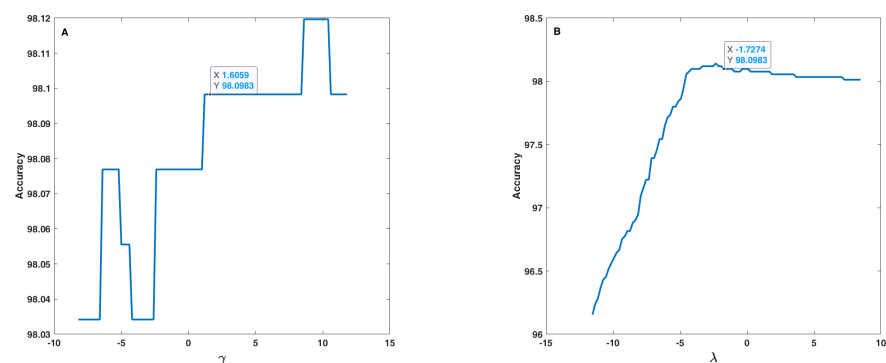


Figure 11. Accuracy in terms of γ (A) and λ (B) parameters used in the fusion step (Equation (14)).

Figures 12–14 show the results obtained as to the 2D sensitivity analysis results. As in the 1D case, the initial values for α , β , γ and λ parameters were equal to 1.6406, 2.8792,

1.8059 and -1.5274, respectively. In each curve, two parameters varied around their initial values and the other two parameters remained constant. Figure 12A shows the behaviour of accuracy in terms of α and β parameters, which have a very similar effect, but based on the 3D plot shown, one can realize that β had higher influence than α . In this figure, the threshold value for changing accuracy from zero to an acceptable value can be defined as a line, between $\alpha \approx 6$, $\beta \approx -7$ and $\alpha \approx -8$, $\beta \approx 7$, which are very different to crisp threshold values found in the case of Figure 10. Figure 12B shows the behaviour of accuracy in terms of α and γ parameters. In each plot of Figure 12, the tagged initial point had a safe distance to the threshold plane where accuracy decreased quickly. Figure 13A shows the behaviour of accuracy in terms of α and λ parameters. The shape of the zero accuracy area is considerably different to the ones of plots of Figure 12, and the threshold value for α in this plot was about -4, which is very similar to the α crisp threshold found for the case of Figure 10A. This similarity suggests that λ has a small effect on α and on accuracy. Figure 13B shows the behaviour of accuracy in terms of β and γ parameters. The observed behaviour is approximately similar to the one depicted in Figure 12B, which indicates the similar effect of α and β parameters. In both plots of Figure 13, the tagged initial point had a safe distance to the threshold plane.

Figure 14A shows the behaviour of accuracy in terms of β and γ parameters, which is similar to the one observed in Figure 13A and confirms once again the similar effect of α and β parameters. Figure 14B lets one conclude that the behaviour of accuracy in terms of λ and γ is different relative to the ones observed for the other cases. In this plot, the lowest accuracy value is around 86%, which indicates that the initial values of α and β parameters keep the accuracy within an acceptable range, and that the effect of λ and γ parameters are lower than the one of α and β parameters. In the meantime, it is possible to conclude that these two parameters can affect accuracy by about 16%, which is considerable and therefore should not be discarded. In addition, in this case, the tagged point had a safe distance to the threshold plane. In conclusion, the sensitivity analysis lets us conclude that the terms including α and β parameters are the main parts of the fusion step, and that the terms including λ and γ parameters can considerably improve the accuracy of the proposed method.

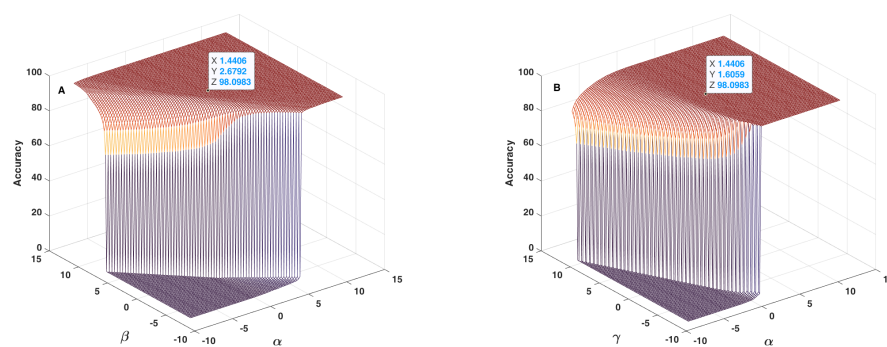


Figure 12. Accuracy in terms of α and β parameters (A) and of α and γ parameters (B) used in the fusion step (Equation (14)).

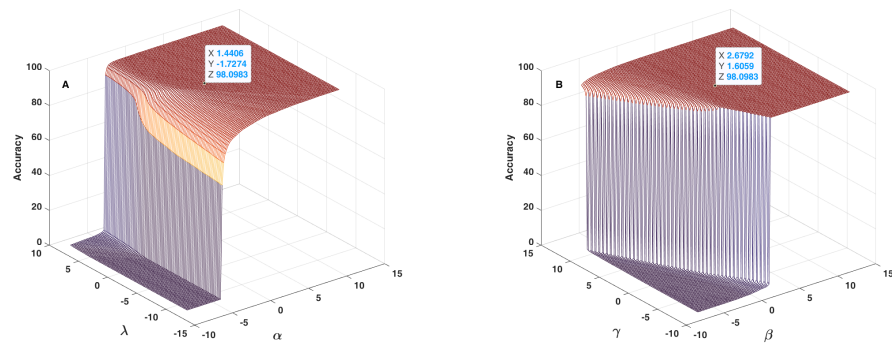


Figure 13. Accuracy in terms of α and λ parameters (A) and of β and γ parameters (B) used in the fusion step (Equation (14)).

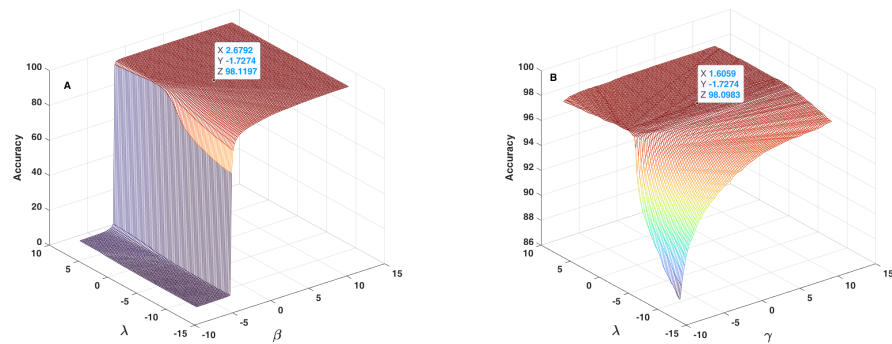


Figure 14. Accuracy in terms of β and λ parameters (A) and of λ and γ parameters (B) used in the fusion step (Equation (14)).

5.3. Result Comparison with Previous Studies

The results of the proposed method were compared with the ones obtained by previous studies that also used the TUT Acoustic Scenes 2017 dataset. Usually, the related methods used Mel based features or raw data. The training phase accuracy of all methods under comparison are generally good, but the suggested method showed a considerable improvement in the test phase. Table 2 presents the total accuracy of the methods as to the evaluation dataset obtained under similar conditions. Hence, the methods were trained using TUT Acoustic Scenes 2017 development dataset, and were tested on evaluation of the TUT Acoustic Scenes 2017 dataset. Therefore, the train and test conditions were fully the same for all methods under comparison, and the values as to total accuracy reported in the literature for the related methods were used in Table 2. The accuracy of the proposed method in the training and test phases was 98.1 and 95%, respectively. These results compared to the state-of-the-art methods under comparison indicates that the suggested fusion scheme can be a potential solution for future ASC systems using stereo audio signals data as input.

Table 2. Results of different ASC methods and of the proposed ASC method on the evaluation set of the DCASE 2017 dataset.

Ref.	Year	Test Accuracy	Detection Approach
[68]	2017	70	Deep residual CNN
[69]	2017	70.6	DNN
[70]	2017	70.6	CNN
[71]	2017	71.7	Recurrent Neural Network (RNN)
[72]	2017	72.6	CNN
[73]	2017	73.8	CNN
[74]	2017	74.1	CNN
[75]	2017	77.7	DCNN
[76]	2017	80.4	CNN
[77]	2017	83.3	GAN
[78]	2018	64	Deep scalogram representations
[79]	2018	69.9	SVM
[80]	2019	69.3	CNN
[81]	2019	75.4	CNN
[82]	2019	77.1	DCNN
[1]	2020	70	SVM
[23]	2021	80	CNN and Ensemble classifiers
[27]	2021	72.6	DNN
Channel A		72.04	One Ensemble classifier
Channel B		76.36	One Ensemble classifier
Our Method		95	Two Ensemble classifiers and Fusion

6. Conclusions

In recent years, sound based intelligent systems such as Acoustic scene classification have received a large amount of attention in different applications. This study has presented a two-step method for ASC in stereo signals, which consists of two ensemble classifiers and a novel fusion step. The proposed robust method uses wavelet scattering transform as a stable, time shift invariant transformation. In the proposed method, firstly, two ensemble classifiers were trained using two channels of stereo input. The output of these classifiers was effectively combined using a nonlinear transform to improve the final classification accuracy. The suggestion of a proper, nonlinear transform that satisfies the fusion step conditions and finding the unknown parameters of this transform using a heuristic method, mainly a genetic algorithm, is the main novelty of the proposed method. The classification accuracy obtained using the proposed system in the DCASE 2017 dataset overcome considerably (at least, a 15% of improvement was achieved) the current state of the art as to this well known public dataset. Based on the obtained results, and on its improvement comparing to other ASC methods, the application of the proposed fusion scheme is suggested for other acoustic classification and detection applications such as acoustic event detection.

Author Contributions: Conceptualization, funding acquisition, and supervision by J.M.R.S.T.; investigation, data collection, formal analysis, and writing—original draft preparation by V.H. and A.A.G.; writing—review and editing by P.M.C., M.C.F., J.J.M.M. and J.M.R.S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This article is a result of the project Safe Cities—“Inovação para Construir Cidades Seguras”, with reference POCI-01-0247-FEDER-041435, cofunded by the European Regional Development Fund (ERDF), through the Operational Programme for Competitiveness and Internationalization (COMPETE 2020), under the PORTUGAL 2020 Partnership Agreement. The first author would like to thank “Fundação para a Ciência e Tecnologia (FCT)”, in Portugal, for his PhD grant with reference 2021.08660.BD.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Waldekar, S.; Saha, G. Two-level fusion-based acoustic scene classification. *Appl. Acoust.* **2020**, *170*, 107502. doi:10.1016/j.apacoust.2020.107502.
2. Ren, Z.; Kong, Q.; Han, J.; Plumbley, M.D.; Schuller, B.W. CAA-Net: Conditional Atrous CNNs with Attention for Explainable Device-robust Acoustic Scene Classification. *IEEE Trans. Multimed.* **2020**, *23*, 10–15. doi:10.1109/tmm.2020.3037534.
3. Abeßer, J. A Review of Deep Learning Based Methods for Acoustic Scene Classification. *Appl. Sci.* **2020**, *10*, 2020. doi:10.3390/app10062020.
4. Liu, Y.; Jiang, S.; Shi, C.; Li, H. Acoustic scene classification using ensembles of deep residual networks and spectrogram decompositions. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE 2019), New York University, NY, USA, 25–26 October 2019.
5. Naranjo-Alcazar, J.; Perez-Castanos, S.; Zuccarello, P.; Cobos, M. Acoustic Scene Classification With Squeeze-Excitation Residual Networks. *IEEE Access* **2020**, *8*, 112287–112296. doi:10.1109/access.2020.3002761.
6. Peeters, G.; Richard, G. Deep Learning for Audio and Music. In *Multi-faceted Deep Learning*; Springer: Cham, 2021; pp. 231–266.
7. Serizel, R.; Bisot, V.; Essid, S.; Richard, G. Acoustic Features for Environmental Sound Analysis. In *Computational Analysis of Sound Scenes and Events*; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 71–101. doi:10.1007/978-3-319-63450-0_4.
8. Vilouras, K. Acoustic scene classification using fully convolutional neural networks and per-channel energy normalization. Technical Report, Detection and Classification of Acoustic Scenes and Events 2020 Challenge; 1 March–1 July 2020.
9. Hajihashemi, V.; Alavigharabagh, A.; Oliveira, H.S.; Cruz, P.M.; Tavares, J.M.R. Novel Time-Frequency Based Scheme for Detecting Sound Events from Sound Background in Audio Segments. In *Iberoamerican Congress on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 402–416. doi:10.1007/978-3-030-93420-0_38.
10. McDonnell, M.; UniSA, S. *Low-Complexity Acoustic Scene Classification Using One-Bit-per-Weight Deep Convolutional Neural Networks*; Technical Report, Detection and Classification of Acoustic Scenes and Events 2020 Challenge; 1 March–1 July 2020.
11. Jiang, S.; Shi, C.; Li, H. Acoustic Scene Classification Technique for Active Noise Control. In Proceedings of the 2019 International Conference on Control, Automation and Information Sciences (ICCAIS), Chengdu, China, 23–26 October 2019; pp. 1–5.
12. Ma, X.; Shao, Y.; Ma, Y.; Zhang, W.Q. Deep Semantic Encoder-Decoder Network for Acoustic Scene Classification with Multiple Devices. In Proceedings of the 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, 7–10 December 2020; pp. 365–370.
13. Zhang, T.; Liang, J.; Ding, B. Acoustic scene classification using deep CNN with fine-resolution feature. *Expert Syst. Appl.* **2020**, *143*, 113067.
14. Yang, L.; Tao, L.; Chen, X.; Gu, X. Multi-scale semantic feature fusion and data augmentation for acoustic scene classification. *Appl. Acoust.* **2020**, *163*, 107238. doi:10.1016/j.apacoust.2020.107238.
15. He, N.; Zhu, J. A Weighted Partial Domain Adaptation for Acoustic Scene Classification and Its Application in Fiber Optic Security System. *IEEE Access* **2021**, *9*, 2244–2250. doi:10.1109/access.2020.3044153.
16. Nguyen, T.; Pernkopf, F.; Kosmider, M. Acoustic Scene Classification for Mismatched Recording Devices Using Heated-Up Softmax and Spectrum Correction. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; doi:10.1109/icassp40776.2020.9053582.
17. Zhang, L.; Han, J.; Shi, Z. Learning Temporal Relations from Semantic Neighbors for Acoustic Scene Classification. *IEEE Signal Process. Lett.* **2020**, *27*, 950–954. doi:10.1109/lsp.2020.2996085.
18. Mezza, A.I.; Habets, E.A.; Müller, M.; Sarti, A. Feature Projection-Based Unsupervised Domain Adaptation for Acoustic Scene Classification. In Proceedings of the 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP), Espoo, Finland, 21–24 September 2020; pp. 1–6.
19. Mezza, A.I.; Habets, E.A.P.; Muller, M.; Sarti, A. Unsupervised Domain Adaptation for Acoustic Scene Classification Using Band-Wise Statistics Matching. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; doi:10.23919/eusipco47968.2020.9287533.

20. Takeyama, S.; Komatsu, T.; Miyazaki, K.; Togami, M.; Ono, S. Robust Acoustic Scene Classification to Multiple Devices Using Maximum Classifier Discrepancy and Knowledge Distillation. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; doi:10.23919/eusipco47968.2020.9287734.
21. Ooi, K.; Peksi, S.; Gan, W.S. *Ensemble of Pruned Low-Complexity Models for Acoustic Scene Classification*; Technical Report, In Proceedings of 5th the Workshop on Detection and Classification of Acoustic Scenes and Events 2020 (DCASE 2020), Tokyo, Japan, 2–4 November 2020.
22. Kwiatkowska, Z.; Kalinowski, B.; Kośmider, M.; Rykaczewski, K. Deep Learning Based Open Set Acoustic Scene Classification. In *Interspeech 2020*; ISCA: Singapore 2020; pp. 1216–1220. doi:10.21437/interspeech.2020-3092.
23. Alamir, M.A. A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers. *Appl. Acoust.* **2021**, *175*, 107829. doi:10.1016/j.apacoust.2020.107829.
24. Abrol, V.; Sharma, P. Learning Hierarchy Aware Embedding from Raw Audio for Acoustic Scene Classification. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2020**, *28*, 1964–1973. doi:10.1109/taslp.2020.3001969.
25. Wu, Y.; Lee, T. Time-Frequency Feature Decomposition Based on Sound Duration for Acoustic Scene Classification. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 716–720. doi:10.1109/icassp40776.2020.9053194.
26. Leng, Y.; Zhao, W.; Lin, C.; Sun, C.; Wang, R.; Yuan, Q.; Li, D. LDA-based data augmentation algorithm for acoustic scene classification. *Knowl.-Based Syst.* **2020**, *195*, 105600. doi:10.1016/j.knosys.2020.105600.
27. Pham, L.; Phan, H.; Nguyen, T.; Palaniappan, R.; Mertins, A.; McLoughlin, I. Robust acoustic scene classification using a multi-spectrogram encoder-decoder framework. *Digit. Signal Process.* **2021**, *110*, 102943. doi:10.1016/j.dsp.2020.102943.
28. Nguyen, T.; Ngo, D.; Pham, L.; Tran, L.; Hoang, T. A Re-trained Model Based On Multi-kernel Convolutional Neural Network for Acoustic Scene Classification. In Proceedings of the 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh City, Vietnam, 14–15 October 2020; doi:10.1109/rivf48685.2020.9140774.
29. Gao, W.; McDonnell, M.; UniSA, S. *Acoustic Scene Classification Using Deep Residual Networks with Focal Loss and Mild Domain Adaptation*; Technical Report, Detection and Classification of Acoustic Scenes and Events 2020 Challenge; 1 March–1 July 2020.
30. Lee, Y.; Lim, S.; Kwak, I.Y. CNN-Based Acoustic Scene Classification System. *Electronics* **2021**, *10*, 371. doi:10.3390/electronics10040371.
31. Seo, S.; Kim, C.; Kim, J.H. *Multi-Channel Feature Using Inter-Class and Inter-Device Standard Deviations for Acoustic Scene Classification*; Technical Report, Detection and Classification of Acoustic Scenes and Events 2020 Challenge; 1 March–1 July 2020.
32. Hu, H.; Yang, C.H.H.; Xia, X.; Bai, X.; Tang, X.; Wang, Y.; Niu, S.; Chai, L.; Li, J.; Zhu, H.; et al. A Two-Stage Approach to Device-Robust Acoustic Scene Classification. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2020; pp. 1–5.
33. McDonnell, M.D.; Gao, W. Acoustic Scene Classification Using Deep Residual Networks with Late Fusion of Separated High and Low Frequency Paths. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 141–145. doi:10.1109/icassp40776.2020.9053274.
34. Hu, H.; Yang, C.H.H.; Xia, X.; Bai, X.; Tang, X.; Wang, Y.; Niu, S.; Chai, L.; Li, J.; Zhu, H.; et al. Device-robust acoustic scene classification based on two-stage categorization and data augmentation. *arXiv* **2020**, arXiv:2007.08389.
35. Bai, X.; Du, J.; Pan, J.; shun Zhou, H.; Tu, Y.H.; Lee, C.H. High-Resolution Attention Network with Acoustic Segment Model for Acoustic Scene Classification. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 656–660. doi:10.1109/icassp40776.2020.9053519.
36. Singh, A.; Rajan, P.; Bhavsar, A. SVD-based redundancy removal in 1D CNNs for acoustic scene classification. *Pattern Recognit. Lett.* **2020**, *131*, 383–389. doi:10.1016/j.patrec.2020.02.004.
37. Paseddula, C.; Gangashetty, S.V. Acoustic Scene Classification using Single Frequency Filtering Cepstral Coefficients and DNN. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–6. doi:10.1109/ijcnn48605.2020.9207257.
38. Lostanlen, V.; Andén, J. Binaural scene classification with wavelet scattering. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), Tampere University of Technology, Tampere, Finland, September 2016.
39. Shim, H.J.; Jung, J.W.; Kim, J.H.; Yu, H.J. Capturing scattered discriminative information using a deep architecture in acoustic scene classification. *arXiv* **2020**, arXiv:2007.04631.
40. Jung, J.W.; Heo, H.S.; Shim, H.J.; Yu, H.J. Knowledge Distillation in Acoustic Scene Classification. *IEEE Access* **2020**, *8*, 166870–166879. doi:10.1109/access.2020.3021711.
41. Nguyen, T.; Pernkopf, F. Acoustic Scene Classification Using a Convolutional Neural Network Ensemble and Nearest Neighbor Filters. In Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, Surrey, UK, 19–20 November 2018.
42. Jung, J.W.; Heo, H.S.; Shim, H.J.; Yu, H. DNN based multi-level feature ensemble for acoustic scene classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), Surrey, UK, 19–20 November 2018; pp. 113–117.
43. Singh, A.; Thakur, A.; Rajan, P.; Bhavsar, A. A layer-wise score level ensemble framework for acoustic scene classification. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; pp. 837–841.

44. Sakashita, Y.; Aono, M. Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE 2018), Surrey, UK, 19–20 November 2018.
45. Mars, R.; Pratik, P.; Nagisetty, S.; Lim, C. Acoustic scene classification from binaural signals using convolutional neural networks. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), New York University, NY, USA, 25–26 October 2019.
46. Huang, J.; Lu, H.; Lopez Meyer, P.; Cordourier, H.; Del Hoyo Ontiveros, J. Acoustic scene classification using deep learning-based ensemble averaging. Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), New York University, NY, USA, 25–26 October 2019.
47. Wang, W.; Liu, M.; Li, Y. The SEIE-SCUT systems for acoustic scene classification using CNN ensemble. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), New York University, NY, USA, 25–26 October 2019.
48. Ding, B.; Liu, G.; Liang, J. Acoustic scene classification based on ensemble system. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), New York University, NY, USA, 25–26 October 2019.
49. Xu, K.; Zhu, B.; Kong, Q.; Mi, H.; Ding, B.; Wang, D.; Wang, H. General audio tagging with ensembling convolutional neural networks and statistical features. *J. Acoust. Soc. Am.* **2019**, *145*, EL521–EL527.
50. Gao, L.; Xu, K.; Wang, H.; Peng, Y. Multi-representation knowledge distillation for audio classification. *Multimed. Tools Appl.* **2022**, 1–24. doi: 10.1007/s11042-021-11610-8.
51. Wang, M.; Wang, R.; Zhang, X.L.; Rahardja, S. Hybrid constant-Q transform based CNN ensemble for acoustic scene classification. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 1511–1516.
52. Lopez-Meyer, P.; Ontiveros, J.d.H.; Stemmer, G.; Nachman, L.; Huang, J. Ensemble of convolutional neural networks for the DCASE 2020 acoustic scene classification challenge. In Proceedings of 5th the Workshop on Detection and Classification of Acoustic Scenes and Events 2020 (DCASE 2020), Tokyo, Japan, 2–4 November 2020.
53. Chin, C.S.; Kek, X.Y.; Chan, T.K. Scattering Transform of Averaged Data Augmentation for Ensemble Random Subspace Discriminant Classifiers in Audio Recognition. In Proceedings of the 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 19–20 March 2021; Volume 1, pp. 454–458.
54. Wang, Q.; Zheng, S.; Li, Y.; Wang, Y.; Wu, Y.; Hu, H.; Yang, C.H.H.; Siniscalchi, S.M.; Wang, Y.; Du, J.; et al. A Model Ensemble Approach for Audio-Visual Scene Classification. In Proceedings of the 6th Workshop on Detection and Classification of Acoustic Scenes and Events 2021 (DCASE 2021), Online, 15–19 November 2021. doi: 10.5281/zenodo.5770113.
55. Sarman, S.; Sert, M. Audio based violent scene classification using ensemble learning. In Proceedings of the 2018 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, Turkey, 22–25 March 2018; pp. 1–5.
56. Paseddula, C.; Gangashetty, S.V. Late fusion framework for Acoustic Scene Classification using LPCC, SCMC, and log-Mel band energies with Deep Neural Networks. *Appl. Acoust.* **2021**, *172*, 107568. doi:10.1016/j.apacoust.2020.107568.
57. Mallat, S. Group Invariant Scattering. *Commun. Pure Appl. Math.* **2012**, *65*, 1331–1398. doi:10.1002/cpa.21413.
58. Anden, J.; Mallat, S. Deep Scattering Spectrum. *IEEE Trans. Signal Process.* **2014**, *62*, 4114–4128. doi:10.1109/tsp.2014.2326991.
59. Zhu, H.; Wong, T.; Lin, N.; Lung, H.; Li, Z.; Theodoridis, S. A New Target Classification Method for Synthetic Aperture Radar Images based on Wavelet Scattering Transform. In Proceedings of the 2020 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Macau, China, 21–24 August 2020; pp. 1–6. doi:10.1109/icspcc50002.2020.9259471.
60. Ghezaiel, W.; Brun, L.; Lezoray, O. Wavelet Scattering Transform and CNN for Closed Set Speaker Identification. In Proceedings of the 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 21–24 September 2020; pp. 1–6. doi:10.1109/mmmsp48831.2020.9287061.
61. Adiga, A.; Magimai, M.; Seelamantula, C.S. Gammatone wavelet Cepstral Coefficients for robust speech recognition. In Proceedings of the 2013 IEEE International Conference of IEEE Region 10 (TENCON 2013), Xi'an, China, 22–25 October 2013; pp. 1–4. doi:10.1109/tencon.2013.6718948.
62. Anden, J.; Lostanlen, V.; Mallat, S. Joint time-frequency scattering for audio classification. In Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, 17–20 September 2015; pp. 1–6. doi:10.1109/mlsp.2015.7324385.
63. Kreyszig, E. *Advanced Engineering Mathematics*, 10th ed.; Publisher John Wiley & Sons: Hoboken, NJ, USA, 2009.
64. Chaparro, L.; Akan, A. *Signals and Systems Using MATLAB*; Academic Press: Cambridge, MA, USA, 2018.
65. Slaney, M. *An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank*; Apple Computer Technical Report #35; Perception Group, Advanced Technology Group, Apple Computer Inc: Cupertino, Santa Clara County, CA, USA, 1993, Volume 35, pp. 57–64.
66. Ho, T.K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832–844. doi:10.1109/34.709601.
67. Mesaros, A.; Heittola, T.; Diment, A.; Elizalde, B.; Shah, A.; Vincent, E.; Raj, B.; Virtanen, T. DCASE 2017 challenge setup: Tasks, datasets and baseline system. Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2017, Munich, Germany, 16–17 November 2017.

68. Zhao, S.; Nguyen, T.N.T.; Gan, W.S.; Jones, D.L. ADSC submission for DCASE 2017: Acoustic scene classification using deep residual convolutional neural networks. Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2017, Munich, Germany, 16–17 November 2017.
69. Jung, J.W.; Heo, H.S.; Yang, I.; Yoon, S.H.; Shim, H.J.; Yu, H.J. DNN-based audio scene classification for DCASE 2017: Dual input features, balancing cost, and stochastic data duplication. In Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2017, Munich, Germany, 16–17 November 2017.
70. Piczak, K.J. The details that matter: Frequency resolution of spectrograms in acoustic scene classification. In Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2017, Munich, Germany, 16–17 November 2017.
71. Kukanov, I.; Hautamäki, V.; Lee, K.A. Recurrent neural network and maximal figure of merit for acoustic event detection. In Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2017, Munich, Germany, 16–17 November 2017.
72. Park, S.; Mun, S.; Lee, Y.; Ko, H. Acoustic scene classification based on convolutional neural network using double image features. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE 2017), Munich, Germany, 16–17 November 2017; pp. 98–102.
73. Lehner, B.; Eghbal-Zadeh, H.; Dorfer, M.; Korzeniowski, F.; Koutini, K.; Widmer, G. Classifying short acoustic scenes with I-vectors and CNNs: Challenges and optimisations for the 2017 DCASE ASC task. In Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2017, Munich, Germany, 16–17 November 2017.
74. Hyder, R.; Ghaffarzagdegan, S.; Feng, Z.; Hasan, T. Buet Bosch consortium (B2C) acoustic scene classification systems for DCASE 2017 challenge. In Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2017, Munich, Germany, 16–17 November 2017.
75. Zheng, W.; Jiantao, Y.; Xing, X.; Liu, X.; Peng, S. Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion. In Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2017, Munich, Germany, 16–17 November 2017.
76. Han, Y.; Park, J.; Lee, K. Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification. In proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2017, Munich, Germany, 16–17 November 2017.
77. Mun, S.; Park, S.; Han, D.K.; Ko, H. Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane. In proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2017, Munich, Germany, 16–17 November 2017.
78. Ren, Z.; Qian, K.; Zhang, Z.; Pandit, V.; Baird, A.; Schuller, B. Deep Scalogram Representations for Acoustic Scene Classification. *IEEE/CAA J. Autom. Sin.* **2018**, *5*, 662–669. doi:10.1109/jas.2018.7511066.
79. Waldekar, S.; Saha, G. Wavelet Transform Based Mel-scaled Features for Acoustic Scene Classification. In *Interspeech 2018*. ; ISCA: Singapore, 2018; pp. 3323–3327. doi:10.21437/interspeech.2018-2083.
80. Yang, Y.; Zhang, H.; Tu, W.; Ai, H.; Cai, L.; Hu, R.; Xiang, F. Kullback–Leibler Divergence Frequency Warping Scale for Acoustic Scene Classification Using Convolutional Neural Network. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 840–844. doi:10.1109/icassp.2019.8683000.
81. Wu, Y.; Lee, T. Enhancing Sound Texture in CNN-based Acoustic Scene Classification. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 815–819. doi:10.1109/icassp.2019.8683490.
82. Chen, H.; Zhang, P.; Yan, Y. An Audio Scene Classification Framework with Embedded Filters and a DCT-based Temporal Module. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 835–839. doi:10.1109/icassp.2019.8683636.
83. Mesaros, A.; Heittola, T.; Benetos, E.; Foster, P.; Lagrange, M.; Virtanen, T.; Plumbley, M.D. Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *26*, 379–393.