Intelligent Audio Systems

Prof. Cheng Siong CHIN

cheng.chin@ncl.ac.uk

Newcastle University

Dr Cheng Siong Chin

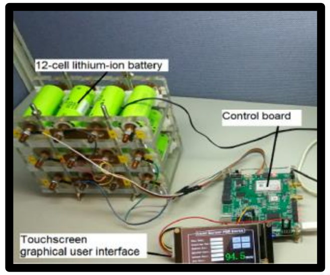
Research Projects

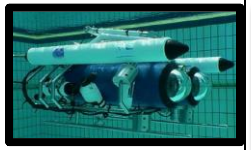
Research Interests:

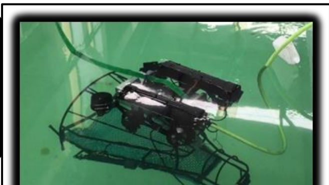
Intelligent Systems Design of complex systems in uncertain environment involving Predictive Analytics (data mining, predictive modelling and machine learning).

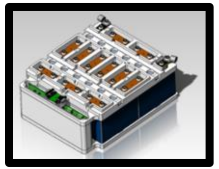
Research Projects:

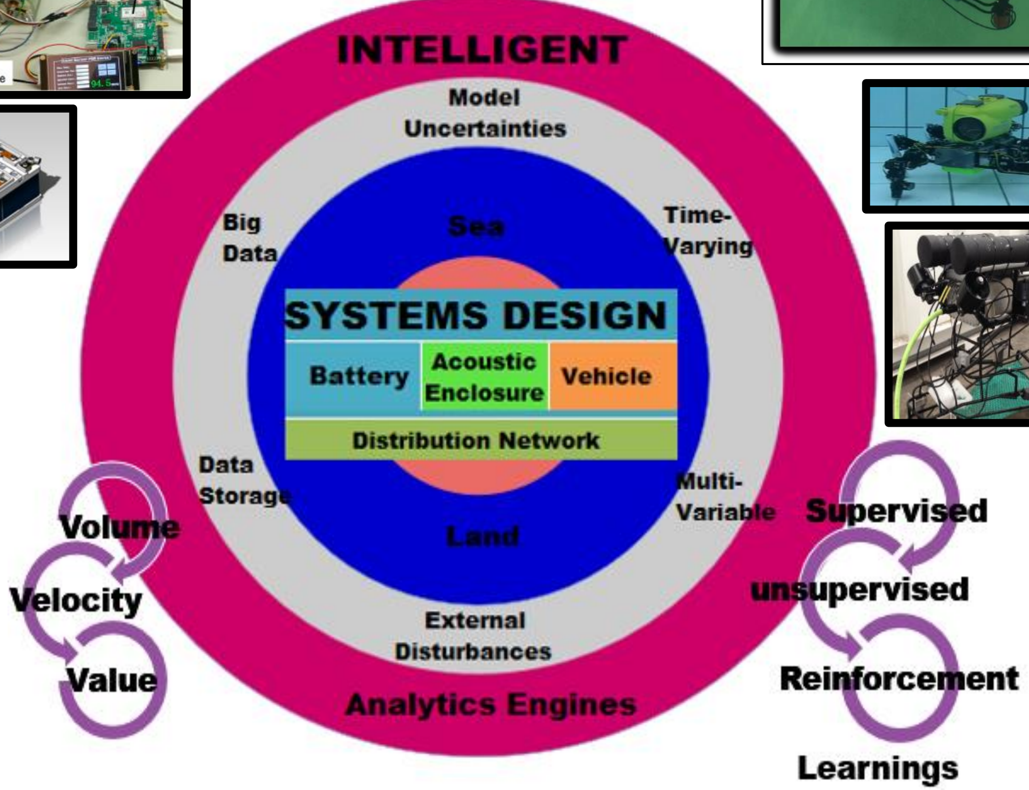
- 2019-present: EMA grant on **AI** System for **Energy Storage in** Hot and Humid Climate
- 2020-present: EDB-IPP grant on **Smart Manufacturing** for Yield Improvement
- 2018- present: EDB-IPP grant on **Smart** and High Precision Leakage Localization.
- 2018 to present: EDB-IPP grant on **Intelligent** Leakage Warning System.
- 2016 to 2019: SMI on **Intelligent** Software Tool for **Noise** Modelling and Prediction.
- 2013 to 2016: SMI on the **Battery** Power System for **ROV**.
- 2013 to 2015: Defence Innovative Research Programme Project on **AUV** Docking Hoop Control
- 2013 to 2017: EDB-IPP grant on **Noise** and Vibration Control of Offshore Structure.
- 2013 to 2018: EDB-IPP grant on Vibration and Psycho-**Acoustic** Parameters in Hard Disk Drive.

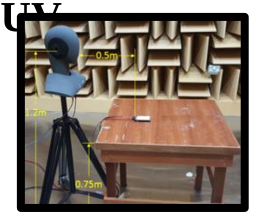

12-cell lithium-ion battery
Control board
Touchscreen graphical user interface

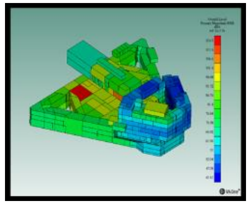


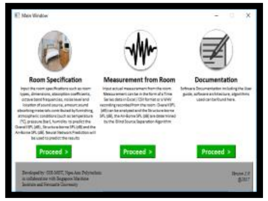












Main Window

Room Specification

Input the room specifications such as room types, dimensions, absorption coefficients, octave band frequencies, noise level and location of sound source, amount sound absorbing materials contributed by furnishing, atmospheric conditions (such as temperature (°C), pressure (bar), humidity to predict the Overall SPL (dB), Structure-borne SPL (dB) and the Air-Borne SPL (dB). Neural Network Prediction will be used to predict the results.

Proceed >

Measurement from Room

Input actual measurement from the room. Measurement can be in the form of a Time Series data in Excel / CSV format or a WAV recording recorded from the room. Overall SPL (dB) can be analysed and the Structure-borne SPL (dB), the Air-Borne SPL (dB) are determined by the Blind Source Separation Algorithm.

Proceed >

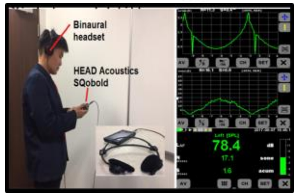
Documentation

Software Documentation including the User guide, software architecture, algorithms used can be found here.

Proceed >

Developed by: [illegible], Ngee Ann Polytechnic
in collaboration with Singapore Maritime
Institute and Newcastle University

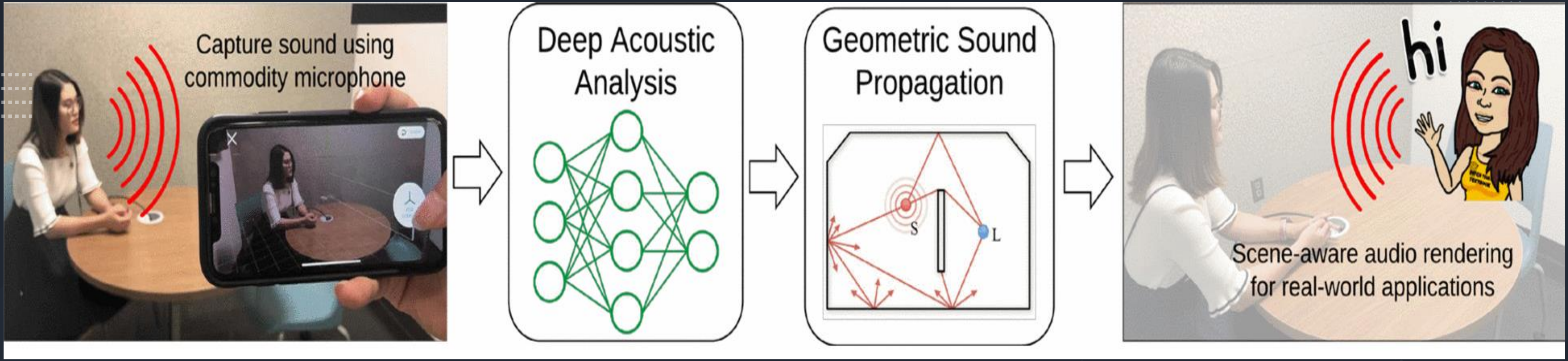
Version 2.0
©2017



Email: cheng.chin@ncl.ac.uk

Intelligent Systems→ Problem, Need and Industrial Relevance

- Intelligent Systems heavily rely of context awareness
- Context awareness is the ability of a system or system component to gather information about its environment at any given time and adapt behaviors accordingly.
- Being able to detect and classify sounds and events from the surroundings is an important aspect of context awareness.


Capture sound using commodity microphone
Deep Acoustic Analysis
Geometric Sound Propagation
S
L
hi
Scene-aware audio rendering for real-world applications

https://www.computer.org/csdl/journal/tg/2020/05/08998301/1hpPBqG2djy

Background

□ Sound Classification is one of the most widely used applications in Audio Deep Learning.

□ It involves learning to classify sounds and to predict the category of that sound.

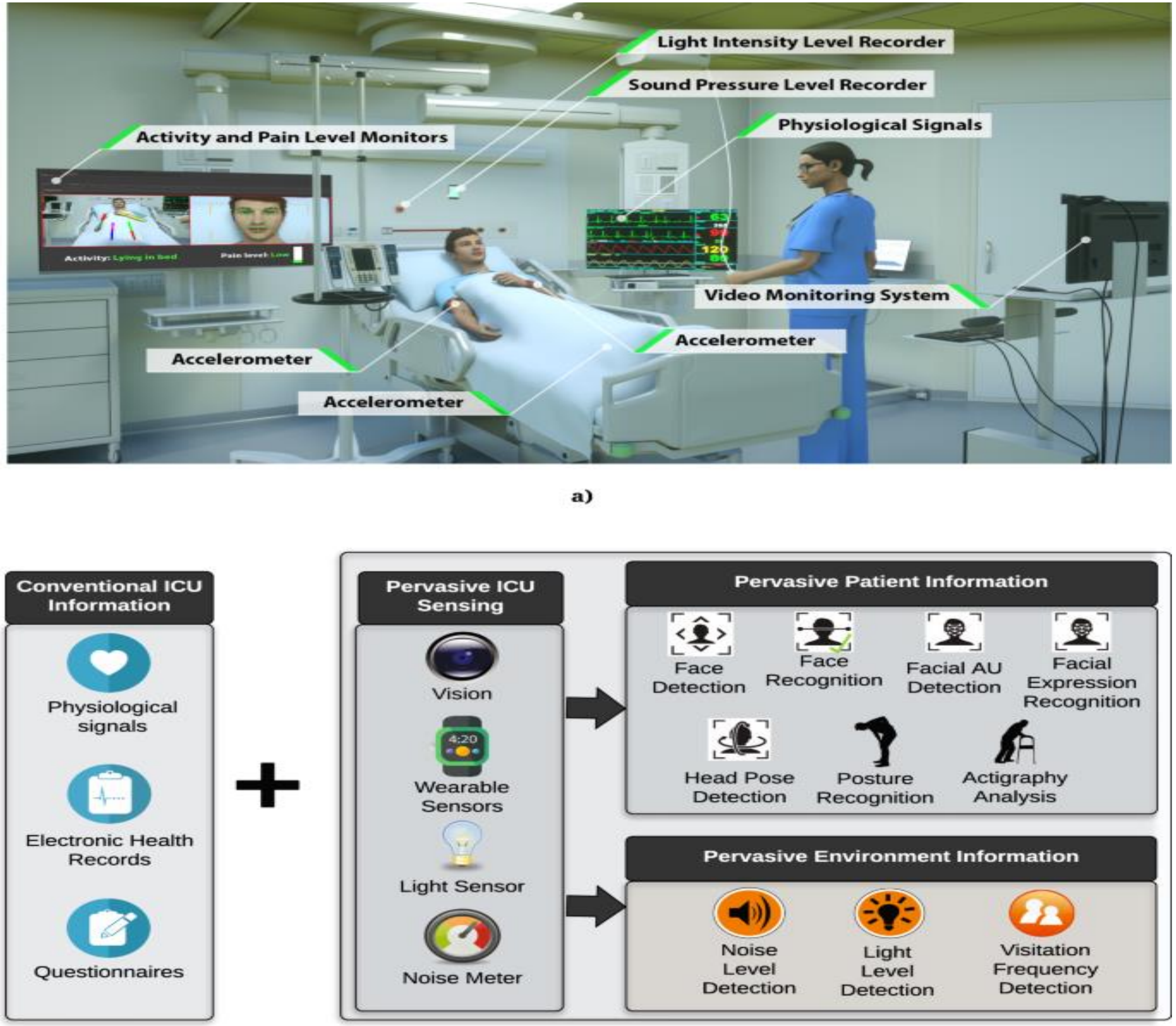
□ This type of problem can be applied to many practical scenarios e.g. classifying music clips to identify the genre of the music, acoustic scenes classification, and sound events detection.

□ It serves many important applications.

For example, to capture the acoustic characteristics of real-world rooms using commodity devices and use the captured characteristics to generate similar sounding sources with virtual models.

Background

- AI in the critical care setting could reduce nurses’ workload to allow them to spend time on more critical tasks, and could also augment human decision-making by offering low-cost and high capacity intelligent data processing.
- How pervasive sensing technology and AI can be used for monitoring patients and their environment in the ICU.
- Wearable accelerometer sensors, a light sensor, a sound sensor, and a high-resolution camera to capture data on patients and their environment in the ICU



a)

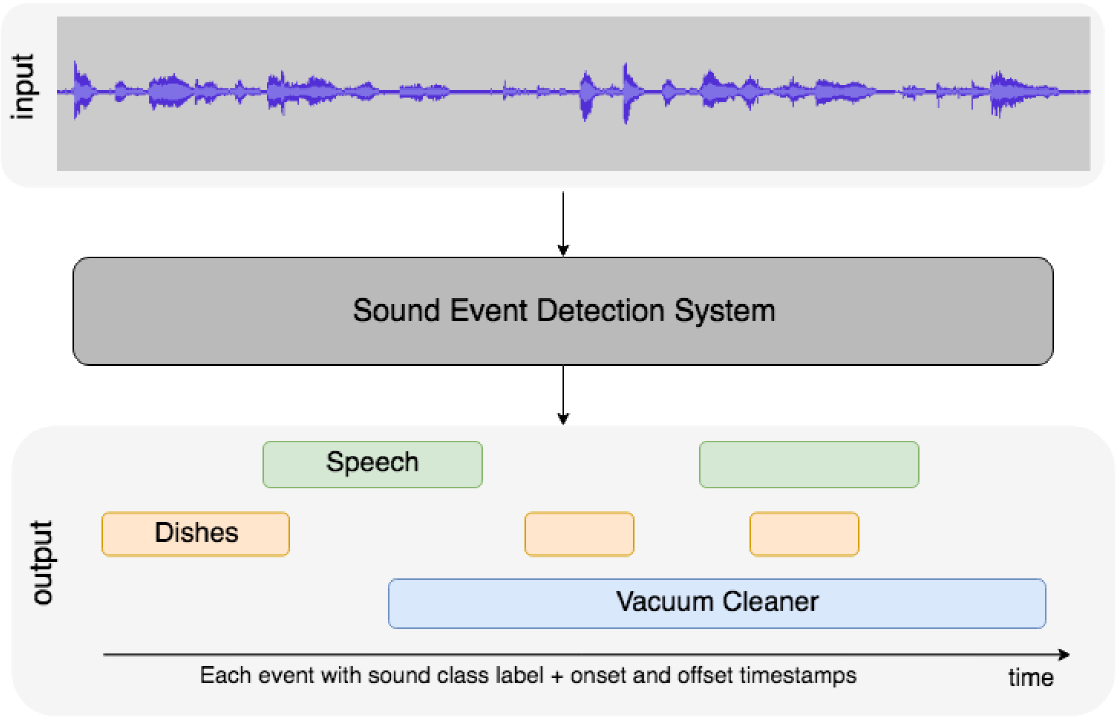
b)

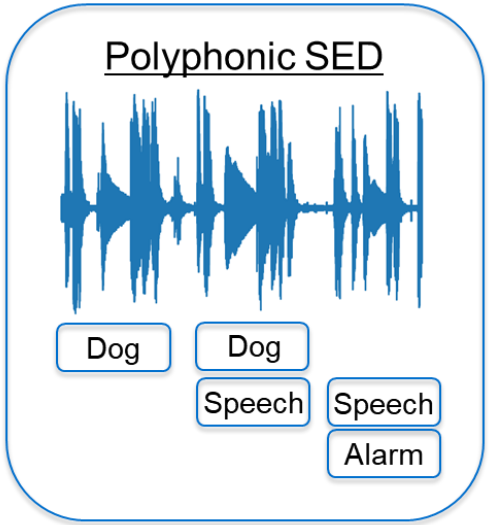
https://www.nature.com/articles/s41598-019-44004-w

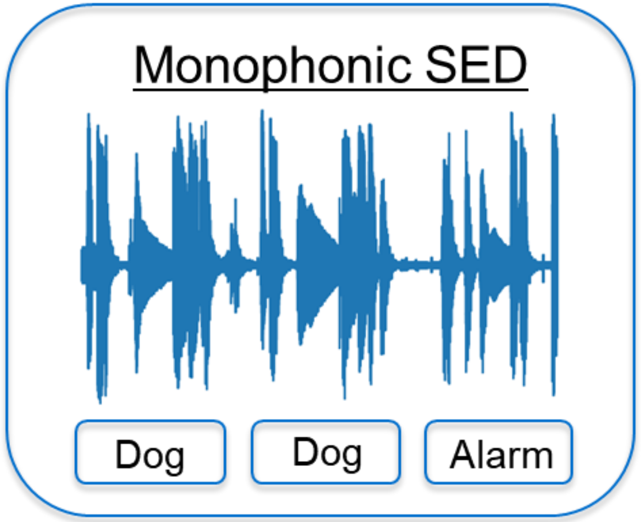
Sound Events Detection (SED)

Sound Event Detection

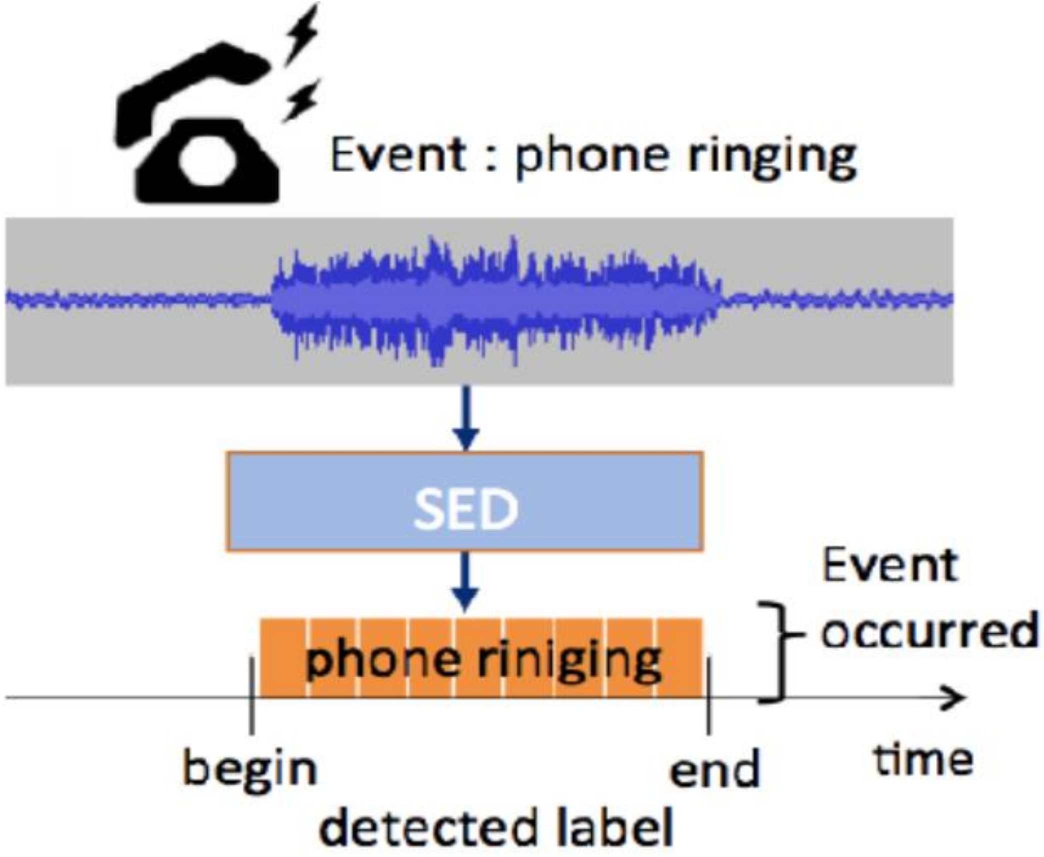
- ❑ A Sound Event Detection (SED) refers to a system that is capable of
    - ❑ Detecting the type of acoustic event present in an audio clip (i.e., audio tagging).
    - ❑ Returning the onset and offset of the identified acoustic event (i.e., temporal localization).
- ❑ A SED system can be categorized as a monophonic or polyphonic system.
- ❑ A polyphonic SED system is more appropriate in real-life application as an audio clip is more likely to contain multiple sound events.


input
Sound Event Detection System
Speech
Dishes
Vacuum Cleaner
output
Each event with sound class label + onset and offset timestamps
time

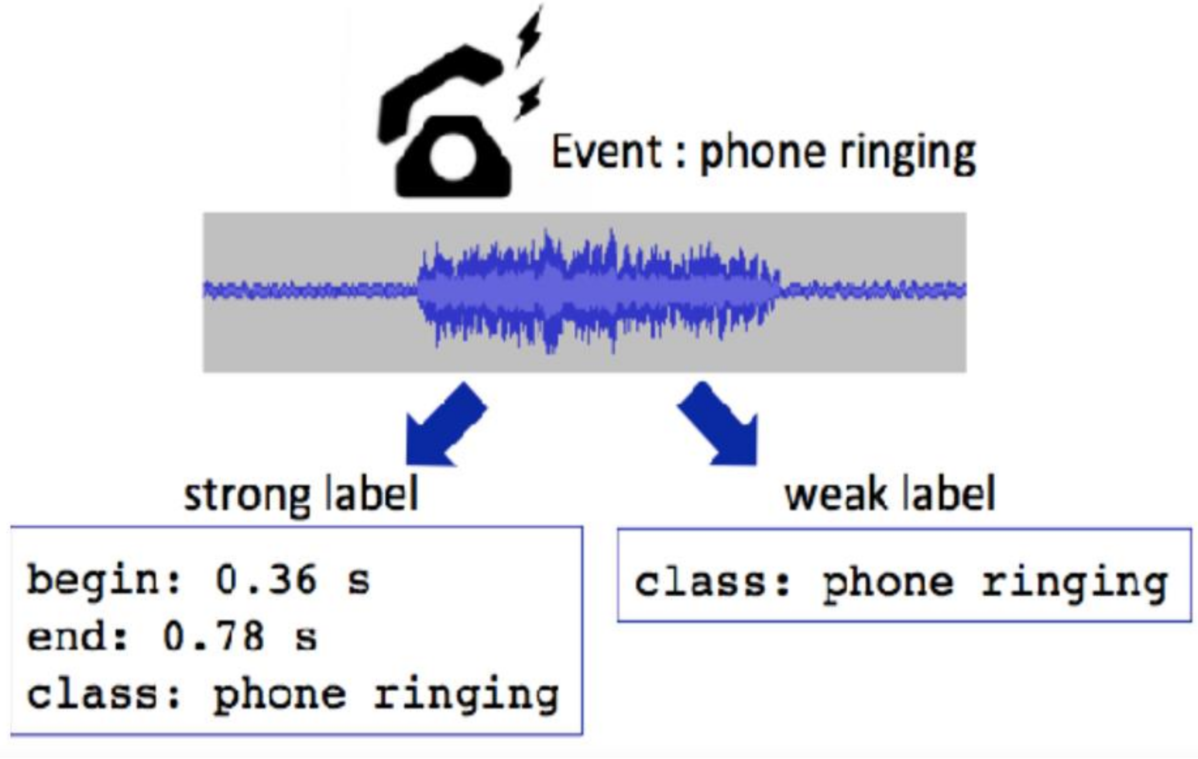




Sound Event Detection



https://www.semanticscholar.org/paper/Weakly-Labeled-Learning-Using-BLSTM-CTC-for-Sound-Matsuyoshi-Komatsu/401296f7d4058ef1702a77874c95a759f3868fc4



Weakly labeled data- where only the event tags are known with certainty → can compromise the solution.

Sound Event Detection

- ❑ How is a SED system useful?
  - ❑ Not affected by the degree of illumination.
  - ❑ Some events can only be detected by sound.
  - ❑ Sound can capture the immediate attention of an individual.
  - ❑ An audio clip requires lesser computational resources than an image or video

Challenges for Sound Event Detection

- ❑System development may require a large amount of strongly labeled data, where event tags and their corresponding onset and offsets are known with certainty
    - ❑Difficult and time consuming to collect.
    - ❑Accuracy of onset and offset annotation is ambiguous due to the fade in and fade out effect.
    - ❑Such dataset is usually limited to minutes or a few hours.

DEMO

- Sound Classification System- apply the audio tagging system to build a sound event detection (SED) system.
- The SED prediction is obtained by applying the audio tagging system on consecutive 2-second segments.

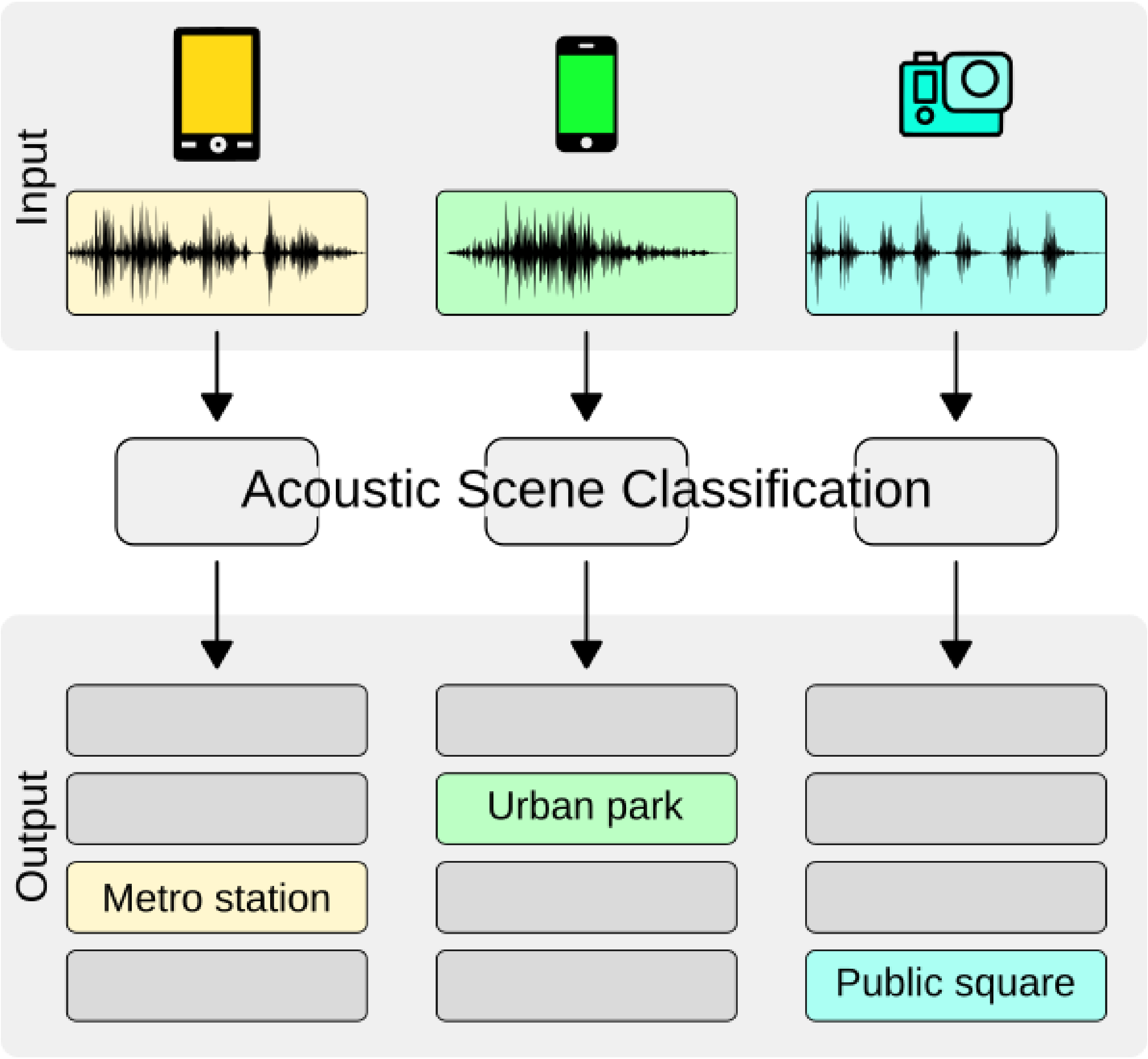
VIDEO

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, Mark D. Plumbley. "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition." arXiv preprint arXiv:1912.10211 (2019).

Acoustic Scenes Classification (ASC)

Acoustic Scene Classification

- ❑Acoustic scene classification (ASC) categorizes an audio file based on the environment in which it has been recorded.
- ❑ For example, →adjustment of a smartphone ring volume when its owner moves from a quiet acoustic environment into a noisier one.



Challenges for Acoustic Scene Classification

❑ Time consuming to collect for each scene.

❑ Dataset is usually limited to minutes or a few hours. Thus, short audio segments provide less information, making ASC difficult.

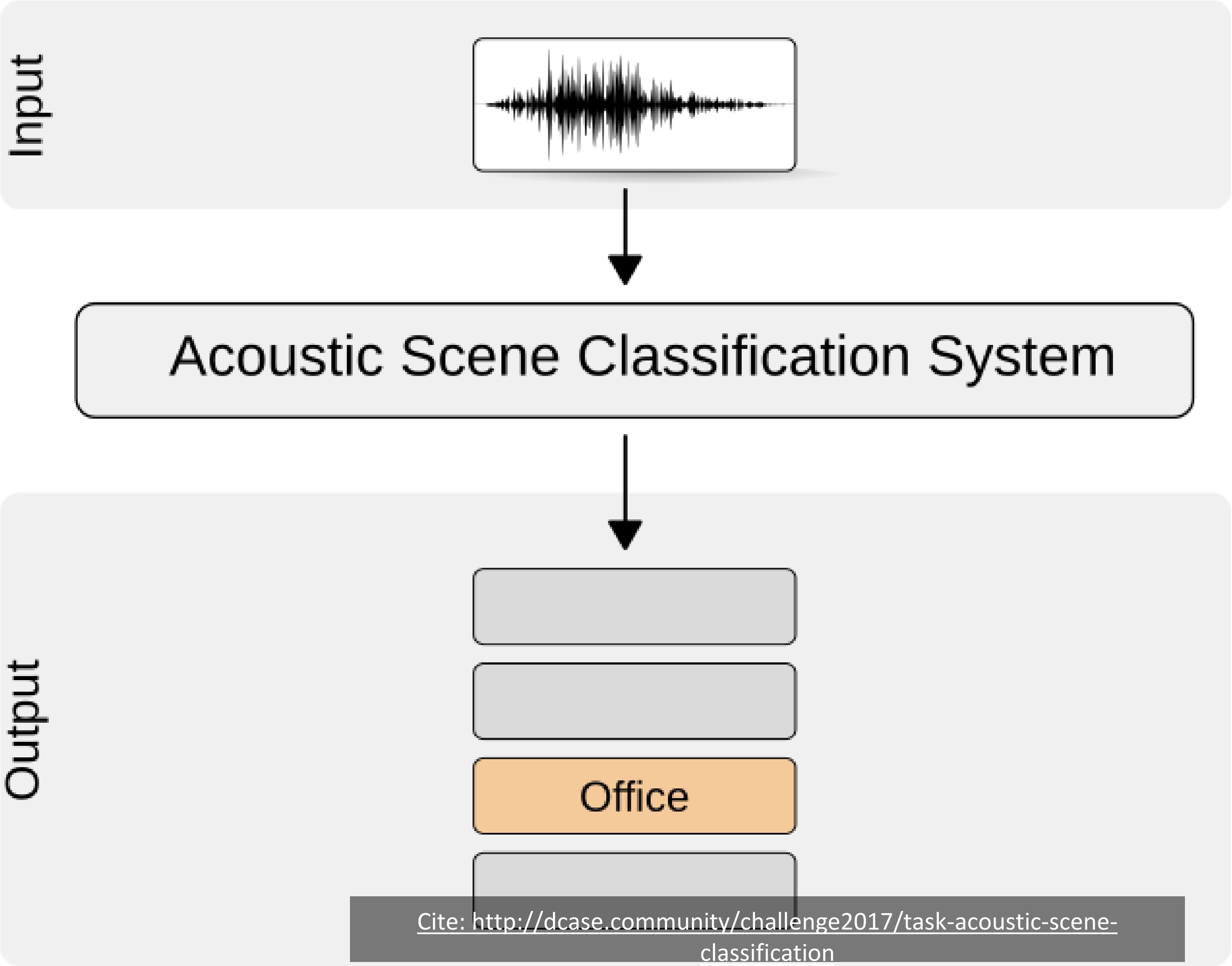
❑ Acoustic scene can share very similar acoustic profile making it difficult to be classified.

❑ Ambient sound scenes typically comprise multiple short events occurring on top of a somewhat stationary background.

❑ foreground-background ambient sound scene separation is challenging.

Introduction

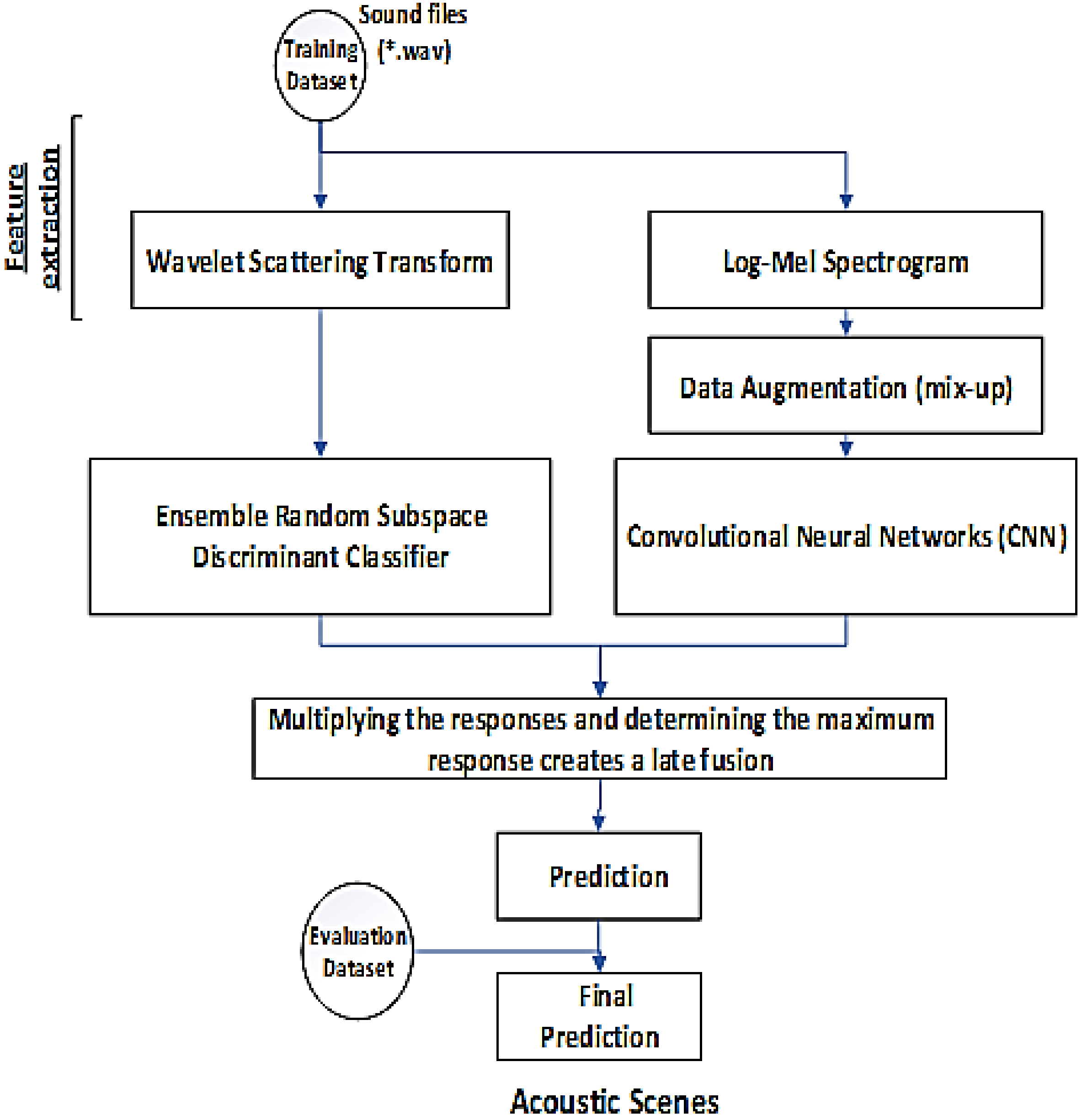
- The top-performing systems in Detection and Classification of Acoustic Scenes and Events (DCASE) challenge consist of multiple models using CNN.
- Instead of log-Mel spectrogram, for feature extraction, wavelet time scattering has been used.
- The extracted features are insensitive to translation, rotation, and small deformation.
- Wavelet scattering transform produce a more accurate classifier in dealing with variation in the ASC.



Cite: http://dcase.community/challenge2017/task-acoustic-scene-classification

Introduction

- **Overview:** We will start with sound files, convert them into spectrograms, input them into a CNN plus Linear Classifier model, and produce predictions about the class to which the sound belongs.
- The **wavelet features** are used for ensemble classifiers (different discriminant analysis learners-linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), Regularized linear discriminant analysis (RDA)).
- To reduce the overfitting of the dataset in CNN, a Mixup algorithm is used. They are used to train a CNN.
- A multi-model late fusion is then used to fuse → log-Mel CNN & Wavelet ensemble classifier.
- It will provide a time-frequency representation of audio to capture Spectro-temporal modulation patterns for identifying various acoustic scenes.


Sound files
(*.wav)
Training Dataset
Feature extraction
Wavelet Scattering Transform
Log-Mel Spectrogram
Data Augmentation (mix-up)
Ensemble Random Subspace Discriminant Classifier
Convolutional Neural Networks (CNN)
Multiplying the responses and determining the maximum response creates a late fusion
Prediction
Evaluation Dataset
Final Prediction
Acoustic Scenes

Datasets

1) bus
2) forest path
3) home
4) city center
5) cafe
6) lakeside beach (outdoor)
7) library (indoor)
8) car
9) grocery store
10) urban park (outdoor)
11) office: multiple persons (indoor)
12) metro station (indoor)
13) train (traveling, vehicle)
14) residential area (outdoor)
15) tram (traveling, vehicle)

The TUT Acoustic Scenes 2017 dataset was used for the experiments.

It consists of 10-seconds audio segments from 15 acoustic scenes

Each acoustic scene has 312 segments giving a total of 52 (=312*10/60) minutes of audio.

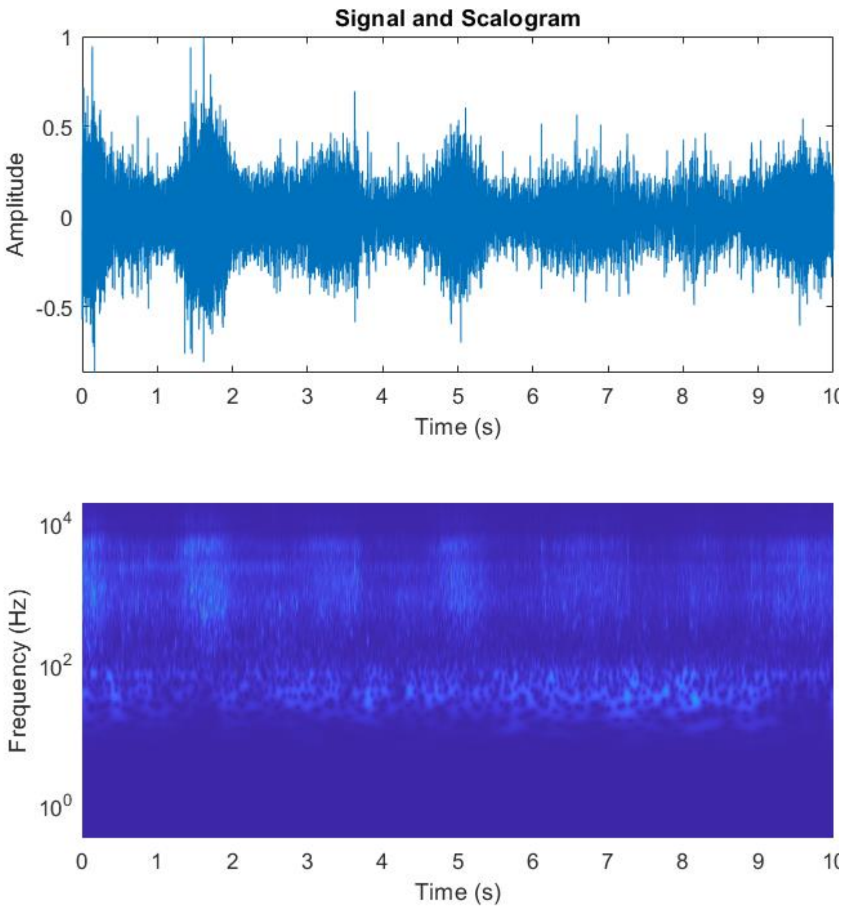
Sound recordings were performed via different devices at 24-bit resolution and 44100Hz sampling rate. The microphones are worn during recording.

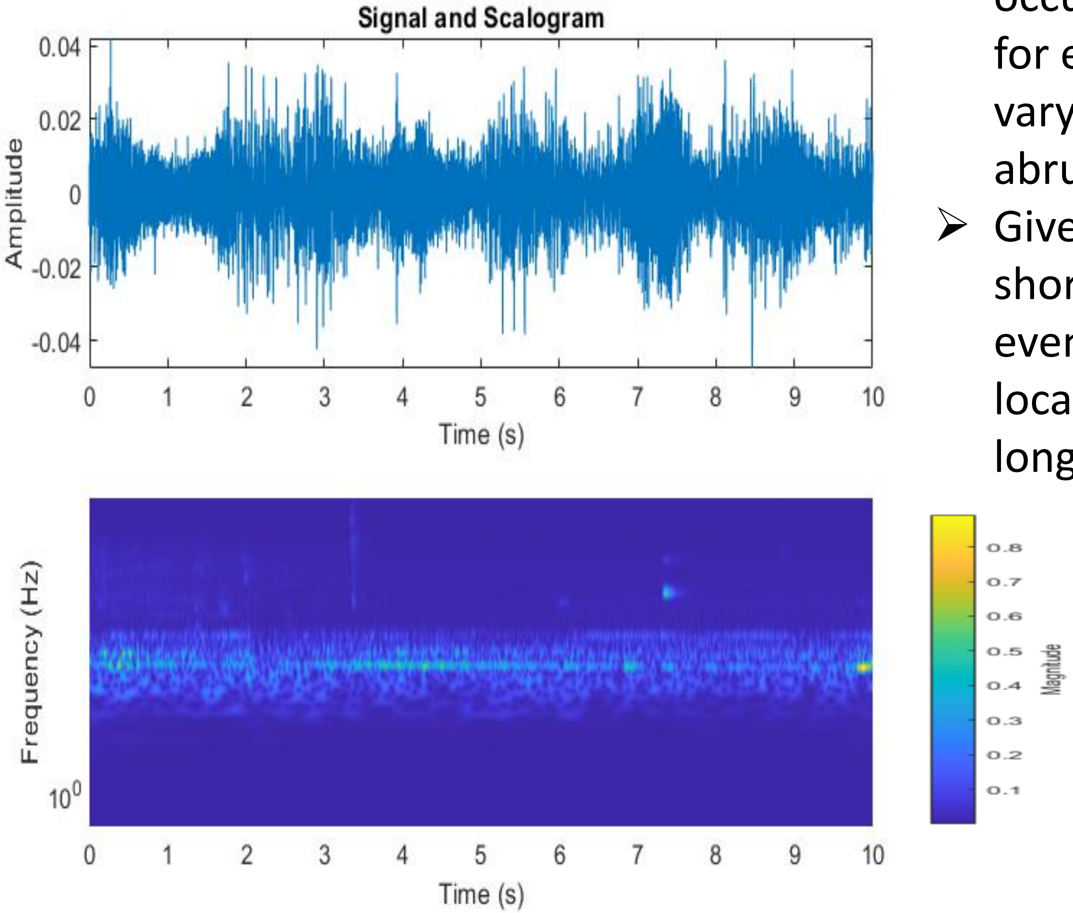
https://doi.org/10.5281/zenodo.400515

Sample Data

Beach

Train

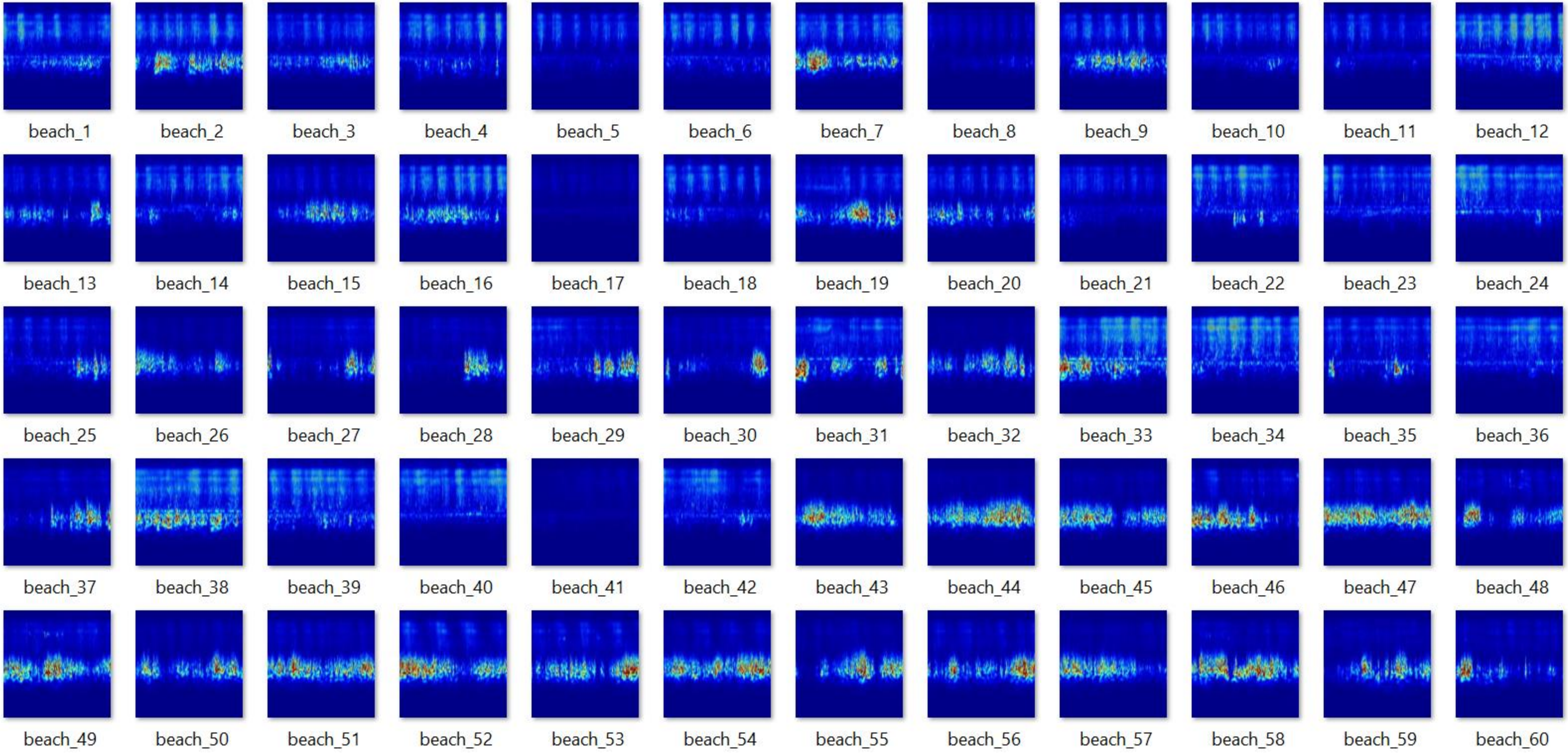

Signal and Scalogram


Signal and Scalogram

- ➢ The scalogram is the absolute value of the continuous wavelet transform (CWT) of a signal, plotted as a function of time and frequency.
- ➢ The scalogram analyzes real-world signals with features occurring at different scales — for example, signals with slowly varying events punctuated by abrupt transients.
- ➢ Give better time localization for short-duration, high-frequency events, and better frequency localization for low-frequency, longer-duration events.

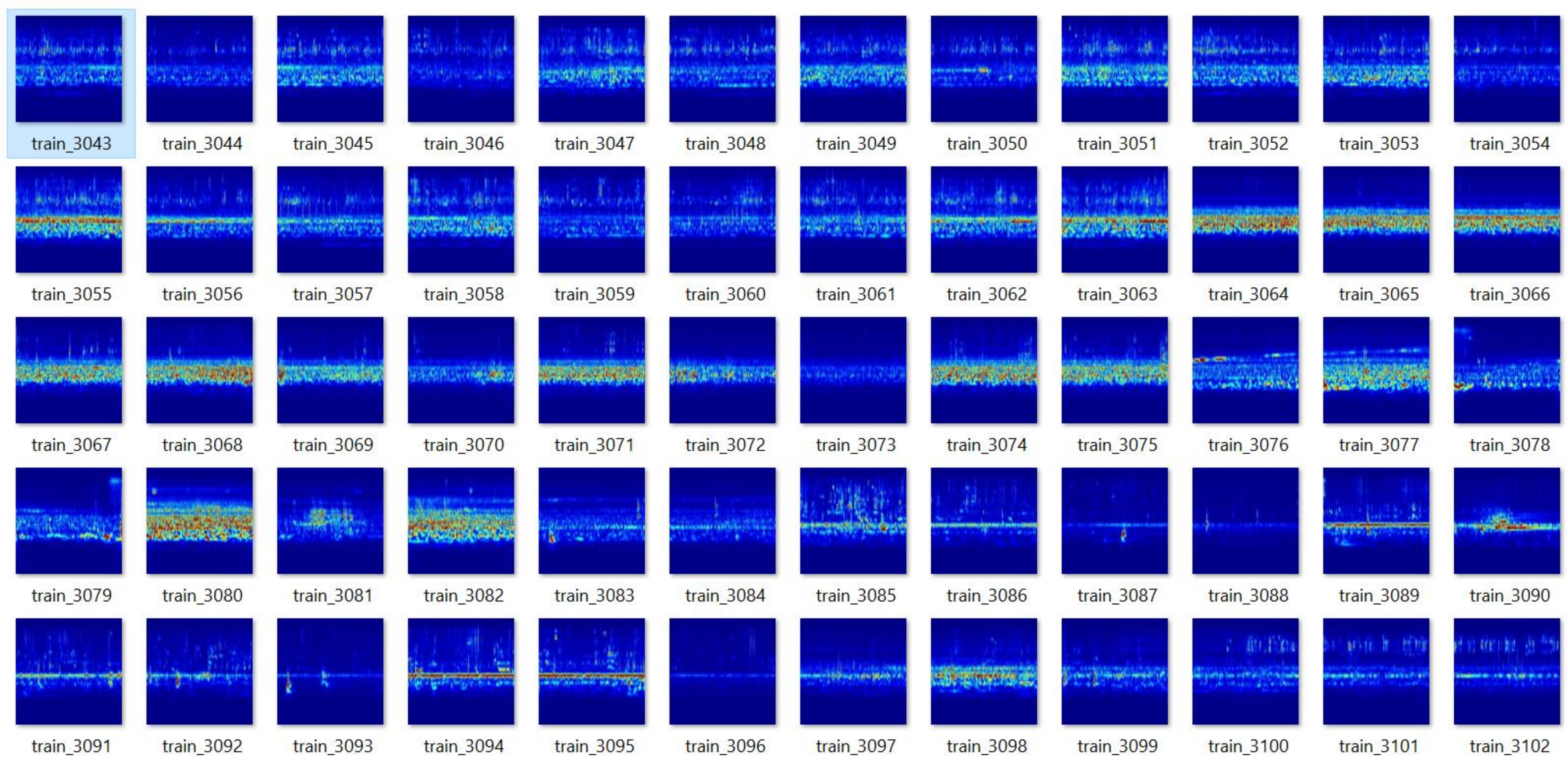
Sample Data

Beach



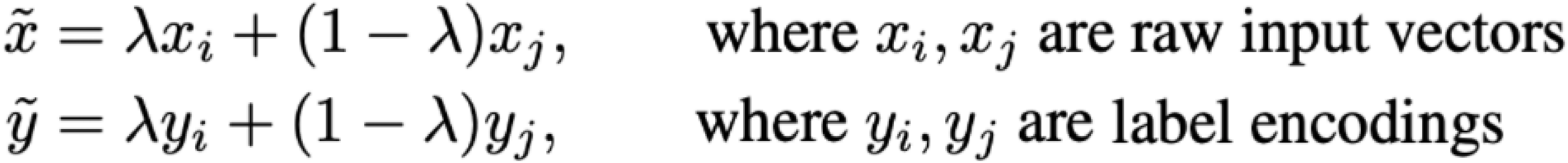
Sample Data

Train



Data Augmentation-Mixup

The technique literally mixing up the features and their corresponding labels.

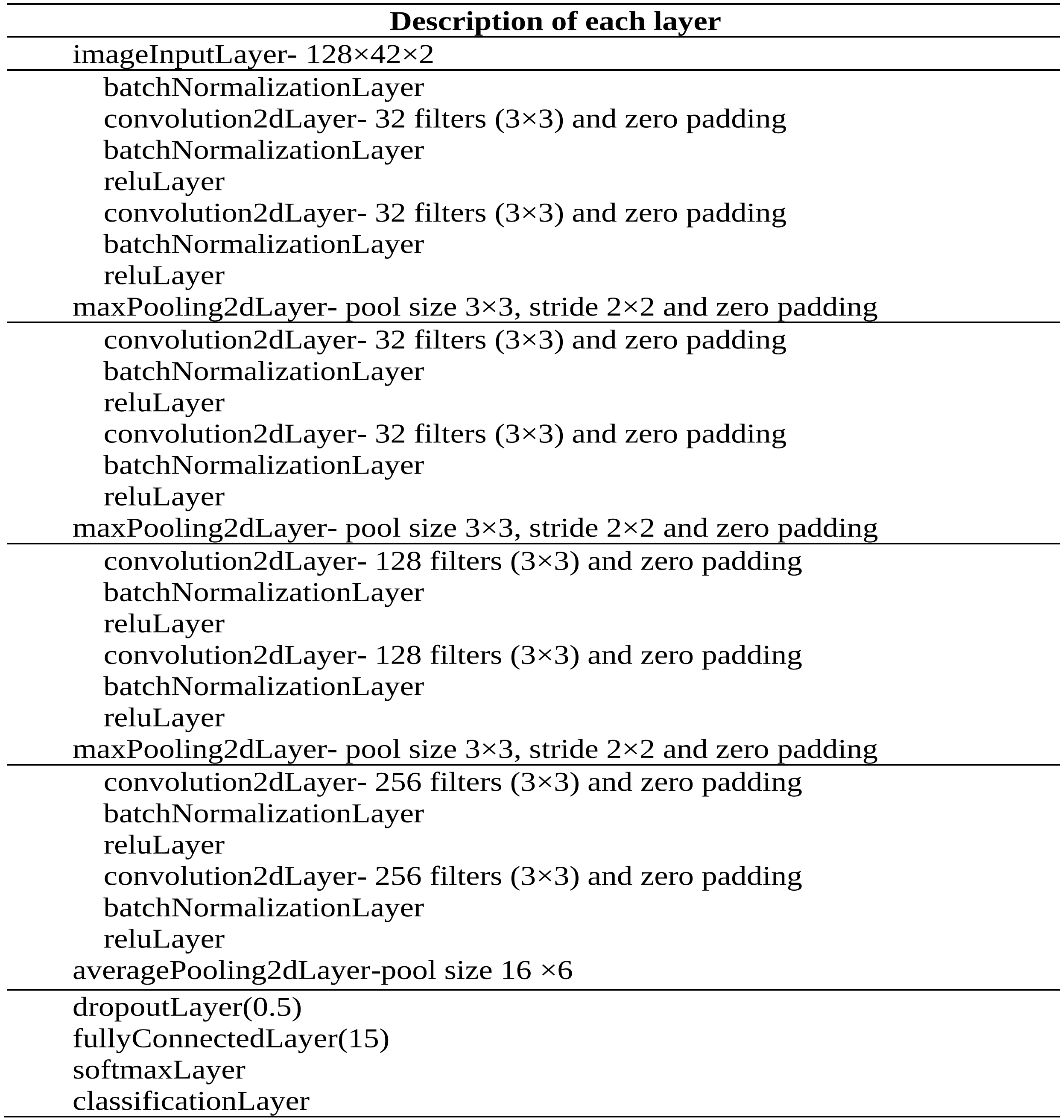


$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j, \qquad \text{where } x_i, x_j \text{ are raw input vectors}$$
$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j, \qquad \text{where } y_i, y_j \text{ are label encodings}$$

Note that the lambda values are values with the [0, 1] range

1) Convolutional Neural Network (CNN)

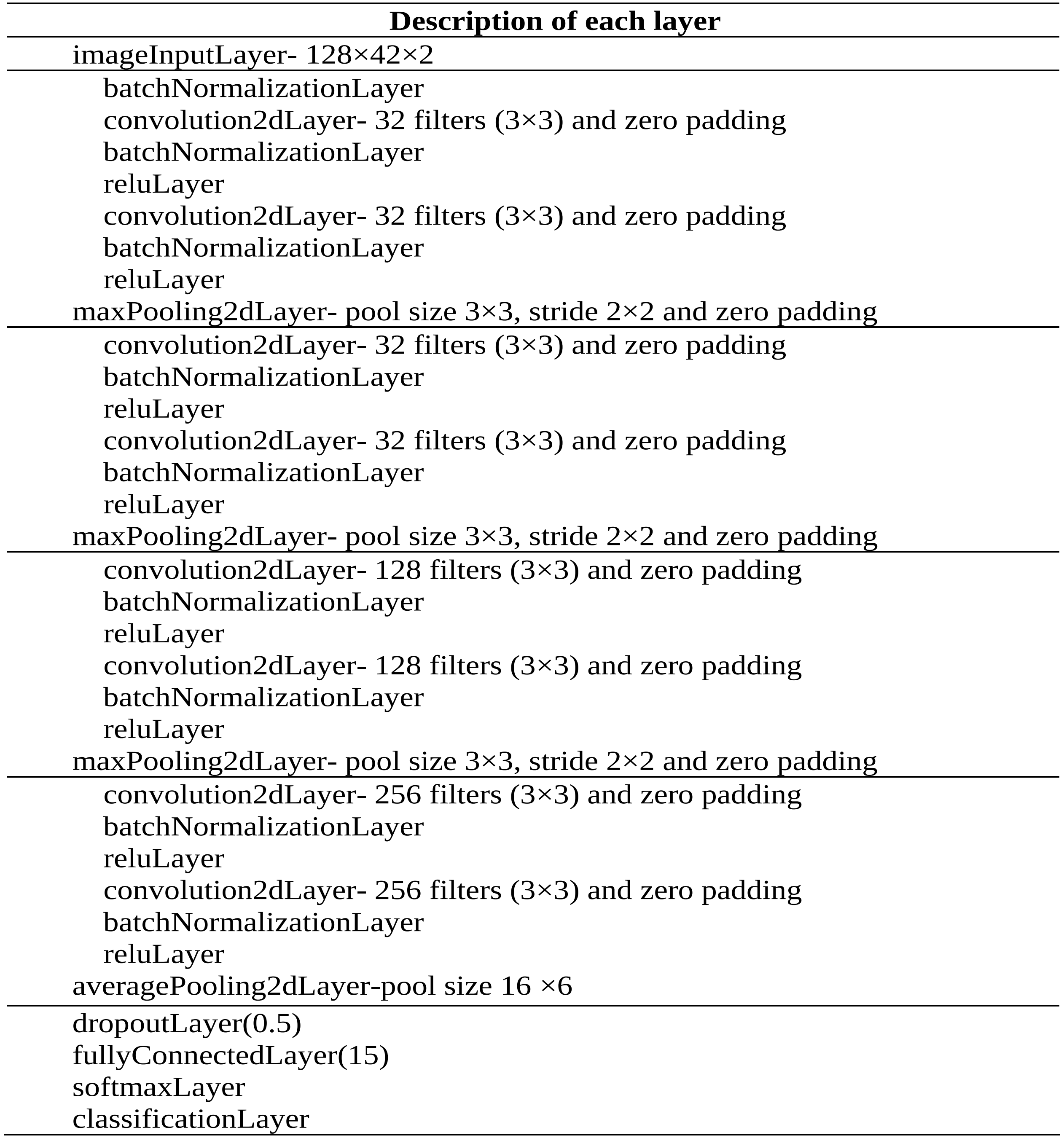
- ➢ The Batch Normalization (BN) and rectified linear unit (ReLU) are used.
- ➢ The ReLU increases the non-linearity in the images.
- ➢ The batch normalization learning is used as a regularization to prevent overfitting.
- ➢ The activation function and BN are located before the convolution layer to improve the acoustic classification accuracy.
- ➢ The max-pooling layers come after the convolution process.



Description of each layer
imageInputLayer- 128×42×2
batchNormalizationLayer
convolution2dLayer- 32 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
convolution2dLayer- 32 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
maxPooling2dLayer- pool size 3×3, stride 2×2 and zero padding
convolution2dLayer- 32 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
convolution2dLayer- 32 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
maxPooling2dLayer- pool size 3×3, stride 2×2 and zero padding
convolution2dLayer- 128 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
convolution2dLayer- 128 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
maxPooling2dLayer- pool size 3×3, stride 2×2 and zero padding
convolution2dLayer- 256 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
convolution2dLayer- 256 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
averagePooling2dLayer-pool size 16 ×6
dropoutLayer(0.5)
fullyConnectedLayer(15)
softmaxLayer
classificationLayer

1)Convolutional Neural Network (CNN)

- ➢ The feature map that includes a prominent feature is obtained from the output of the max-pooling layer.
- ➢ The average pooling reduces the activation by combining the non-maximal activations.
- ➢ The last few layers consist of a dropout layer that removes 50% of the visible and hidden units to reduce overfitting.
- ➢ The fully connected layer is compiled the data to form the output for the last second layer that uses the softmax activation function to obtain probabilities of the input from the 15 classes.
- ➢ Lastly, the last classification layer produces the final classification



Description of each layer
imageInputLayer- 128×42×2
batchNormalizationLayer
convolution2dLayer- 32 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
convolution2dLayer- 32 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
maxPooling2dLayer- pool size 3×3, stride 2×2 and zero padding
convolution2dLayer- 32 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
convolution2dLayer- 32 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
maxPooling2dLayer- pool size 3×3, stride 2×2 and zero padding
convolution2dLayer- 128 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
convolution2dLayer- 128 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
maxPooling2dLayer- pool size 3×3, stride 2×2 and zero padding
convolution2dLayer- 256 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
convolution2dLayer- 256 filters (3×3) and zero padding
batchNormalizationLayer
reluLayer
averagePooling2dLayer-pool size 16 ×6
dropoutLayer(0.5)
fullyConnectedLayer(15)
softmaxLayer
classificationLayer

2) Wavelet Scattering

The next step involves feature extraction using wavelet scattering for subsequent ensemble classifiers.

The parameters of the transform are the filter-bank (using 1D Morlet wavelets) resolutions Q1=1 and Q2=4.

The duration 0.75s of the averaging filter (or invariance scale) is used for the modulation structure duration.

The sampling frequency is 44100 Hz.

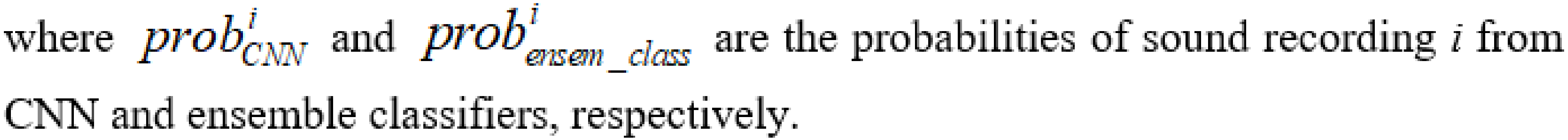
3) Ensemble Classifiers

- ➢The proposed ensemble classifiers include different discriminant analysis learners, such as linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), and Regularized linear discriminant analysis (RDA) with other predictors covariance treatments.
- ➢The random subspace learning method is used to increase the acoustic classification accuracy.
- ➢In the random subspace, the feature subspaces are chosen randomly from the original feature space.
- ➢The final prediction of these individual classifiers is then obtained using majority voting.

4) Fusion of CNN and Classifiers

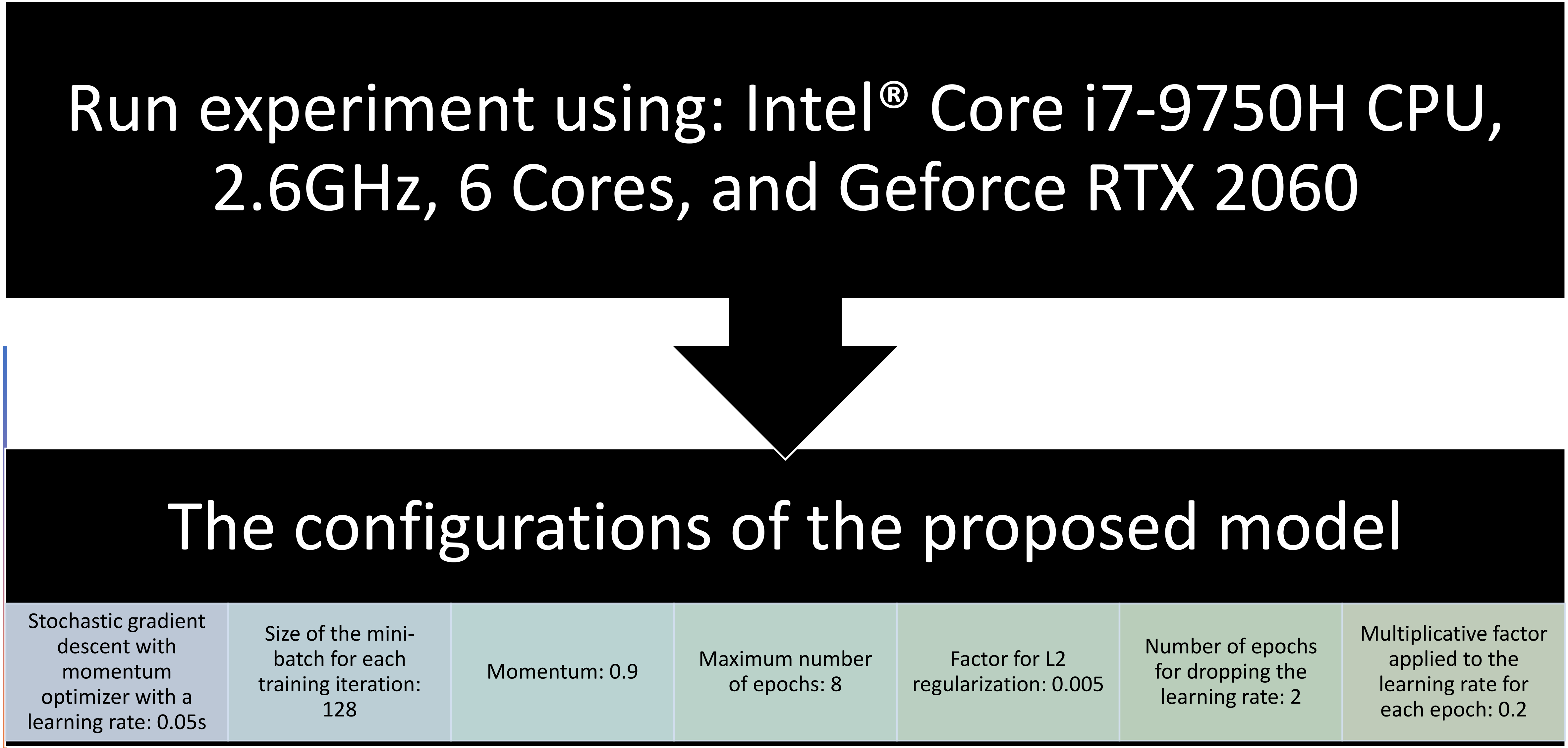
- ➢The fusion of the CNN and classifier prediction results indicates the relative confidence of their prediction.
- ➢Multiplying the responses and determining the maximum response creates a late fusion system that inherent in the merits of each method.

$$class\_pred^{i} = \operatorname{argmax}\left(prob^{i}_{CNN}, prob^{i}_{ensem\_class} \right)$$



where $prob^{i}_{CNN}$ and $prob^{i}_{ensem\_class}$ are the probabilities of sound recording i from CNN and ensemble classifiers, respectively.

Results and Discussion



Run experiment using: Intel® Core i7-9750H CPU, 2.6GHz, 6 Cores, and Geforce RTX 2060

The configurations of the proposed model

Stochastic gradient descent with momentum optimizer with a learning rate: 0.05s

Size of the mini-batch for each training iteration: 128

Momentum: 0.9

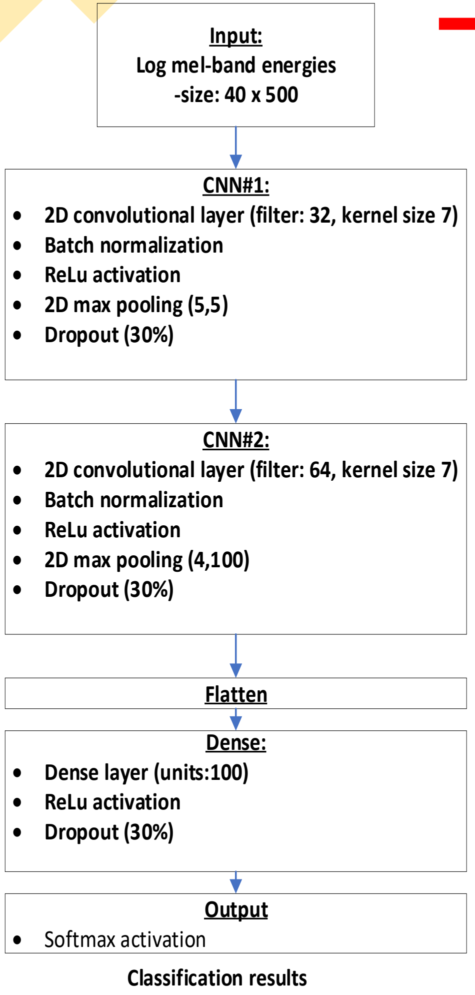
Maximum number of epochs: 8

Factor for L2 regularization: 0.005

Number of epochs for dropping the learning rate: 2

Multiplicative factor applied to the learning rate for each epoch: 0.2

Results and Discussion
-Comparison with Baseline Model



Input:
Log mel-band energies
-size: 40 x 500

CNN#1:
- 2D convolutional layer (filter: 32, kernel size 7)
- Batch normalization
- ReLu activation
- 2D max pooling (5,5)
- Dropout (30%)

CNN#2:
- 2D convolutional layer (filter: 64, kernel size 7)
- Batch normalization
- ReLu activation
- 2D max pooling (4,100)
- Dropout (30%)

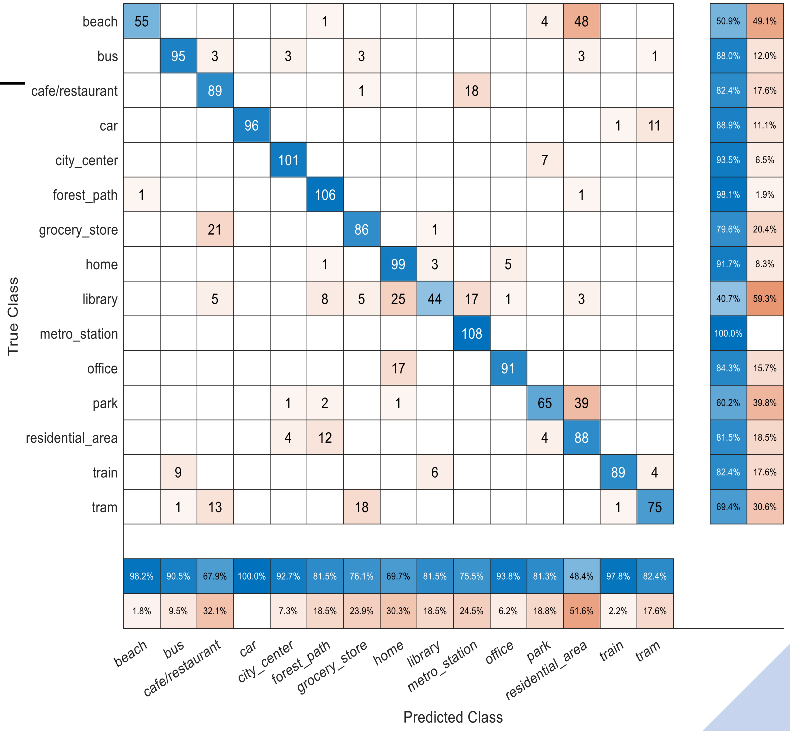
Flatten

Dense:
- Dense layer (units:100)
- ReLu activation
- Dropout (30%)

Output
- Softmax activation

Classification results

Scenes	Acoustic Classification Accuracy (%)			
	Baseline Model [32]	**CNN Model**	**Ensemble Classifiers Model**	**Fused Model**
beach	40.7	73.1	37.9	50.9
bus	38.9	58.3	92.5	87.9
cafe/restaurant	43.5	74.0	82.4	82.4
car	64.8	100	76.8	88.8
city-center	79.6	88.8	91.6	93.5
forest path	85.2	97.2	96.2	98.1
grocery store	49.1	70.3	79.6	79.6
home	76.9	89.8	76.8	91.6
library	30.6	49.0	36.1	40.7
metro station	93.5	100	95.3	100
office	73.1	80.5	83.3	84.2
park	32.4	20.3	68.5	60.1
residential area	77.8	63.8	77.7	81.4
train	72.2	76.8	85.1	82.4
tram	57.4	57.4	64.8	69.4
Average	61.0	73.3	76.3	79.4

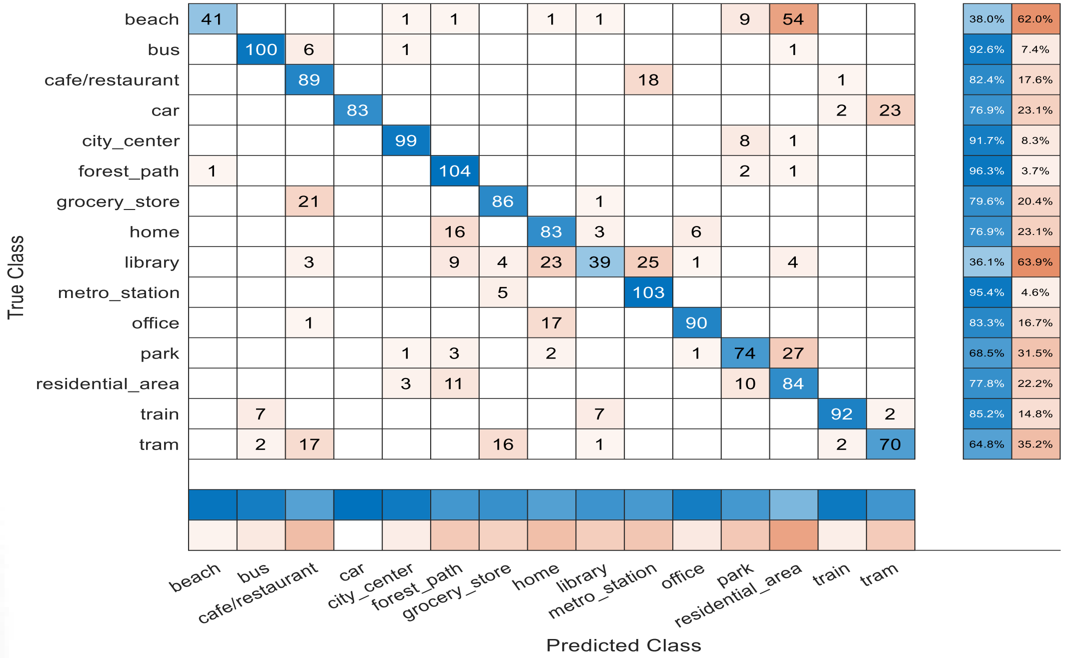


True Class	beach	bus	cafe/restaurant	car	city_center	forest_path	grocery_store	home	library	metro_station	office	park	residential_area	train	tram		
beach	55					1						4	48			50.9%	49.1%
bus		95	3		3		3						3		1	88.0%	12.0%
cafe/restaurant			89				1			18						82.4%	17.6%
car				96										1	11	88.9%	11.1%
city_center					101							7				93.5%	6.5%
forest_path	1					106							1			98.1%	1.9%
grocery_store			21				86		1							79.6%	20.4%
home						1		99	3		5					91.7%	8.3%
library			5			8	5	25	44	17	1		3			40.7%	59.3%
metro_station										108						100.0%	
office								17			91					84.3%	15.7%
park					1	2		1				65	39			60.2%	39.8%
residential_area					4	12						4	88			81.5%	18.5%
train		9							6					89	4	82.4%	17.6%
tram		1	13				18							1	75	69.4%	30.6%
	98.2%	90.5%	67.9%	100.0%	92.7%	81.5%	76.1%	69.7%	81.5%	75.5%	93.8%	81.3%	48.4%	97.8%	82.4%		
	1.8%	9.5%	32.1%		7.3%	18.5%	23.9%	30.3%	18.5%	24.5%	6.2%	18.8%	51.6%	2.2%	17.6%		

- ➢ The ensemble classifiers have a higher acoustic classification accuracy than CNN.
- ➢ Compared to the baseline model (consists of 2 layers × 50 hidden units, 20% dropout), the fused model exhibits 18.4% higher accuracy.

Results and Discussion
-Comparison with CNN only

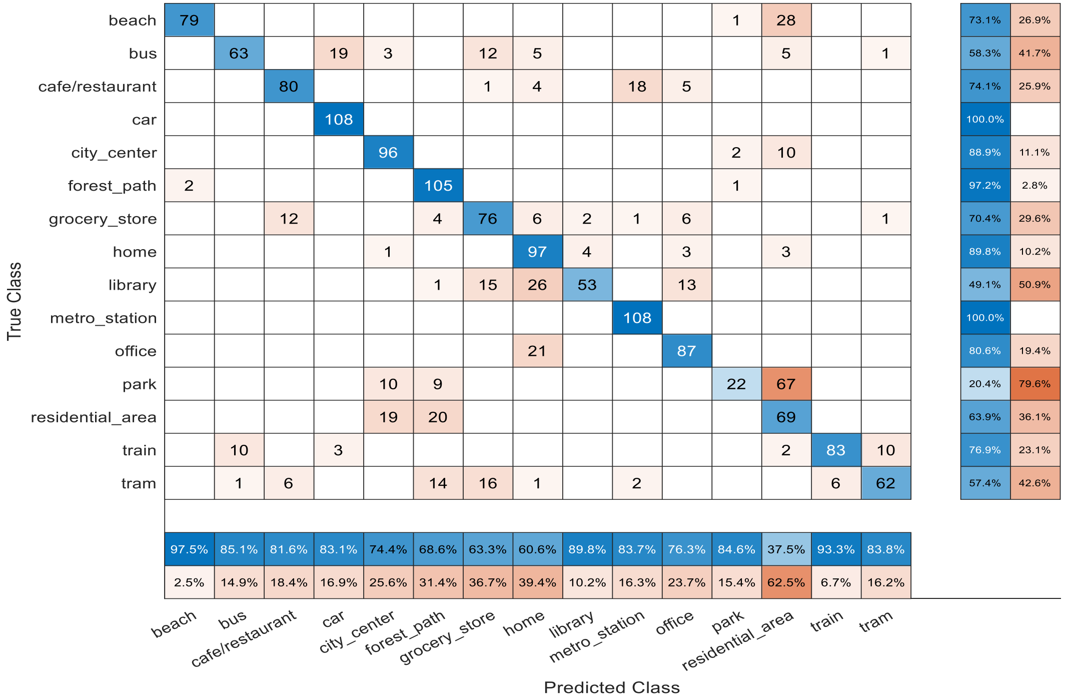
- Although the result of the scene (i.e., beach) using ensemble classifiers (37.96%) is quite poor as compared to CNN (73.14%), the fused model managed to increase the acoustic classification accuracy to 50.92%.
- Conversely, the scene (i.e. park) using CNN model is relatively low compared to the ensemble classifiers. The fused model increases it to 60.18%.
- The confusion matrix of CNN, ensemble classifiers, and the fused model are shown.
- The average acoustic classification accuracy of the fused model is computed as 79.43%.



	beach	bus	cafe/restaurant	car	city_center	forest_path	grocery_store	home	library	metro_station	office	park	residential_area	train	tram		
beach	41				1	1		1	1			9	54			38.0%	62.0%
bus		100	6		1								1			92.6%	7.4%
cafe/restaurant			89							18				1		82.4%	17.6%
car				83										2	23	76.9%	23.1%
city_center					99							8	1			91.7%	8.3%
forest_path	1					104						2	1			96.3%	3.7%
grocery_store			21				86		1							79.6%	20.4%
home						16		83	3		6					76.9%	23.1%
library			3			9	4	23	39	25	1		4			36.1%	63.9%
metro_station							5			103						95.4%	4.6%
office			1					17			90					83.3%	16.7%
park					1	3		2			1	74	27			68.5%	31.5%
residential_area					3	11						10	84			77.8%	22.2%
train		7							7					92	2	85.2%	14.8%
tram		2	17				16		1					2	70	64.8%	35.2%

True Class

Predicted Class



	beach	bus	cafe/restaurant	car	city_center	forest_path	grocery_store	home	library	metro_station	office	park	residential_area	train	tram		
beach	79											1	28			73.1%	26.9%
bus		63		19	3		12	5					5		1	58.3%	41.7%
cafe/restaurant			80				1	4		18	5					74.1%	25.9%
car				108												100.0%	
city_center					96							2	10			88.9%	11.1%
forest_path	2					105						1				97.2%	2.8%
grocery_store			12			4	76	6	2	1	6				1	70.4%	29.6%
home					1			97	4		3		3			89.8%	10.2%
library						1	15	26	53		13					49.1%	50.9%
metro_station										108						100.0%	
office								21			87					80.6%	19.4%
park					10	9						22	67			20.4%	79.6%
residential_area					19	20							69			63.9%	36.1%
train		10		3									2	83	10	76.9%	23.1%
tram		1	6			14	16	1		2				6	62	57.4%	42.6%
	97.5%	85.1%	81.6%	83.1%	74.4%	68.6%	63.3%	60.6%	89.8%	83.7%	76.3%	84.6%	37.5%	93.3%	83.8%		
	2.5%	14.9%	18.4%	16.9%	25.6%	31.4%	36.7%	39.4%	10.2%	16.3%	23.7%	15.4%	62.5%	6.7%	16.2%		

True Class

Predicted Class

Conclusion

The multi-model late fusion system model consisting of the log-Mel spectrogram for convolutional neural network and wavelet time scattering for ensemble of subspace discriminant classifiers was proposed.

Based on the dataset from the TUT Acoustic Scenes, it demonstrated that the fused model gives good acoustic classification accuracy of 79.43%.

The proposed multi-model late fusion system exhibits 18.4% higher acoustic classification accuracy than the baseline model despite relatively low performance in a few scenes such as the beach and library.

Nevertheless, the multi-model late fusion system shows good acoustic classification accuracy for most of the scenes.

Related Publications (2019-Present)

1) T.K. Chan, C.S. Chin, Multi Branch Convolutional Macaron Net for Sound Event Detection, IEEE Transactions on Audio, Speech and Language Processing, vol. 29, pp. 2972-2985, 2021.
2) T.K. Chan, C.S. Chin, Y. Li, Semi-Supervised NMF-CNN For Sound Event Detection, IEEE Access, vol. 9, pp. 130529-130542, 2021.
3) T.K. Chan, C.S. Chin, Lightweight Convolutional-iConformer For Sound Event Detection, IEEE Transactions on Audio, Speech and Language Processing, 2021, submitted.
4) T. K. Chan and C. S. Chin, A Comprehensive Review of Polyphonic Sound Event Detection, IEEE Access, 8, 103339-103373, 2020.
5) TK Chan, CS Chin, Detecting Sound Events Using Convolutional Macaron Net With Pseudo Strong Labels, IEEE 23rd International Workshop on Multimedia Signal Processing, Tampere, Finland, 6-8 October 2021.
6) CS. Chin, JF. Xiao, Max-Fusion of Random Ensemble Subspace Discriminant with Aggregation of MFCCs and High Scalogram Coefficients for Acoustics Classification, 20th IEEE/ACIS International Summer Semi-Virtual Conference on Computer and Information Science, (ICIS 2021-Summer), Shanghai, China, 23-25 Jun 2021.
7) CS Chin, JH Zhang, Wavelet Scattering Transform for Multiclass Support Vector Machines in Audio Devices Classification System, 20th IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Aula Conference Centre TU Delft, Delft, The Netherlands, 12-16 Jul 2021.

Related Publications (2019-Present)

8) XY Kek, CS Chin, Y. Li, An Investigation on Multiscale Normalised Deep Scattering Spectrum with Deep Residual Network for Acoustic Scene Classification, 22nd IEEE/ACIS International Fall Virtual Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing,Taichung, Taiwan, 24-26 November 2021.

9) XY Kek, CS Chin, Y.Li, Deep Scattering Spectrum with Mobile Network for Low Complexity Acoustic Scene Classification, Detection and Classification of Acoustic Scenes and Events 2021(DCASE2021), Technical Report, 2021.

10) CS Chin, JH Zhang, Late Fusion of Convolutional Neural Network with Wavelet-based Ensemble Classifier for Acoustic Scene Classification, Intelligent Systems Conference (IntelliSys) 2021, Springer, Amsterdam, The Netherlands, 2-3 Sept 2021.

11) CS Chin, XY Kek, TK Chan, Wavelet Scattering Based Gated Recurrent Units for Binaural Acoustic Scenes Classification, IEEE International Conference on Internet of Things and Intelligent Applications, Zhenjiang, China, 27-29 Nov, 2020.

12) CS Chin, XY Kek, TK Chan, Scattering Transform of Averaged Data Augmentation for Ensemble Random Subspace Discriminant Classifiers in Audio Recognition, 7th International Conference on Advanced Computing and Communication Systems, Coimbatore, India, 19 – 20 March, 2021.

13) Chan TK, Chin CS, Li Y. Non-Negative Matrix Factorization-Convolution Neural Network (NMF-CNN) for Sound Event Detection. IEEE Workshop on Detection and Classification of Acoustic Scenes and Events 2019 (DCASE 2019), New York, USA, 2019.

14) XY. Kek, C. S. Chin, Y Li, Acoustic Scene Classification using Bilinear Pooling on Time-Liked and Frequency-Liked Convolution Neural Network, IEEE Symposium Series on Computational Intelligence, Xiamen, China, 6-9 Dec, 2019.

Thank you!