Date of publication JUN-30, 2022, date of current version APR-08, 2022.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.21.2.2595

Sound Context Classification based on Joint Learning Model and Multi-Spectrogram Features

DAT NGO[1], LAM PHAM[2], ANH NGUYEN[3], TIEN LY[4], KHOA PHAM[5], THANH NGO[6]

[1]School of Computer Science and Electronic Engineering, University of Essex, UK, (e-mail: dn22678@essex.ac.uk)
[2]Center for Digital Safety & Security, Austrian Institute of Technology, Austria, (e-mail: lam.pham@ait.ac.at)
[3]Electrical and Electronics Department, Ho Chi Minh City University of Technology, Vietnam, (e-mail: anh.nguyenk2017@hcmut.edu.vn)
[4]Engineering Science, University of Oxford, UK, (e-mail: ktien@oxfordrobotics.institute)
[5]Department of Computer and Communication Engineering, University of Technology and Education HCMC, Vietnam, (e-mail: khoapv@hcmute.edu.vn)
[6]Department of Electrical Engineering, Da Nang University of Science and Technology, Vietnam, (e-mail: ndthanh@dut.udn.vn)

Corresponding author: Dat Ngo (e-mail: dn22678@essex.ac.uk).

ABSTRACT This article presents a deep learning framework applied for Acoustic Scene Classification (ASC), the task of classifying different environments from the sounds they produce. To successfully develop the framework, we firstly carry out a comprehensive analysis of spectrogram representation extracted from sound scene input, then propose the best multi-spectrogram combination for front-end feature extraction. In terms of back-end classification, we propose a novel joint learning model using a parallel architecture of Convolutional Neural Network (CNN) and Convolutional Recurrent Neural Network (C-RNN), which is able to learn efficiently both spatial features and temporal sequences of a spectrogram input. The experimental results have proved our proposed framework general and robust for ASC tasks by three main contributions. Firstly, the most effective spectrogram combination is indicated for specific datasets that none of publication previously analyzed. Secondly, our joint learning architecture of CNN and C-RNN achieves better performance compared with the CNN only which is proposed for the baseline in this paper. Finally, our framework achieves competitive performance compared with the state-of-the-art systems on various benchmark datasets of IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 Task 1, 2017 Task 1, 2018 Task 1A & 1B, and LITIS Rouen.

KEYWORDS Acoustic scene classification; Spectrogram; Convolutional neural network; Recurrent neural network; Joint learning architecture; Feature extraction.

I. INTRODUCTION

ACOUSTIC Scene Classification (ASC), which aims to identify a sound scene context, vitally contributes to a variety of real-life applications ranging from security [1], surveillance [2] and context-aware consumer services [3–5]. Although ASC research is very close to Automatic Speech Recognition (ASR) and Speaker Recognition System (SRS) due to exploring audio signals, ASC currently presents various and different challenges. Firstly, there is a wide range of acoustic events in real-world environments, and these occur in different ways. Some sound events constitute natural auditory scenes that presents an acoustic mixture signal. For instance, *bird* sounds and the sounds of *leaves*, *grass*, or *trees blowing* in the wind clearly indicate certain context like *in a park* or *on a field*. However, it is more difficult to handle some sound events that are not context specific such as *engine*, or *talking*. This kind of context causes a confusion for even human (only listening) to recognize exactly *in a street, on a transportation* such as *car, bus, tube* or *in a station*. Indeed, experimental results in [6] indicate that a proposed ASC system configured by Mel Frequency Cepstral Coefficients (MFCCs) – Hidden Markov Model (HMM) significantly outperformed human ability for recognizing everyday acoustic. Secondly, if sound events are considered as signal that mixed in diverse scenes as noise, there are different levels of signal-to-noise ratio (SNR) due to environmental conditions, distance of recording devices and so on. Moreover, these sounds exist across a wide range of frequency bands. Some occupy narrow frequency bands, while some spread over wide bands, and many sounds have frequency bands that overlap each other. Finally, natural sounds in ASC research do not follow any structure, unlike a speech signal.

To deal with these challenges, the state-of-the-art systems tend to make use of multi-input features. In particular, systems approaching frame-based features make effort to combine

frequency and temporal features to maximize the chance of correct feature representation. For instance, MFCCs [7], one of most used frequency features, is combined with a wide range of temporal features such as loudness, average short time energy, sub-band energy, zero-crossing rate, spectral flux, or spectral centroid in [8–10]. Similarly, an effective combination of MFCCs with a variety of features such as perceptual linear prediction (PLP) coefficients, power normalized cepstral coefficients (PNCC), robust compressive gamma-chirp filter bank cepstral coefficients (RCGCC) or subspace projection cepstral coefficients (SPPCC) was proposed in [11] that helps to achieve the top-three system in DCASE 2016 challenge. As usual, ASC systems approaching frame-based feature representation use traditional machine learning models for back-end classification, such as Hidden Markov Model (HMM) [6], Support Vector Machine (SVM) [12], [13], and Gaussian Mixture Model (GMM) [14].

However, frame-based representation shows its limitation to fully capture information of sound signals compared to spectrogram representation [15], [16], [17]. Therefore, two-dimensional spectrograms appear as a more effective way for low-level feature representation, and have been exploited by the state-of-the-art ASC systems. In particular, spectrograms such as short-term Fourier transform (STFT) [7], log-Mel [18], [19], MFCC [20], constant-Q transform (CQT) [21], and Gammatone spectrograms (GAM) [15], [22] are the most frequent low-level features used. To further investigate on this advantage, multi-spectrogram combinations are widely proposed. For instances, log-Mel is combined with a different type of spectrograms such as Mel-based nearest neighbor filter (NNF) spectrogram [23], [24], CQT [25], or two spectrograms such as MFCC and GAM in [15], [26]. However, no one analyzes and concludes which combination of spectrogram general and robust achieves high performance across a variety of benchmark datasets.

To explore two-dimensional spectrogram representation, ASC systems usually deploy complicated classification models, mainly coming from deep learning techniques. For examples, Yang *et al.* [27] proposed a complicated CNN-based architecture called the *Xception* network. This is inspired by the fact that a deep learning network trained by a wide range of feature scales and over separated channels can result in a very powerful model. Besides, Truc *et al.* [23] applied a parallel CNNs to learn from two types of spectrogram (log-Mel and NNF). Next, the two outputs of the CNNs are concatenated to generate high-performed features that were thus explored by a DNN and achieved the highest accuracy rate in DCASE 2018 Task 1B challenge. Approach Recurrent Neural Network (RNN) based architecture, Zang *et al*. [28–30] provided a deep analysis of the application of Long Short-Term Memory (LSTM) for ASC. Other examples prove effectiveness in exploiting RNN-based networks for ASC were published by Huy *et al.* [22], [31], [32]. However, rather than focus only on learning spatial feature by CNN-based networks or temporal information with sequence models as state-of-the-art methods performed in [23], [31], and [32], it is necessary to propose a novel joint learning model that can make the most of its advantages in learning effectively features in both spatial and temporal domains.

With the analysis of the-state-of-the-art as well as some limitation ASC systems needed to be solved, we propose a robust framework that uses spectrogram representation for low-level feature input and explore a joint learning model architecture for classification. In particular, we mainly contribute:

- Although spectrogram-based ASC systems explore multi-spectrogram input features to deal with ASC challenges, none of research has analyzed and indicated the most effective combination of spectrograms. In this paper, we, therefore, provide a comprehensive analysis on spectrograms by conducting experiments on five common types of spectrograms, comprising of Short-time Fourier Transform (STFT), log-Mel, Mel Frequency Cepstral Coefficient (MFCC), Constant Q Transform (CQT), and Gammatone filter (GAM). To this end, we firstly introduce a baseline C-DNN deep learning model, likely VGG-9 [33]. Consequently, we evaluate individual spectrograms on the baseline C-DNN network proposed, thus indicate the most effective combination of spectrograms via the late fusion of individual spectrogram.
- Next, we improve the baseline C-DNN model by adding a parallel C-RNN architecture to efficiently learn the structure of temporal sequences of spectrograms. By using a parallel C-DNN and C-RNN network, we create a joint learning architecture that is very useful to deploy the two-dimensional spectrogram input.
- To evaluate ASC systems, researchers normally did experiments on one dataset [9], [11]. Some proposed to evaluate on two datasets [18], [34]. This may not conclude ASC systems proposed general or powerful. We, therefore, conduct extensive experiments, evaluating our proposed systems over five ASC datasets of DCASE 2016 Task 1, DCASE 2017 Task 1, DCASE 2018 Task 1A & 1B, and LITIS Rouen published recently. Competitive results obtained on various datasets showing different category number, recording time, and wide range of real-life environments strongly prove our proposed system general and robust.

II. DATASET AND SETTING

Our experiments are conducted over a variety of published ASC datasets, comprising of LITIS Rouen [35] and IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 Task 1 [36], 2017 Task 1 [20], 2018 Task 1A & 1B [37].

LITIS Rouen dataset was recorded at a sample rate of 22050 Hz with 3026 segments, each presents 30-s duration. This dataset contains totally 25.51 recording hours for 19 urban scene categories, showing unbalanced data. Following the mandated settings, the dataset is separated and organized for 20-fold cross validation, reporting the final classification accuracy by averaging over the 20 testing folds.

DCASE 2016 Task 1 and DCASE 2017 Task 1 similarly present 15 categories and were recorded at 44100 Hz. While each segment in DCASE 2016 is 30 s, 10-s duration is presented in DCASE 2017. Noticeably, DCASE 2017 reuses all DCASE 2016 and adds new data recorded. Due to the recommended setting, we train our proposed system on development set (Dev.) and evaluate on the evaluation set (Eva.). As regards DCASE 2018 Task 1A, it was recorded at 44800 kHz, spanning 10 categories and using one recording device namely A. DCASE 2018 Task 1B reuses all data from DCASE 2018 Task 1A, and adds more data recorded by two different devices, namely B and C. Noticeably, the total recording time spent on device B and C is much less than

device A (denoted as DCASE 2018 Task 1A), reporting totally 4 hours on B&C compared to 24 hours in device A. As a result, DCASE 2018 Task 1B dataset involves issues of mismatched recording devices and unbalanced data in terms of recording devices. Therefore, DCASE 2018 Task 1B challenge only compares systems' results on device B&C with less recording time. As DCASE 2018 Task 1A and 1B have not released labels of evaluation set, we separate development set into two sub-sets, namely Training and Test sets for training and testing processes respectively. While DCASE 2016 Task 1 and DCASE 2017 Task 1 are balanced, little unbalanced data is shown in DCASE 2018 Task 1A and 1B.

Because this paper evaluates ASC datasets of LITIS Rouen and DCASE in years of 2016, 2017, and 2018, the evaluation metric used in this paper follows these challenges. In particular, if C is considered as the number of audio segments which are correctly predicted, and the total number of audio segments is T, the classification accuracy (Acc.%) mentioned in these challenges is:

$$Acc.(\%) = 100\frac{C}{T}$$

III. HIGH-LEVEL ARCHITECTURE AND BASELINE C-DNN NETWORK PROPOSED

Our proposed deep learning framework applied for ASC, in general, is described in Fig. 1. As Fig. 1 shows, the framework is separated into low-level feature extraction (the upper part) and back-end classification (the lower part).

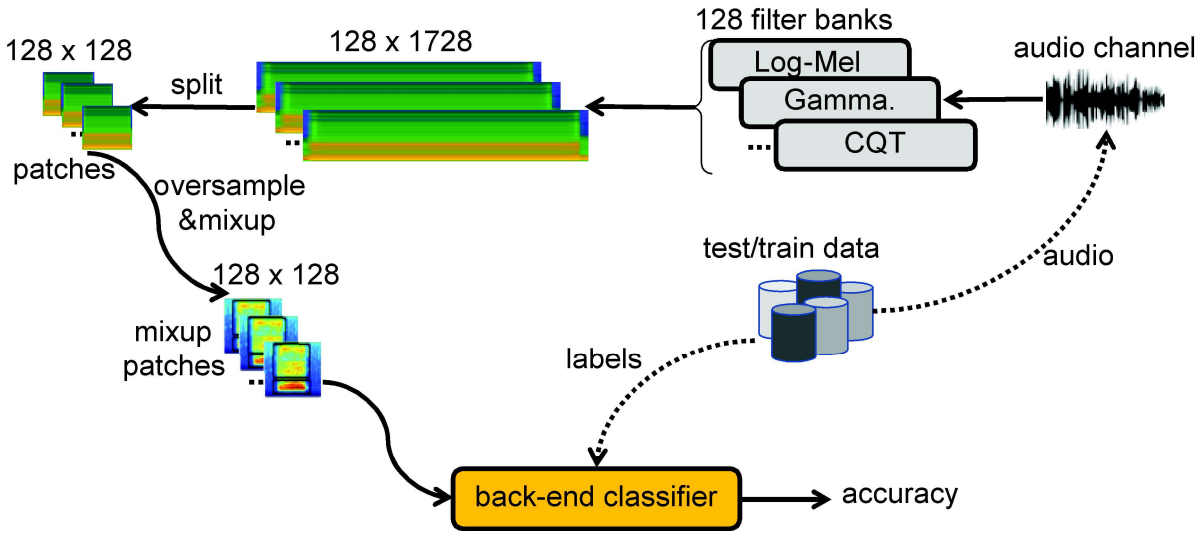


Figure 1. High-level architecture of our ASC system.

In particular, the draw audio from Channel 1 is firstly transformed into spectrogram representation, using 128 filter banks. The entire spectrogram is thus split into non-overlapped image patches of 128x128. To deal with unbalanced data issue, we apply two data augmentation techniques on the image patches. Firstly, we randomly oversample image patches, which belong to categories with less audio segments. Next, the mix-up data augmentation [38] is applied to generate new image patches. Let us consider two original image patches as $\mathbf{X}_1$, $\mathbf{X}_2$ and expected labels as $\mathbf{y}_1$, $\mathbf{y}_2$, new image patches are generated as below equations:

$$\mathbf{X}_{mp1} = \mathbf{X}_1\gamma + \mathbf{X}_2(1-\gamma) \quad (1)$$

$$\mathbf{X}_{mp2} = \mathbf{X}_1(1-\gamma) + \mathbf{X}_2\gamma \quad (2)$$

$$\mathbf{y}_{mp1} = \mathbf{y}_1\gamma + \mathbf{y}_2(1-\gamma) \quad (3)$$

$$\mathbf{y}_{mp2} = \mathbf{y}_1(1-\gamma) + \mathbf{y}_2\gamma\,, \quad (4)$$

where γ is random coefficient from both unit and beta distribution, $\mathbf{X}_{mp1}$, $\mathbf{X}_{mp2}$ and $\mathbf{y}_{mp1}$, $\mathbf{y}_{mp2}$ are new image patches and new labels generated, respectively. Eventually, the mix-up patches are fed into a back-end classifier for classification.

By using the high-level architecture mentioned above, we evaluate five individual spectrograms (STFT, log-Mel, MFCC, GAM, and CQT), thus indicating which kind of spectrograms and their combinations is the most influencing on our system's performance. Besides, to inspire that each spectrogram contains discriminative and complementary features, we fuse the individual systems' accuracy results, thus indicating which combination of spectrograms is effective to improve the performance.

Note that we use the same setting with window size = 1290, hop size = 256, frequency minimum f_{min} = 10 Hz, and filter bank number = 128 to generate same-size spectrograms.

In order to evaluate individual and multiple spectrograms, we proposed a C-DNN network architecture, which is considered as the baseline back-end classification. As Fig. 2 and Table 1 show, the baseline C-DNN architecture comprises of CNN and DNN parts in order. CNN part is described by six blocks, namely Vg-Cv, performed by Batch Normalization (BN), Convolution (Cv[kernel size]), Rectified Linear Unit (ReLU), Dropout (Dr(Percentage dropped)), Average Pooling (AP [kernel size]), Global Average Pooling (GAP) layers as showed in the top of Table 1. Meanwhile, DNN part in Fig. 2 is configured by three blocks, namely Vg-Fl with Fully-connect (Fl), ReLU, Dropout (Dr(Percentage dropped)), and Softmax layers, as described in the bottom of Table 1.

Table 1. Network layers used in C-DNN architecture

	Blocks	Layers	Output Shape
CNN Architecture	Input		128x128x1
	Vg-Cv Block 01	BN – Cv [9 x 9] – ReLU – BN – AP [2 x 2] – Dr (10%)	64x64x32
	Vg-Cv Block 02	Cv [7 x 7] – ReLU – BN – AP [2 x 2] – Dr (15%)	32x32x64
	Vg-Cv Block 03	Cv [5 x 5] – ReLU – BN – Dr (20%)	32x32x128
	Vg-Cv Block 04	Cv [5 x 5] – ReLU – BN – AP [2 x 2] – Dr (20%)	16x16x128
	Vg-Cv Block 05	Cv [3 x 3] – ReLU – BN – Dr (25%)	16x16x256
	Vg-Cv Block 06	Cv [3 x 3] – ReLU – BN – GAP – Dr (25%)	256
DNN Architecture	Vg-Fl Block 01	Fl – ReLU – Dr (30%)	512
	Vg-Fl Block 02	Fl – ReLU – Dr (30%)	1024
	Vg-Fl Block 03	Fl – Softmax	10

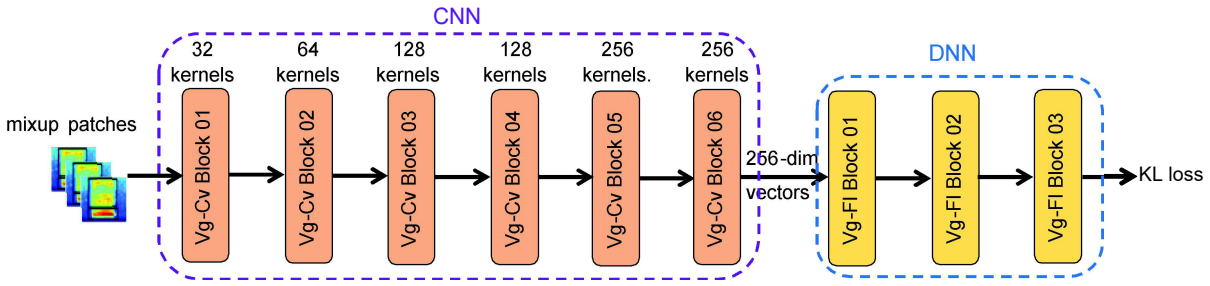


Figure 2. Block-level architecture of the baseline C-DNN network.

It can be seen that CNN part helps to map input image patches to condensed and discriminative vectors, referred to as high-level features. Each high-level feature vector presents 256 dimensions due to the number of kernels used in the final convolutional layer of Vg-Cv block 06. Next, DNN part explores the high-level features, thus classifies into 10 categories (the category number in DCASE 2018 Task 1B dataset evaluated) and reports the classification accuracy.

Eventually, we evaluate individual spectrogram and combinations of spectrograms by using the framework with C-DNN baseline architecture over DCASE 2018 Task 1B. Next, we not only compare these performances of baseline architecture to this DCASE baseline but also indicate which kind of spectrograms and their combinations is the most influencing on our system's performance as details in the next section 6.1.

IV. AN ANALYSIS OF SPECTROGRAMS

A. SPECTROGRAM REPRESENTATION AND THEIR COMBINATIONS PROPOSED

To evaluate individual spectrograms and their combinations, formulas of spectrograms are firstly presented below:

Short-Time Fourier Transform (STFT): The first STFT spectrogram evaluated applies Fourier Transform to extract Frequency content of local section of input signal over short time duration. Let us consider $\mathbf{s}[n]$ as digital audio signal with length of N, a pixel value at central frequency f and time frame t of STFT spectrogram **STFT**[F, T] is computed as:

$$\mathbf{STFT}[f,t]=\sum_{n=0}^{N-1}\mathbf{s}[n]\mathbf{w}[t]e^{-j2\pi fn}, \qquad (5)$$

where $\mathbf{w}[t]$ is a window function, typically Hamming, while time resolution (T) of STFT spectrogram is set by window size and hop size, the frequency resolution (F) equals to the number of central frequencies set to 2048. The frequency resolution, eventually, re-scales into 128 that is the same as other spectrograms.

log-Mel: To generate log-Mel spectrogram, draw audio signal is firstly transformed into STFT spectrogram recently mentioned. Next, a Mel filter bank, which simulates the overall frequency selectivity of the human auditory system using the frequency warping $F_{mel} = 2595log(1+F/700)$ [7], is applied to generate a Mel spectrogram $\mathbf{MEL}[F_{mel}, T]$ (noting that frequency resolution (F_{mel}) is the Mel filter number set to 128). Eventually, logarithmic scaling is applied to obtain the log-Mel spectrogram. Let us consider $\mathbf{COE}[F_{mel},F]$ as matrix storing coefficients of Mel filters, log-Mel spectrogram likely a matrix is computed by:

$$\mathbf{log-Mel}[F_{mel},T]=$$

$$log(\mathbf{COE}[F_{mel},F]\times\mathbf{STFT}[F,T]) \qquad (6)$$

Mel Frequency Cepstral Coefficient (MFCC): From log-Mel spectrogram, Discrete Cosine Transform (DCT) is used to extract a sequence of uncorrelated coefficients crossing frequency dimension, reducing log-Mel frequency resolution into smaller space. A pixel value $\boldsymbol{DCT}[f_{dct,}\, t_{dct}]$ of DCT matrix $\boldsymbol{DCT}[F_{dct,}\, T_{dct}]$, where F_{dct} and T_{dct} are frequency and time resolutions, is computed by:

$$\mathbf{DCT}[f_{dct},t_{dct}]=\left(\frac{2}{F_{mel}}\right)^{\frac{1}{2}}\left(\frac{2}{T}\right)^{\frac{1}{2}}.\sum_{f_{mel}=0}^{F_{mel}-1}.\sum_{t=0}^{T-1}\Lambda(f_{mel}).cos\left[\frac{\pi f_{dct}}{F_{mel}}(2f_{mel}+1)\right].\Lambda(t)cos\left[\frac{\pi t_{dct}}{T}(2t+1)\right].\mathbf{log-Mel}[f_{mel},t], \quad (7)$$

where

$$\Lambda(x)=\begin{cases}\frac{1}{\sqrt{2}} & \text{if } x=0\\ 1 & \text{otherwise}\end{cases} \quad (8)$$

T and F_{mel} are time and frequency resolution of log-Mel spectrogram.

Next, delta coefficients per time frame showing difference of DCT coefficients over time are computed, shown in Equation (9).

$$\mathbf{DELTA}[F_{dct},t]=\frac{1}{2}(\mathbf{DCT}[F_{dct},t-1]-\mathbf{DCT}[F_{dct},t+1]) \quad (9)$$

Eventually, $\mathbf{DELTA}[F_{dct,}\,T_{dct}]$ is concatenated with DCT spectrogram $\mathbf{DCT}[F_{dct,}\,T_{dct}]$ across frequency dimension to generate MFCC spectrogram as expression $\mathbf{MFCC}[F_{mfcc,}\,T_{dct}]$ (note that MFCC frequency resolution (F_{mfcc}) doubles frequency resolution of DCT (F_{dct}) and equals to 128, and T_{dct} is set to equal to T resolution of log-Mel spectrogram).

Constant Q transform (CQT): This spectrogram applies a bank of filters corresponding to tonal spacing, where each filter is equivalent to a subdivision of an octave, with central frequencies given by:

$$f_k=(2^{\frac{1}{b}})^k f_{min} \quad for\ 1\le k\le K, \quad (10)$$

where f_k denotes the frequency of k^{th} spectral component, f_{min} is minimum frequency set to 10 Hz, b is the number of filters per octave as 24, and K is frequency resolution of CQT, which is 128. As the name suggest, the Q value is the ratio of central frequency to bandwidth, is constant computed as:

$$Q=\frac{f_k}{\Delta f_k}=\frac{f_k}{f_{k+1}-f_k}=\left(2^{\frac{1}{b}}-1\right)^{-1}. \quad (11)$$

Like STFT, CQT spectrogram is extracted using Fourier-based transformation, described as Equation (12):

$$\mathbf{CQT}[f_k,t]=\frac{1}{N(k)}\sum_{n=0}^{N(k)-1}\mathbf{s}[n].\mathbf{w}[k,n-t]e^{-i2\pi\frac{nQ}{N(k)}}, \quad (12)$$

where

$$N(k)=Q\frac{fs}{f_k} \quad (13)$$

$$\mathbf{w}[k,n]=\alpha+(1-\alpha)cos\frac{2\pi n}{N(k)-1}, \quad (14)$$

where, f_s is sample rate of digital input signal $\mathbf{s}[n]$, $\mathbf{w}[k,\,n]$ is window function with α set to 0.54. To generate STFT, log-Mel, MFCC, and CQT, we use a popular audio toolbox, namely Librosa [39].

Gammatone (GAM): Gammatone filters are designed to model the frequency-selective cochlea activation response of the human inner ear [40], in which filter output simulates the frequency response of the basilar membrane. The impulse response is given by:

$$g(t)=t^{P-1}e^{-2lt\pi}cos(2ft\pi+\theta), \quad (15)$$

where t is time, P is the filter order, θ is the phase of the carrier, l is filter bandwidth, and f is central frequency. The filter bank was then formulated as ERB scale [41] as follows:

$$ERB=24.7(4.37.10^{-3}f+1) \quad (16)$$

To quickly generate Gamma spectrogram, we apply a toolbox developed by Ellis *et al.* [42], namely Gammatone-like spectrogram. Firstly, audio signal is transformed into STFT spectra recently mentioned above. Next, gammatone weighting $\mathbf{COE}[F_{gam,}\,\mathrm{F}]$ is applied on STFT to obtain the Gamma spectrogram.

$$\mathbf{GAM}[F_{gam},T]=\mathbf{COE}[F_{gam},F]\times\mathbf{STFT}[F,T], \quad (17)$$

where F_{gam} resolution of GAM spectrogram is Gammatone filter number of 128.

As spectrogram formulas described, we construct a spectrogram tree as shown in Fig. 3. Firstly, although both of CQT and STFT spectrograms are built on Fourier Transform theory, they extract different central frequencies. From the root tree, we therefore separate into two main branches of CQT and STFT. From the branch of STFT spectrogram, we continuously divide into log-Mel and GAM spectrograms due to applying different Mel and Gammatone filters, respectively.

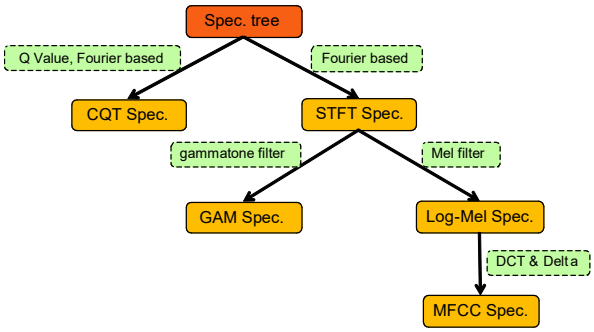


Figure 3. Constructed spectrogram tree based on difference of central frequencies and auditory models applied.

Eventually, MFCC is an extended branch from log-Mel due to extracting DCT and Delta from this spectrogram. It can be seen that five spectrograms proposed either extract different central frequencies or apply different auditory models. Therefore, each spectrogram may contain its own distinct and complimentary information. This inspires us to conduct experiments to indicate how individual spectrograms and their combinations affect an ASC system's performance.

Based on the tree shown in Fig. 3, we propose a variety of combinations as denoted in Table 2. In particular, two-spectrogram combinations are inspired from two main branches from the root tree, each extracts specific central frequencies. Thus, we create groups of CQT+STFT, CQT+GAM, CQT+log-Mel, and CQT+MFCC.

Table 2. Spectrogram combinations proposed.

Group of	Combinations of
Two spectrograms	CQT+STFT, CQT+GAM, CQT+log-Mel, CQT+MFCC,
Three spectrograms	CQT+GAM +log-Mel, CQT+GAM+MFCC,
Four spectrograms	CQT+GAM+STFT+MFCC, CQT+GAM+STFT+ log-Mel,
Five spectrograms	CQT +GAM+ STFT+MFCC+log-Mel

There are two third-spectrogram groups of CQT+GAM+log-Mel and CQT+GAM+MFCC evaluated that inspires exploring different central frequencies between CQT & STFT branches and different auditory models used among MFCC, log-Mel, and GAM. Given that that applying auditory models on STFT may destroy discriminative features on this spectrogram and two spectrograms of MFCC, log-Mel may contain very similar features due to coming from same Mel filter banks, we propose two four-spectrogram combinations, which are CQT+GAM+STFT+MFCC and CQT+GAM+STFT+log-Mel. Eventually, the combination of all five spectrograms is also evaluated.

B. LATE FUSION STRATEGY TO EVALUATE SPECTROGRAM COMBINATIONS

As the back-end classification works on smaller patches, the posterior probability of an entire spectrogram is computed by averaging of all patches' posterior probabilities. Let us consider $P^n = (P_1^n, P_2^n, \ldots, P_C^n)$ with C being the category number and the n^{th} out of N patches fed into learning model, as the probability of a test sound instance, then the mean classification probability is denoted as $\overline{p} = (\overline{p}_1, \overline{p}_2, \ldots, \overline{p}_c)$, where

$$\overline{p}_c = \frac{1}{N}\sum_{n=1}^{N} p_c^n \quad for\ 1 \le n \le N \qquad (18)$$

and the predicted label $\hat{y}$ for an individual spectrogram evaluated is determined using:

$$\hat{y} = argmax(\overline{p}_1, \overline{p}_2, ..., \overline{p}_C). \qquad (19)$$

To evaluate the combinations of spectrograms, we proposed a late fusion scheme, namely *Mean* fusion. In particular, we conduct experiments over individual spectrograms, thus obtaining posterior probability of each spectrogram as $\overline{p}_S = (\overline{p}_{S1}, \overline{p}_{S2}, \ldots, \overline{p}_{SC})$, where C is the category number and the s^{th} out of S spectrograms evaluated. Next, the posterior probability after late fusion $p_{f-mean} = (p_1, p_2, \ldots, p_c)$ is obtained from by:

$$p_c = \frac{1}{S}\sum_{s=1}^{S} \overline{p}_{sc} \quad for\ 1 \le s \le S. \qquad (20)$$

Eventually, the predicted label $\hat{y}$ is determined by:

$$\hat{y} = argmax(p_1, p_2, ..., p_C). \qquad (21)$$

C. HYPERPARAMETER SETTING AND DATASET USED TO EVALUATE SPECTROGRAMS

In this work, we adopt Tensorflow framework to build deep leaning models with learning rate of 0.0001, a batch size of 50, epoch number of 100, and Adam method [43] for learning rate optimization. As using mix-up data augmentation, the labels are not one-hot format. Therefore, we use Kullback-Leibler (KL) divergence loss [44] instead of the standard cross-entropy loss as shown in Equation below:

$$Loss_{KL}(\theta) = \sum_{n=1}^{N} \mathbf{y}_n \log(\frac{\mathbf{y}_n}{\hat{\mathbf{y}}_n}) + \frac{\lambda}{2} \| \theta \|_2^2, \qquad (22)$$

where $Loss_{KL}(\theta)$ is KL-loss function, θ describes the trainable parameters of the network trained, λ denotes the ℓ_2-norm regularization coefficient experimentally set to 0.0001, N is the batch size, $\mathbf{y}_n$ and $\hat{\mathbf{y}}_n$ are the ground-truth and the network recognized output, respectively. Note that we use only DCASE 2018 Task 1B dataset to analyze individual spectrograms and their combinations proposed.

V. FURTHER IMPROVED BACK-END CLASSIFICATION

A. JOINT LEARNING DEEP NEURAL NETWORK ARCHITECTURE PROPOSED

To further enhance ASC system's performance, we focus on exploring back-end classification in this section. Therefore, we

propose a joint learning model, as described in Fig. 4. As Fig. 4 shows, we reuse the CNN part with six Vg-Cv blocks from the baseline C-DNN architecture. These Convolutional blocks help to capture spatial features from spectrogram input, thus transform image patches into condensed and discriminative vectors with 256 dimension. Additionally, we add a parallel C-RNN architecture (the lower part of Fig. 4) that is used to capture structures of temporal sequences from spectrogram input. As Table 3 shows, the C-RNN proposed, input patches of 128x128 are fed into sub-blocks Cv, BN, ReLU, AP and Dr that are similar to those used in the CNN part. However, we adjust settings of these sub-blocks to allow the C-RNN network be able to learn time-sequential features. In particular, convolutional layers (Cv) with kernel size set to [4 x 1] are applied to learn the difference between frequency banks in each temporal frame. Next, average pooling layers (AP [16 x 1]) are used to scale the frequency dimension of the spectrogram but remain time resolution of 128. As a result, frequency dimension is scaled into 1, generating a sequence of 128-temporal frames after four Re-Cv blocks. Each temporal frame is represented by a 256-dimensional vector. Next the temporal sequence is fed into bi-GRU layer in Re-Bi-GRU Block, which learns the temporal sequence structure from two directions. The output of Re-Bi-GRU Block is a matrix of 128x256 with 128 temporal frames and 256 dimension each frame.

Table 3. C-RNN network architecture.

Blocks	Layers	Output Shape
	Input	128x128x1
Re-Cv Block 01	BN – Cv [4 x 1] – ReLU – BN – AP [2 x 1] – Dr (10%)	64x128x32
Re-Cv Block 02	Cv [4 x 1] – ReLU – BN – AP [2 x 1] – Dr (15%)	32x128x64
Re-Cv Block 03	Cv [4 x 1] – ReLU – BN – AP [2 x 1] – Dr (20%)	16x128x128
Re-Cv Block 04	Cv [4 x 1] – ReLU – BN – AP [16 x 1] – Dr (20%)	128x256
Re-Bi-GRU Block	Bi-GRU (64 hidden states, 30% dropout)	128x256
Re-GlAv Block	GAP	128

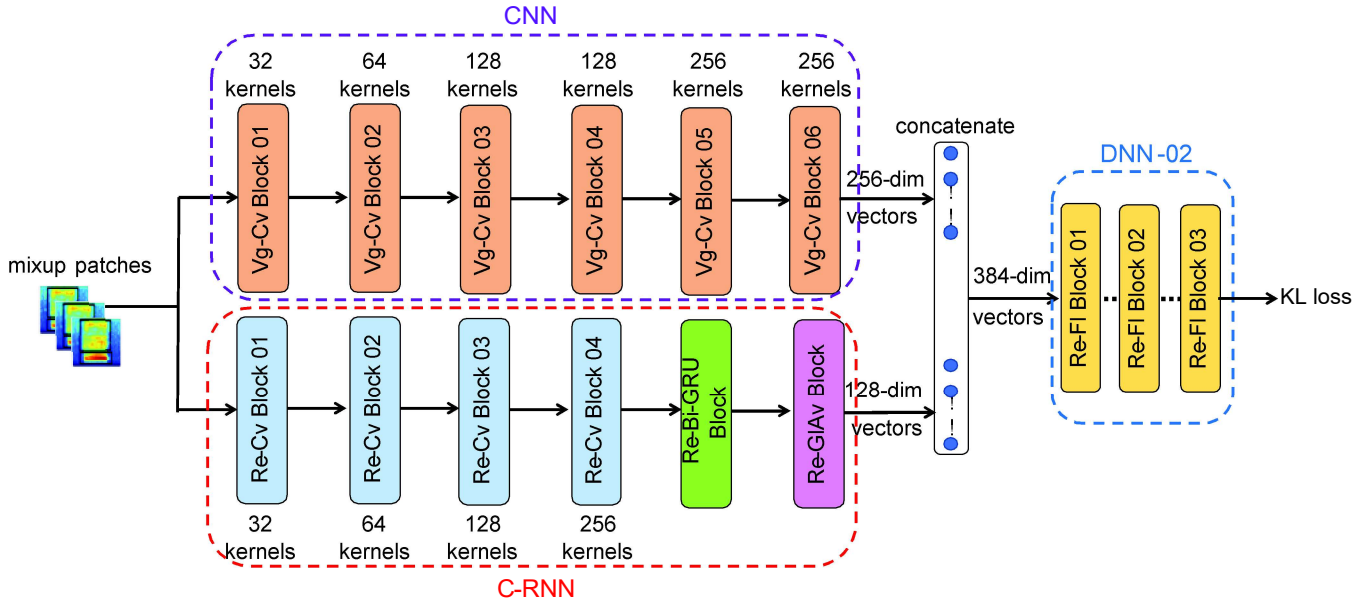


Figure 4. Joint learning network architecture

Next, a Global Average Pooling layer in Re-GlAv block is applied on each temporal frame to get average results, generating a 128-dimensional vector. Both output of C-RNN and CNN are thus concatenated, generate 384-dimensional vectors. Next, these vectors are fed into a DNN-02 architecture, as shown in Table 4, configured by Fl, ReLU, Dr, and Softmax layers for classification. Note that output layer number C depends on specific ASC task due to various datasets evaluated.

Table 4. DNN-02 network architecture.

Blocks	Layers	Output Shape
	Input	384
Re-Fl Block 01	Fl – ReLU – Dr (30%)	2048
Re-Fl Block 02	Fl – ReLU – Dr (30%)	1024
Re-Fl Block 03	Fl – Softmax	C

B. HYPERPARAMETER SETTING FOR THE FRAMEWORK PROPOSED

The joint learning model proposed is built by Tensorflow framework and reused all hyper-parameter setting from C-DNN network experiments. To evaluate the effect of spectrograms, we conduct experiments on the best spectrogram combinations indicated in Table 5. Additionally, we do further investigation of late fusion on accuracy. In particular, we compute more two fusion strategies, called *Max* and *Prod* fusions. Let us consider posterior probability of each spectrogram as $\overline{p}_S = (\overline{p}_{S1}, \overline{p}_{S2}, \ldots, \overline{p}_{SC})$ described in Equation (18), where S is specific spectrogram and C is the number of category classified. Next, the posterior probability of combination with *Prod* strategy $\boldsymbol{p_{f-prod}} = (\boldsymbol{p_1}, \boldsymbol{p_2}, \ldots, \boldsymbol{p_c})$ is obtained by

$$p_c = \frac{1}{S}\prod_{s=1}^{S}\overline{p}_{sc} \quad for\ 1 \le s \le S, \tag{23}$$

where S is the number of spectrograms combined.

Table 5. Comparing individual spectrograms and their combinations with C-DNN architecture to DCASE 2018 Task 1B baseline with best results (%) in bold.

Spectrograms	A	B&C	A&B&C
DCASE baseline	58.9	45.6	52.2
Single spectrogram			
MFCC	64.9	**55.0**	59.9
STFT	59.8	42.7	51.3
log-Mel	**68.2**	54.7	**61.4**
CQT	58.4	47.8	53.1
GAM	64.1	48.9	58.1
Two spectrograms			
CQT+STFT	64.2	55.8	60.0
CQT+GAM	70.9	53.3	62.1
CQT+log-Mel	**72.0**	**60.8**	**66.4**
CQT+MFCC	69.8	58.9	64.4
Three spectrograms			
CQT+GAM+log-Mel	**74.1**	**62.5**	**68.3**
CQT+GAM+MFCC	71.9	61.1	66.5
Four spectrograms			
CQT+GAM+STFT+ log-Mel	**74.4**	**62.5**	**68.5**
CQT+GAM+STFT+MFCC	72.7	60.3	66.5
All five spectrograms			
CQT+GAM+STFT+ MFCC +log-Mel	**73.7**	**62.8**	**68.2**

The posterior probability of combination with *Max* strategy $\boldsymbol{p_{f-max}} = (\boldsymbol{p_1}, \boldsymbol{p_2}, \ldots, \boldsymbol{p_c})$ is obtained by

$$p_c = max(\overline{p}_{1c}, \overline{p}_{2c}, \ldots, \overline{p}_{Sc}). \tag{24}$$

Eventually, the predicted label $\hat{y}$ for either *Max* or *Prod* fusions is determined by Equation (21). As regards ASC datasets evaluated, we conduct extensive experiments on five different datasets, comprising of LITIS Rouen, DCASE 2016 Task 1A, DCASE 2017 Task 1A, and DCASE 2018 Task 1A and 1B. Thus, we compare our best results to the state-of-the-art systems.

VI. EXPERIMENTS AND DISCUSSION

A. PERFORMANCE COMPARISON OVER SPECTROGRAMS AND THEIR COMBINATIONS WITH C-DNN BASELINE

Initially, we equally evaluate all of five individual spectrograms with the baseline C-DNN architecture, thus showing category-wise performance comparison of device A and device B&C of DCASE 2018 Task 1B dataset. Due to the obtained results on device A in Fig. 5, log-Mel, GAM, and MFCC generally outperform STFT and CQT on most categories. As regards the average accuracy, log-Mel and GAM stand on the top, showing competitive results of 68.2% and 67.2%, respectively. Meanwhile, STFT and CQT show very low average scores compared with log-Mel and GAM, indicating a gap performance of nearly 10%. Noticeably, CQT spectrogram showed great performance in sound scenes relating to transportation such as *Bus*, *Metro* and *Street Traffic*. Regarding performance on device B&C in Fig. 6, obtained results show similar with top scores of log-Mel, GAM, and MFCC. Again, CQT spectrogram still shows good accuracy rate in related-transformation categories such as *Bus*, *Metro* and *Tram*.

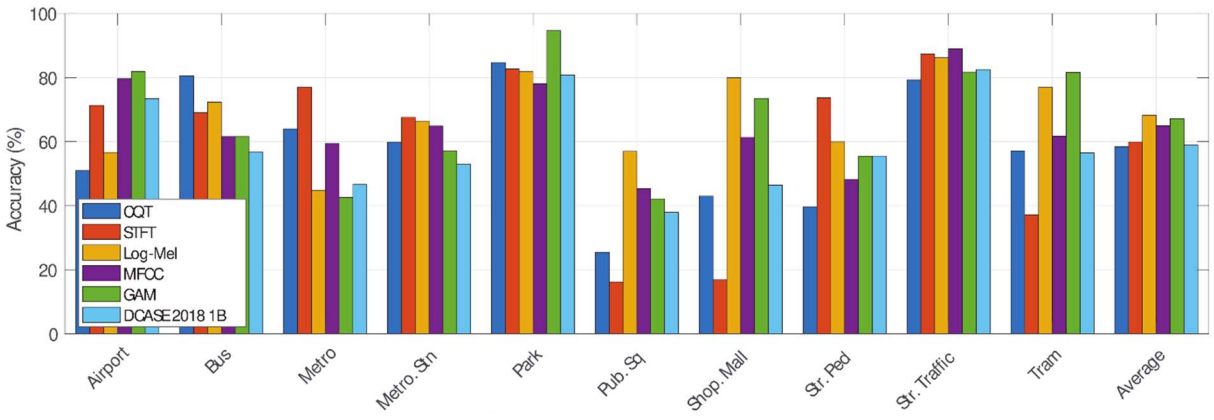


Figure 5. Category-wise performance comparison among spectrograms on device A – DCASE 2018 Task 1B.

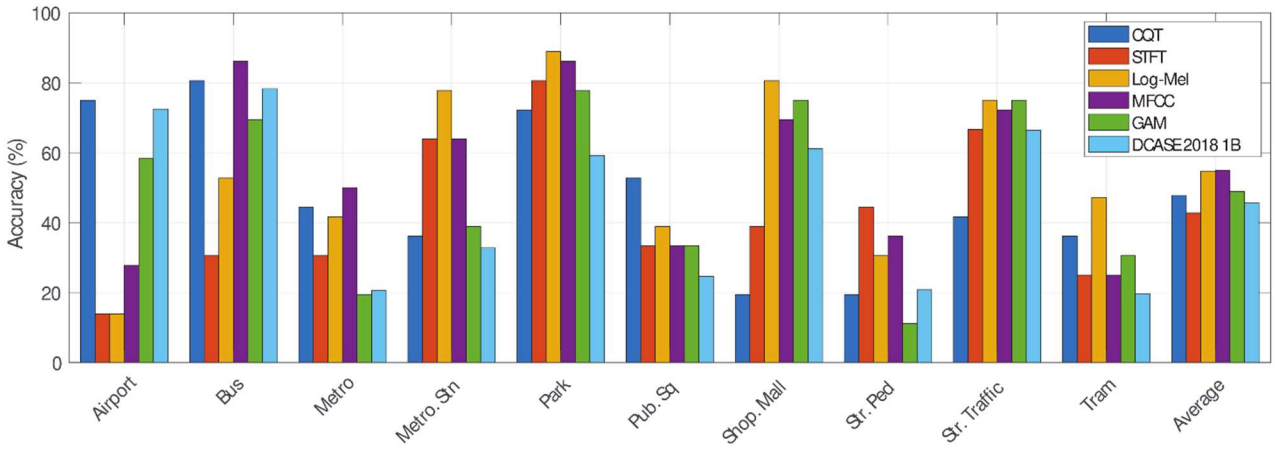


Figure 6. Category-wise performance comparison among spectrograms on device B&C – DCASE 2018 Task 1B.

In general, category-wise performance comparison of device A and device B&C indicates that spectrograms extracted from auditory models such as GAM, log-Mel and MFCC gain high performance. Comparison of these three spectrograms with the DCASE 2018 baseline as regards Task 1B challenge (only device B&C) shows that they outperform the DCASE 2018 baseline over almost categories and achieve an improvement of 3.3%, 9.1%, and 9.4%, respectively in terms of the average result.

Next, we conduct experiments on two-spectrogram combinations and present obtained results in Table 5. As Table 5 shows, CQT+log-Mel achieves the greatest performance on both device A and B&C, improving by 4% and 6%, respectively compared to only log-Mel (the top score of individual spectrogram). Comparing CQT+log-Mel score to DCASE 2018 Task 1B baseline, it is shown a significant improvement of 13.1% and 15.2% over device A and B&C, respectively.

As regards three-spectrogram combinations as shown in Table 5, two analyzed groups of CQT+GAM+MFCC and CQT+GAM +log-Mel show competitive results, reporting 71.9%, 61.1% and 74.1%, 62.5% for device A and B&C, respectively. It indicates that log-Mel and MFCC may contain very similar features.

The results on four-spectrogram combinations of CQT+GAM+STFT+log-Mel witness a minor increase of nearly 0.3% and 0.2% in terms of device A and overall, respectively, compared to CQT+GAM +log-Mel, thanks to the contribution of STFT. Meanwhile, adding STFT into CQT+GAM+MFCC only helps to improve the performance on device A a little, but makes a decrease of 0.8% on device B&C.

As regards the result of all five spectrograms, it even has a downward trend in device A, leading to the decrease of 0.7 %, compared to the accuracy in the best four-spectrogram combination of CQT+GAM+ STFT+log-Mel.

Eventually, we summarized all types of spectrogram combinations, and highlighted which achieved the best scores on all three devices A&B&C. As results show in Table 5, there is a gradual increase in the accuracy rate when combination of spectrograms applied.

In particular, CQT+log-Mel achieves the best performance in two-spectrogram groups, with an increase of nearly 5% compared to the best single spectrogram log-Mel. By adding GAM into group of CQT+log-Mel, it helps to improve by 2% on average. However, a minor increase of 0.2% is observed in the performance of four-spectrogram combination CQT+GAM+STFT+log-Mel until there is no improvement from combination of all five spectrograms. It can be concluded that using multiple spectrograms is effective to improve the performance, thus far exceed the DCASE 2018 Task 1B baseline.

B. PERFORMANCE COMPARISON OVER SPECTROGRAMS AND THEIR COMBINATIONS WITH JOINT LEARNING MODEL

Due to the results obtained in Table 6, adding the C-RNN architecture into the baseline C-DNN network to create the joint learning model helps to improve the performance over both device A and B&C. Especially, the accuracy rate increases when more spectrograms are combined. As regards late fusion methods suggested, *Prod* and *Mean* are very competitive and outperform *Max* fusion scheme in both the baseline C-DNN and joint learning model proposed. Noticeably, joint learning models with *Prod* fusion achieve the best scores for all kinds of spectrogram combinations in terms of B&C performance.

Table 6. Performance comparison (Device A/ B&C %) on DCASE 2018 Task 1B dataset with the highest scores in bold.

Architecture	Joint Learning Model		
	Mean	Prod	Max
CQT+log-Mel	72.4/62.2	**72.1/64.7**	70.9/60.3
CQT+GAM +log-Mel	74.9/65.0	**74.5/66.4**	73.4/61.9
CQT+GAM+STFT +log-Mel	76.2/64.4	**76.5/66.7**	73.6/60.6
CQT+GAM+STFT+ MFCC +log-Mel	76.0/65.3	**76.4/67.5**	74.0/64.7
Architecture	**C-DNN**		
CQT+log-Mel	72.0/60.8	**73.0/62.2**	70.1/59.2
CQT+GAM +log-Mel	74.1/62.5	**74.7/63.3**	72.3/59.2
CQT+GAM+STFT +log-Mel	74.4/62.5	**74.9/60.6**	71.5/58.3
CQT+GAM+STFT+ MFCC +log-Mel	73.7/62.8	**74.6/61.9**	71.5/58.1

Comparing to DCASE 2018 Task 1B baseline, joint learning models proposed outperform DCASE baseline on both

device A and B&C, in particular, the best score of 67.5% over devices B&C, which is obtained from combination of all spectrograms CQT +GAM+STFT+log-Mel+MFCC. This performance is significantly higher than the DCASE baseline by 22%, compared to the lower improved rate of 16% from C-DNN baseline network. It indicates that the strategy of multi-spectrogram input and joint learning model successfully solve the problem of mismatched devices raised in DCASE 2018 Task 1B challenge[1].

C. PERFORMANCE COMPARISON TO THE-STATE-OF-THE-ART SYSTEMS

We continue to evaluate our best proposed systems on various datasets, thus make a comparison to the state-of-the-art at time of writing. As it is shown in detail in Table 7, we achieve the highest accuracy of 99.1% in LITIS Roune dataset. Our performance in DCASE 2016 is 89.2%, which lies in second position on this challenge table, and is ranked as one of the top-three of the state-of-the art systems. However, in DCASE 2017, we are out of top-ten performance in this challenge, with the figure of 67.3%. As regards DCASE 2018 Task 1A dataset, the accuracy of 77.8% obtained ranks in top four and exceeds all state-of-the art systems. Next, we again show our robustness in terms of dealing with mismatched devices issue in DCASE 2018 Task 1B. For instance, we achieve 67.5% in accuracy rate, outperforming systems in state-of-the-art papers and it is very competitive to the top one score in terms of DCASE challenge. It should be noted that there are inconsistencies between the reported results in the DCASE 2018 technical reports and those published in DCASE 2018 challenge website[2]. The accuracy results shown in Table 7, therefore, are collected from the original sources of technical reports.

Table 7. Comparison of state-of-the-art systems with best performance in bold (Upper part: top-ten DCASE challenges; Middle part: State-of-the-art papers; Low part: Our proposed systems using Prod late fusion strategy).

D. 2018 Task 1B (Dev. set)	Acc. (%)	D. 2018 Task 1A (Dev. set)	Acc. (%)	D. 2017 Task 1 (Eva. set)	Acc. (%)	D. 2016 Task 1 (Eva. set)	Acc. (%)	LITIS Roune (20-fold Ave.)	Acc. (%)
Baseline [37]	45.6	Baseline [37]	59.7	Baseline [36]	74.8	Baseline [45]	77.2		
Li [34]	51.7	Li [46]	72.9	Zhao [47]	70.0	Wei [48]	84.1	Bisot [49]	93.4
Tchorz [50]	53.9	Jung [51]	73.5	Jung [51]	70.6	Bae [52]	84.1	Ye [53]	96.0
Kong [54]	57.5	Wang [55]	73.6	Karol [56]	70.6	Kim [57]	85.4	Huy [24]	96.4
Wang [58]	57.5	Christian [59]	74.7	Ivan [60]	71.7	Takahasi [61]	85.6	Yin [62]	96.4
Waldekar [63]	57.8	Zhang [64]	75.3	Park [65]	72.6	Elizalde [66]	85.9	Huy [19]	96.6
Zhao [15]	63.3	Li [34]	76.1	Lehner [67]	73.8	Valenti [68]	86.2	Ye [69]	97.1
Truc [21]	63.6	Dang [70]	76.7	Hyder [71]	74.1	Marchi [9]	86.4	Huy [32]	97.8
		Octave [72]	78.4	Zhengh [73]	77.7	Park [11]	87.2	Zhang [30]	97.9
		Yang [27]	79.8	Han [74]	80.4	Bisot [75]	87.7	Zhang [28]	98.1
		Golubkov [76]	**80.1**	Mun [77]	**83.3**	Hamid [78]	89.7	Huy [31]	98.7
Zhao [15]	63.3	Bai [79]	66.1	Zhao [80]	64.0	Mun [81]	86.3		
Truc [22]	64.7	Gao [82]	69.6	Yang [83]	69.3	Li [84]	88.1		
Truc [85]	66.1	Zhao [15]	72.6	Waldekar [63]	69.9	Hyder [86]	88.5		
Yang [87]	**67.8**	Phaye [16]	74.1	Wu [88]	75.4	Song [89]	89.5		
		Heo [90]	77.4	Chen [91]	77.1	Yin [62]	**91.0**		
log-Mel	58.6	log-Mel	68.0	log-Mel	60.3	log-Mel	80.7	log-Mel	97.9
log-Mel+CQT	64.7	log-Mel+CQT	70.4	log-Mel+CQT	65.8	log-Mel+CQT	89.2	log-Mel+CQT	99.0
CQT+GAM + log-Mel	66.4	CQT+GAM + log-Mel	73.8	CQT+GAM + log-Mel	67.3	CQT+GAM + log-Mel	88.9	CQT+GAM + log-Mel	99.0
CQT+GAM+ STFT+log-Mel	66.7	CQT+GAM+ STFT+log-Mel	77.3	CQT+GAM+ STFT+log-Mel	66.7	CQT+GAM+ STFT+log-Mel	88.7	CQT+GAM+ STFT+log-Mel	**99.1**
CQT+GAM+ STFT+log-Mel +MFCC	67.5	CQT+GAM+ STFT+log-Mel +MFCC	77.8	CQT+GAM+S TFT+log-Mel+MFCC	67.0	CQT+GAM+ST FT+log-Mel+MFCC	88.2	CQT+GAM+S TFT+log-Mel+MFCC	**99.1**

VII. CONCLUSION

This paper has presented a robust framework applying for ASC task. In front-end feature extraction, the idea of providing a comprehensive analysis of low-level spectrogram representation from draw audio signals enables to figure out the effective types of individual spectrograms and their combinations. As regards back-end classification, our novel joint learning network based on parallel convolutional recurrent architecture has facilitated learning both spatial and temporal structural features of spectrograms. By approaching multi-spectrogram input and the joint learning network, we achieve very competitive results compared to the state-of-the-art systems on various ASC datasets of LITIS Rouen and DCASE challenges in three consecutive years 2016, 2017 and 2018, thus prove a general and robust framework for ASC tasks.

References

[1] P. Zwan, "Automatic sound recognition for security purposes," in *Audio Engineering Society Convention*, 124. Audio Engineering Society, 2008, pp. 7387.

[2] X. Valero and Francesc Alias, "Gammatone wavelet features for sound classification in surveillance applications," *Proceedings of the EUSIPCO*, 2012, pp. 1658–1662.

[3] B. N. Schilit, N. Adams, R. Want, et al., Context-aware computing applications, *Xerox Corporation*, Palo Alto Research Center, 1994. https://doi.org/10.1109/WMCSA.1994.16.

[4] T. Heittola, A. Mesaros, A. Eronen, T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1, 2013. https://doi.org/10.1186/1687-

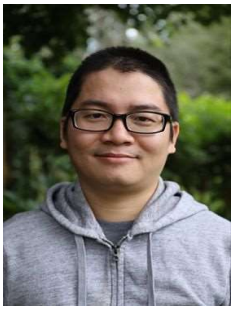
[1] http://dcase.community/challenge2018/

[2] http://dcase.community/challenge2018/

4722-2013-1.
[5] Y. Xu, W. J. Li, K. K. Caramon Lee, *Intelligent Wearable Interfaces*, Wiley Online Library, 2008. https://doi.org/10.1002/9780470222867.
[6] L. Ma, D. J. Smith, B. P. Milner, "Context awareness using environmental noise classification," *Proceedings of the EUROSPEECH*, 2003, pp. 2237-2240.
[7] I. V. McLoughlin, *Speech and Audio Processing: a MATLAB-based Approach*, Cambridge University Press, 2016. https://doi.org/10.1017/CBO9781316084205.
[8] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, S. Squartini, B. Schuller, "Pairwise decomposition with deep neural networks and multiscale kernel subspace learning for acoustic scene classification," *DCASE Technical Report*, 2016, pp. 65–69.
[9] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, B. Schuller, "The up system for the 2016 DCASE challenge using deep recurrent neural network and multiscale kernel subspace learning," *DCASE Technical Report*, 2016.
[10] A. Vafeiadis, D. Kalatzis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, R. Hamzaoui, "Acoustic scene classification: From a hybrid classifier to deep learning," *DCASE Technical Report*, 2017.
[11] S. Park, S. Mun, Y. Lee, H. Ko, "Score fusion of classification systems for acoustic scene classification," *DCASE Technical Report*, 2016.
[12] J. T. Geiger, M. A. Lakhal, B. Schuller, G. Rigoll, "Learning new acoustic events in an hmm-based system using map adaptation," *Proceedings of the INTERSPEECH*, 2011, pp. 293-296. https://doi.org/10.21437/Interspeech.2011-113.
[13] J. T. Geiger, B. Schuller, G. Rigoll, "Large-scale audio feature extraction and SVM for acoustic scene classification," *Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, *WASPAA*, 2013, pp. 1–4. https://doi.org/10.1109/WASPAA.2013.6701857.
[14] L. Vuegen, B. V. D. Broeck, P. Karsmakers, J. F. Gemmeke, B. Vanrumste, H. V. Hamme, "An MFCC-GMM approach for event detection and classification," *Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, *WASPAA*, 2013, pp. 1–3.
[15] H. Phan, L. Hertel, M. Maass, P. Koch, R. Mazur, A. Mertins, "Improved audio scene classification based on label-tree embeddings and convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1278–1290, 2017. https://doi.org/10.1109/TASLP.2017.2690564.
[16] H. Phan, O. Y. Chén, L. Pham, P. Koch, M. De Vos, I. McLoughlin, A. Mertins, "Spatio-temporal attention pooling for audio scene classification," arXiv preprint arXiv:1904.03543, 2019. https://doi.org/10.21437/Interspeech.2019-3040.
[17] H. Phan, O. Y. Chén, P. Koch, L. Pham, I. McLoughlin, A. Mertins, M. De Vos, "Beyond equal-length snippets: How long is sufficient to recognize an audio scene," arXiv preprint arXiv:1811.01095, 2018.
[18] Z. Ren, Q. Kong, J. Han, M. D Plumbley, B. W. Schuller, "Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes," *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, 2019, pp. 56–60. https://doi.org/10.1109/ICASSP.2019.8683434.
[19] S. S. R. Phaye, E. Benetos, Y. Wang, "Subspectralnet – using sub-spectrogram based convolutional neural networks for acoustic scene classification," *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, 2019, pp. 825–829. https://doi.org/10.1109/ICASSP.2019.8683288.
[20] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, T. Virtanen, "Dcase 2017 challenge setup: Tasks, datasets and baseline system," *DCASE Technical Report*, 2017.
[21] T. Lidy, A. Schindler, "CQT-based convolutional neural networks for audio scene classification," *DCASE Technical Report*, vol. 90, pp. 1032–1048, 2016.
[22] H. Phan, P. Koch, F. Katzberg, M. Maass, R. Mazur, A. Mertins, "Audio scene classification with deep recurrent neural networks," arXiv preprint arXiv:1703.04770, 2017. https://doi.org/10.21437/Interspeech.2017-101.
[23] T. Nguyen, F. Pernkopf, "Acoustic scene classification using a convolutional neural network ensemble and nearest neighbor filters," *DCASE Technical Report*, 2018. https://doi.org/10.1109/ICMLA.2019.00151.
[24] T. Nguyen, F. Pernkopf, "Acoustic scene classification with mismatched recording devices using mixture of experts layer," *Proceedings of the 2019 IEEE International Conference on Multimedia and Expo ICME*, 2019, pp. 1666–1671. https://doi.org/10.1109/ICME.2019.00287.
[25] H. Zeinali, L. Burget, J. Cernocky, "Convolutional neural networks and x-vector embedding for dcase2018 acoustic scene classification challenge," arXiv preprint arXiv:1810.04273, 2018.
[26] H. Phan, L. Hertel, M. Maass, P. Koch, A. Mertins, "Label tree embeddings for acoustic scene classification," *Proceedings of the 24th ACM International Conference on Multimedia*, 2016, pp. 486–490. https://doi.org/10.1145/2964284.2967268.
[27] L. Yang, X. Chen, L. Tao, "Acoustic scene classification using multi-scale features," *DCASE Technical Report*, 2018.
[28] T. Zhang, K. Zhang, J. Wu, "Temporal transformer networks for acoustic scene classification," *Proceedings of the Interspeech*, 2018, pp. 1349–1353. https://doi.org/10.21437/Interspeech.2018-1152.
[29] T. Zhang, K. Zhang, J. Wu, "Data independent sequence augmentation method for acoustic scene classification," *Proceedings of the Interspeech*, 2018, pp. 3289–3293.
[30] T. Zhang, K. Zhang, J. Wu, "Multi-modal attention mechanisms in lstm and its application to acoustic scene classification," *Proceedings of the Interspeech*, 2018, pp. 3328–3332. https://doi.org/10.21437/Interspeech.2018-1138.
[31] H. Phan, O. Y. Chén, L. Pham, P. Koch, M. De Vos, I. McLoughlin, A. Mertins, "Spatio-temporal attention pooling for audio scene classification," arXiv preprint arXiv:1904.03543, 2019. https://doi.org/10.21437/Interspeech.2019-3040.
[32] H. Phan, O. Y. Ché, P. Koch, L. Pham, I. McLoughlin, A. Mertins, M. De Vos, "Beyond equal-length snippets: How long is sufficient to recognize an audio scene?," arXiv preprint arXiv:1811.01095, 2018.
[33] C. Gousseau, "VGG CNN for urban sound tagging" *DCASE Technical Report*, 2019.
[34] Z. Li, L. Zhang, S. Du, W. Liu, "Acoustic scene classification based on binaural deep scattering spectra with CNN and LSTM," *DCASE Technical Report*, 2018.
[35] A. Rakotomamonjy, G. Gasso, "Histogram of gradients of time–frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, 2014. https://doi.org/10.1109/TASLP.2014.2375575.
[36] A. Mesaros, T. Heittola, T. Virtanen, "TUT database for acoustic scene classification and sound event detection," *Proceedings of the 2016 24th European Signal Processing Conference EUSIPCO*, 2016, pp. 1128-1132. https://doi.org/10.1109/EUSIPCO.2016.7760424.
[37] A. Mesaros, T. Heittola, T. Virtanen, "A multi- device dataset for urban acoustic scene classification," *DCASE Technical Report*, 2018.
[38] J. Salamon, J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017. https://doi.org/10.1109/LSP.2017.2657381.
[39] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, O. Nieto, "librosa: Audio and music signal analysis in python," *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–25. https://doi.org/10.25080/Majora-7b98e3ed-003.
[40] R. D. Patterson, "Auditory filters and excitation patterns as representations of frequency resolution," *Frequency selectivity in hearing*, 1986.
[41] B. R. Glasberg, B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990. https://doi.org/10.1016/0378-5955(90)90170-T.
[42] D. P. W. Ellis, "Gammatone-like spectrograms," 2009. [Online]. Available at: http://www.ee.columbia.edu/dpwe/resources/matlab/gammatonegram.
[43] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[44] S. Kullback, R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951. https://doi.org/10.1214/aoms/1177729694.
[45] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018. https://doi.org/10.1109/TASLP.2017.2778423.
[46] Y. Li, X. Li, Y. Zhang, "The SEIE-SCUT systems for challenge on DCASE 2018: Deep learning techniques for audio representation and classification," *DCASE Technical Report*, 2018.
[47] S. Zhao, T. N. T. Nguyen, W.-S. Gan, D. L. Jones, "Acoustic scene classification using deep residual convolutional neural networks," in DCASE Technical Report, 2017.
[48] D. Wei, J. Li, P. Pham, S. Das, S. Qu, "Acoustic scene recognition with deep neural networks (DCASE challenge 2016)," *DCASE Technical Report*, 2016.
[49] V. Bisot, S. Essid, G. Richard, "Hog and subband power distribution image features for acoustic scene classification," *Proceedings of the 2015 23rd European Signal Processing Conference EUSIPCO*, 2015, pp. 719–

723. https://doi.org/10.1109/EUSIPCO.2015.7362477.
[50] J. Tchorz, M. Weg, "Combination of amplitude modulation spectrogram features and MFCCS for acoustic scene classification," *DCASE Technical Report*, 2018.
[51] J.-W. Jung, H.-S. Heo, I. H. Yang, S.-H. Yoon, H.-J. Shim, H.-J. Yu, "DNN-based audio scene classification for DCASE 2017: dual input features, balancing cost, and stochastic data duplication," *DCASE Technical Report*, 2017.
[52] S. H. Bae, I. Choi, N. S. Kim, "Acoustic scene classification using parallel combination of lstm and CNN," *DCASE Technical Report*, 2016, pp. 11–15.
[53] J. Ye, T. Kobayashi, M. Murakawa, T. Higuchi, "Acoustic scene classification based on sound textures and events," *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015, pp. 1291–1294. https://doi.org/10.1145/2733373.2806389.
[54] Q. Kong, T. Iqbal, Y. Xu, W. Wang, M. D. Plumbley, "DCASE 2018 challenge surrey cross-task convolutional neural network baseline," arXiv preprint arXiv:1808.00773, 2018.
[55] J. Wang, "DCASE 2018 task 1A: Acoustic scene classification by bi-LSTM-CNN-net multichannel fusion," *DCASE Technical Report*, 2018.
[56] K. J. Piczak, "The details that matter: Frequency resolution of spectrograms in acoustic scene classification," *DCASE Technical Report*, 2017.
[57] J. Kim, K. Lee, "Empirical study on ensemble method of deep neural networks for acoustic scene classification," *DCASE Technical Report*, 2016.
[58] J. Wang, S. Li, "Self-attention mechanism based system for DCASE 2018 challenge task1 and task4," *DCASE Technical Report*, 2018, pp. 1–5.
[59] C. Roletscheck, T. Watzka, A. Seiderer, D. Schiller, E. Andre, "Using an evolutionary approach to explore convolutional neural networks for acoustic scene classification," *DCASE Technical Report*, 2018.
[60] I. Kukanov, V. Hautamaki, K. A. Lee, "Recurrent neural network and maximal figure of merit for acoustic event detection," *DCASE Technical Report*, 2017.
[61] G. Takahashi, T. Yamada, S. Makino, N. Ono, "Acoustic scene classification using deep neural network and frame- concatenated acoustic feature," *DCASE Technical Report*, 2016. https://doi.org/10.1109/APSIPA.2017.8282314.
[62] Y. Yin, R. R. Shah, R. Zimmermann, "Learning and fusing multimodal deep features for acoustic scene categorization," *Proceedings of the 26th ACM International Conference on Multimedia*, 2018, pp. 1892–1900. https://doi.org/10.1145/3240508.3240631.
[63] S. Waldekar, G Saha, "Wavelet-based audio features for acoustic scene classification," *DCASE Technical Report*, 2018. https://doi.org/10.21437/Interspeech.2018-2083.
[64] L. Zhang, J. Han, "Acoustic scene classification using multi- layered temporal pooling based on deep convolutional neural network," *DCASE Technical Report*, 2018.
[65] S. Park, S. Mun, Y. Lee, H. Ko, "Acoustic scene classification based on convolutional neural network using double image features," *DCASE Technical Report*, 2017, pp. 1–5.
[66] B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, I. Lane, "Experiments on the DCASE challenge 2016: Acoustic scene classification and sound event detection in real life recording," arXiv preprint arXiv:1607.06706, 2016.
[67] B. Lehner, H. Eghbal-Zadeh, M. Dorfer, F. Korzeniowski, K. Koutini, G. Widmer, "Classifying short acoustic scenes with i-vectors and CNNS: Challenges and optimisations for the 2017 DCASE ASC task," *DCASE Technical Report*, 2017.
[68] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," *DCASE Technical Report*, pp. 95–99, 2016.
[69] J. Ye, T. Kobayashi, N. Toyama, H. Tsuda, M. Murakawa, "Acoustic scene classification using efficient summary statistics and multiple spectro-temporal descriptor fusion," *Applied Sciences*, vol. 8, no. 8, pp. 1363, 2018. https://doi.org/10.3390/app8081363.
[70] A. Dang, T. Vu, J.-C. Wang, "Acoustic scene classification using ensemble of convnets," *DCASE Technical Report*, 2018.
[71] R. Hyder, S. Ghaffarzadegan, Z. Feng, T. Hasan, "Buet Bosch consortium (b2c) acoustic scene classification systems for DCASE 2017," *DCASE Technical Report*, 2017.
[72] O. Mariotti, M. Cord, O. Schwander, "Exploring deep vision models for acoustic scene classification," *DCASE Technical Report*, 2018.
[73] Z. Weiping, Y. Jiantao, X. Xiaotao, L. Xiangtao, P. Shaohu, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," *DCASE Technical Report*, 2017.
[74] Y. Han, J. Park, K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," *DCASE Technical Report*, pp. 1–5, 2017.
[75] V. Bisot, R. Serizel, S. Essid, G. Richard, "Supervised nonnegative matrix factorization for acoustic scene classification," *DCASE Technical Report*, pp. 62–69, 2016.
[76] A. Golubkov, A. Lavrentyev, "Acoustic scene classification using convolutional neural networks and different channels representations and its fusion," *DCASE Technical Report*, 2018.
[77] S. Mun, S. Park, D. K. Han, H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane," *DCASE Technical Report*, pp. 93–97, 2017.
[78] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, G. Widmer, "CP-JKU submissions for DCASE-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks," *DCASE Technical Report*, 2016.
[79] X. Bai, J. Du, Z.-R. Wang, C.-H. Lee, "A hybrid approach to acoustic scene classification based on universal acoustic models," *Proceedings of the Interspeech 2019*, pp. 3619–3623, 2019. https://doi.org/10.21437/Interspeech.2019-2171.
[80] Z. Ren, K. Qian, Y. Wang, Z. Zhang, V. Pandit, A. Baird, B. Schuller, "Deep scalogram representations for acoustic scene classification," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 3, pp. 662–669, 2018. https://doi.org/10.1109/JAS.2018.7511066.
[81] S. Mun, S. Shon, W. Kim, D. K. Han, H. Ko, "Deep neural network based learning and transferring mid-level audio features for acoustic scene classification," *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, 2017, pp. 796–800. https://doi.org/10.1109/ICASSP.2017.7952265.
[82] L. Gao, H. Mi, B. Zhu, D. Feng, Y. Li, Y. Peng, "An adversarial feature distillation method for audio classification," *IEEE Access*, vol. 7, pp. 105319–105330, 2019. https://doi.org/10.1109/ACCESS.2019.2931656.
[83] Y. Yang, H. Zhang, W. Tu, H. Ai, L. Cai, R. Hu, F. Xiang, "Kullback–leibler divergence frequency warping scale for acoustic scene classification using convolutional neural network," *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, 2019, pp. 840–844. https://doi.org/10.1109/ICASSP.2019.8683000.
[84] J. Li, W. Dai, F. Metze, S. Qu, S. Das, "A comparison of deep learning methods for environmental sound detection," *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, 2017, pp. 126–130. https://doi.org/10.1109/ICASSP.2017.7952131.
[85] T. Nguyen, F. Pernkopf, "Acoustic scene classification with mismatched recording devices using mixture of experts layer," *Proceedings of the 2019 IEEE International Conference on Multimedia and Expo ICME*, 2019, pp. 1666–1671. https://doi.org/10.1109/ICME.2019.00287.
[86] R. Hyder, S. Ghaffarzadegan, Z. Feng, J. H. L. Hansen, T. Hasan, "Acoustic scene classification using a CNN-supervector system trained with auditory and spectrogram image features," *Proceedings of the Interspeech*, 2017, pp. 3073–3077. https://doi.org/10.21437/Interspeech.2017-431.
[87] L. Yang, L. Tao, X. Chen, X. Gu, "Multi-scale semantic feature fusion and data augmentation for acoustic scene classification," *Applied Acoustics*, vol. 163, pp. 107238, 2020. https://doi.org/10.1016/j.apacoust.2020.107238.
[88] Y. Wu, T. Lee, "Enhancing sound texture in CNN-based acoustic scene classification," *Proceedings of the 2019 IEEE International Conference on Multimedia and Expo ICASSP*, 2019, pp. 815–819. https://doi.org/10.1109/ICASSP.2019.8683490.
[89] H. Song, J. Han, S. Deng, "A compact and discriminative feature based on auditory summary statistics for acoustic scene classification," arXiv preprint arXiv:1904.05243, 2019. https://doi.org/10.21437/Interspeech.2018-1299.
[90] H.-S. Heo, J.-W. Jung, H.-J. Shim, H.-J. Yu, "Acoustic scene classification using teacher-student learning with soft-labels," arXiv preprint arXiv:1904.10135, 2019. https://doi.org/10.21437/Interspeech.2019-1989.
[91] H. Chen, P. Zhang, Y. Yan, "An audio scene classification framework with embedded filters and a DCT-based temporal module," *Proceedings of the 2019 IEEE International Conference on Multimedia and Expo ICASSP*, 2019, pp. 835–839. https://doi.org/10.1109/ICASSP.2019.8683636.

***DAT NGO**, is currently a PhD student at the School of Computer Science and Electronic Engineering, University of Essex, UK. He particularly investigates topics in the fields of acoustic scene classification and signal processing.*



***LAM PHAM**, received the Bachelor of Engineering, and Master of Science degree in Electronics-Telecommunication Engineering from Ho Chi Minh City University of Technology in 2009 and 2012, respectively. He completed PhD in University of Kent, UK in 2021 and now working as Data Scientist in Austrian Institute of Technology, Austria.*

***ANH NGUYEN**, is a Senior at Ho Chi Minh City University of Technology with major of Electronics and Communication Engineering. His research interests include object detection and classification in sound and image.*

***TIEN LY**, is currently a DPhil student in Engineering Science at the University of Oxford. She received a BEng in Control Engineering and Automation from Ho Chi Minh City University of Technology, Ho Chi Minh City in 2019 and an MSc in Computer Science (Artificial Intelligence) from the University of Nottingham in 2020. She is now working on machine learning techniques in robotics, particularly loco-manipulation systems.*

***KHOA PHAM**, is currently a Lecturer at Department of Computer and Communication Engineering, Ho Chi Minh City, Vietnam. His research interests include Low-power VLSI solutions for Memory, Energy Harvesting, Memristor-based Neuromorphic Computing Systems, Neural Network Accelerators and MP-System-On-Chip based designs.*

***THANH NGO**, works as a Lecturer at Electrical Engineering Department, University of Danang – University of Science and Technology, Danang City, Vietnam. His research interests include run-time mapping techniques for MPSoC, industrial communication networks, supervisory control and data acquisition of embedded multiprocessor systems.*