A Two-Stage Approach to Device-Robust Acoustic Scene Classification

Hu Hu, Chao-Han Huck Yang, Xianjun Xia, Xue Bai, Xin Tang, Yajian Wang,
Shutong Niu, Li Chai, Juanjuan Li, Hongning Zhu, Feng Bao, Yuanjun Zhao,
Sabato Marco Siniscalchi, Yannan Wang, Jun Du, Chin-Hui Lee

ICASSP 2021

Overview

- Introduction

- Two-stage ASC system
  - Two-stage classification procedure
  - CNN classifiers
  - Data augmentation strategies

- Experiments
  - Experimental setup
  - Overall results of proposed systems
  - Evaluation of data augmentation strategies

- Neural Saliency Analysis

- Conclusions

Introduction

- Acoustic Scene Classification (ASC)
  - To identify real-life sounds into environment classes, such as metro station, street traffic, or public square.
  - State-of-the-art performance always comes from CNN based end-to-end system.
    - Input features → CNN classifier → output scene class

- Device robustness issue
  - In real world application, ASC system should be robust to different recording devices.
  - In DCASE 2020[1], task 1a focus on the device robustness issue of ASC.

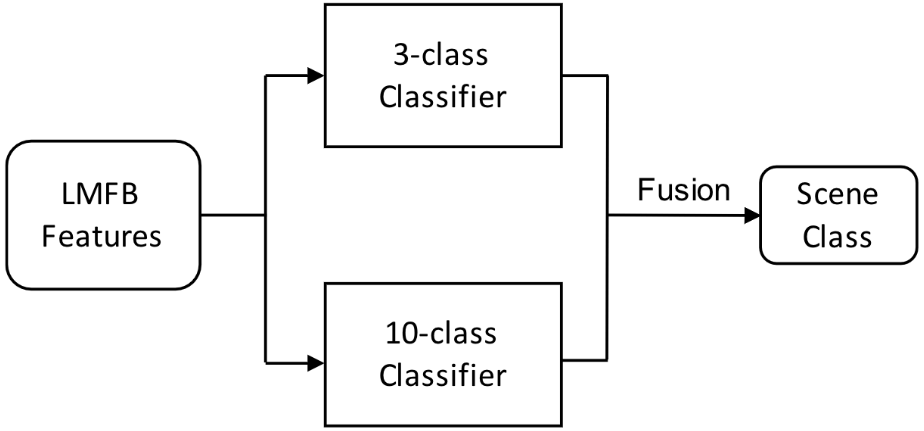
- In this work, we propose a novel **Two-Stage ASC System** to leverage the device robustness issue of ASC.

[1] T. Heittola, A. Mesaros, and T. Virtanen, “Acoustic scene clas-sification in dcase 2020 challenge: generalization across de-vices and low complexity solutions,” in DCASE2020, 2020.

Two-Stage Classification Procedure

- 3-class classifier:
  - Classifies an input scene audio into one of three broad classes: in-door, out-door, and transportation.
  - We expect it to enhance the classification process and leverage the overfitting issue of 10-class classifier.
- 10-class classifier:
  - The main classifier.
  - Assigns a given input audio clip into one of ten target acoustic scene classes
- Score fusion:

$$Class(x) = \underset{q,(p\in\mathbb{C}^1, q\in\mathbb{C}^2, p\supset q)}{\operatorname{argmax}} F_p^1(x) * F_q^2(x)$$



The proposed two-stage ASC system

ASC System Design

- 3 different CNN models are investigated in our work
    - Resnet: A duel path resnet model, where each input feature map is divided into two sub-feature mapping along the frequency dimension.
    - FCNN: A fully convolutional neural networks built with 9 stacked convolutional layers with small-size kernel.
    - fsFCNN: An extension of FCNN model, which mainly has 2 more convolutional layers and reduces max-pooling size in the frequency axis.

- 9 different data augmentation strategies are investigated in our work
    - No generating extra data:
        - Mixup, Random cropping, SpecAugment.
    - Generating extra data:
        - Spectrum correction: Generate new features by a correction coefficient.
        - Reverberation with DRC: Add reverberation by RIR, and apply dynamic range compression.
        - Pitch shift: Randomly shift the pitch of each audio clip based on the uniform distribution.
        - Speed change: Randomly change the audio speed based on the uniform distribution.
        - Random noise: Add random Gaussian noise.
        - Mix audios: Randomly mix two audios from the same scene class.

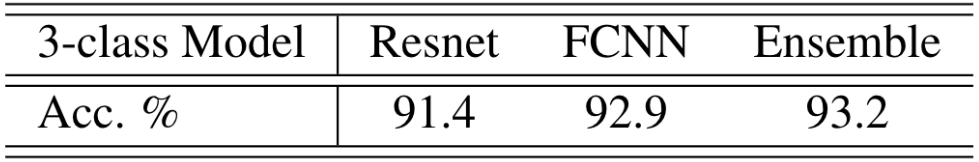
Experimental Setup

- Data set: DCASE 2020 Task1a development data set.
  - ~14K 10-second acoustic scene training audios recorded by 6 different devices.
    - Device A accounts for ~75%, B, C, s1-s3 account for 5%, respectively.
  - The goal is to get good performance on 9 different test devices:
    - Real source devices: device A.
    - Real target devices: device B & C.
    - Simulated seen devices: device s1-s3.
    - Simulated unseen devices: device s4-s6.

- Data processing:
  - 128-D log-mel filter bank (LMFB) features are used.
  - We perform utterance-level scaling operation to scale LMFB features into [0,1]

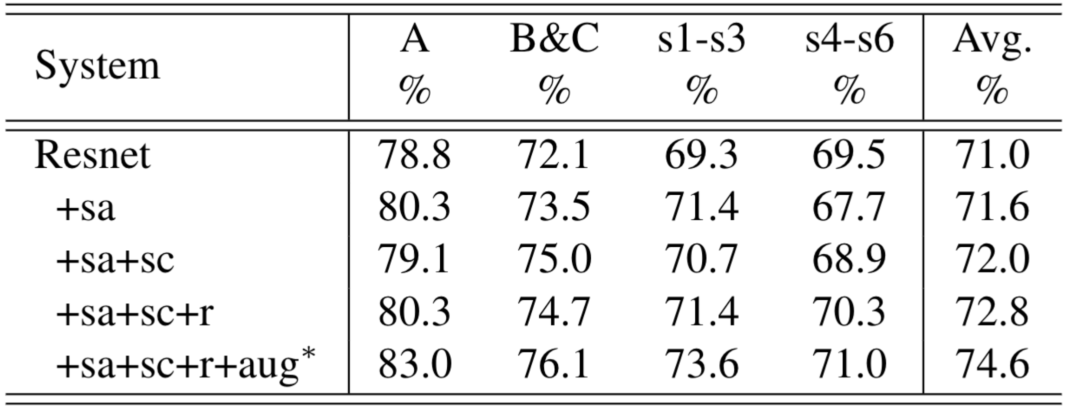
Experimental Results

- 3-class classification results



3-class Model	Resnet	FCNN	Ensemble
Acc. %	91.4	92.9	93.2

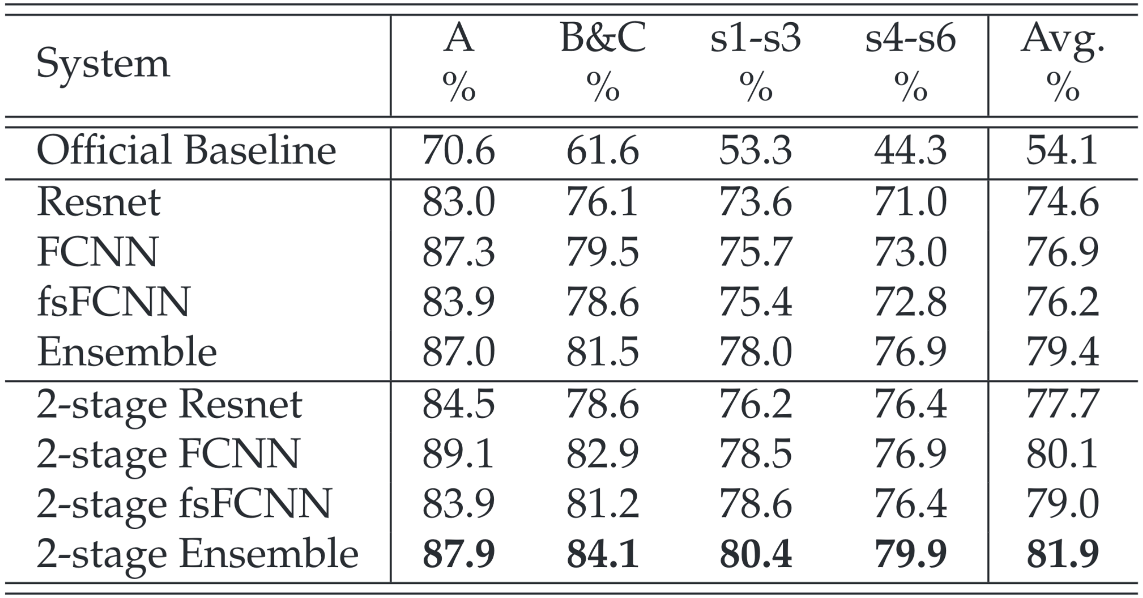
- Evaluation of different data augmentation strategies (mix-up and random cropping are always used)
    - ‘sa’ indicates specAugment.
    - ‘sc’ indicates spectrum correction.
    - ‘r’ indicates reverberation with DRC.
    - ‘aug’ indicates another four augmentation methods, including pitch shift, speed change, random noise and mix audios.



System	A	B&C	s1-s3	s4-s6	Avg.
	%	%	%	%	%
Resnet	78.8	72.1	69.3	69.5	71.0
+sa	80.3	73.5	71.4	67.7	71.6
+sa+sc	79.1	75.0	70.7	68.9	72.0
+sa+sc+r	80.3	74.7	71.4	70.3	72.8
+sa+sc+r+aug*	83.0	76.1	73.6	71.0	74.6

Experimental Results (cont’d)

- Overall results of our proposed system

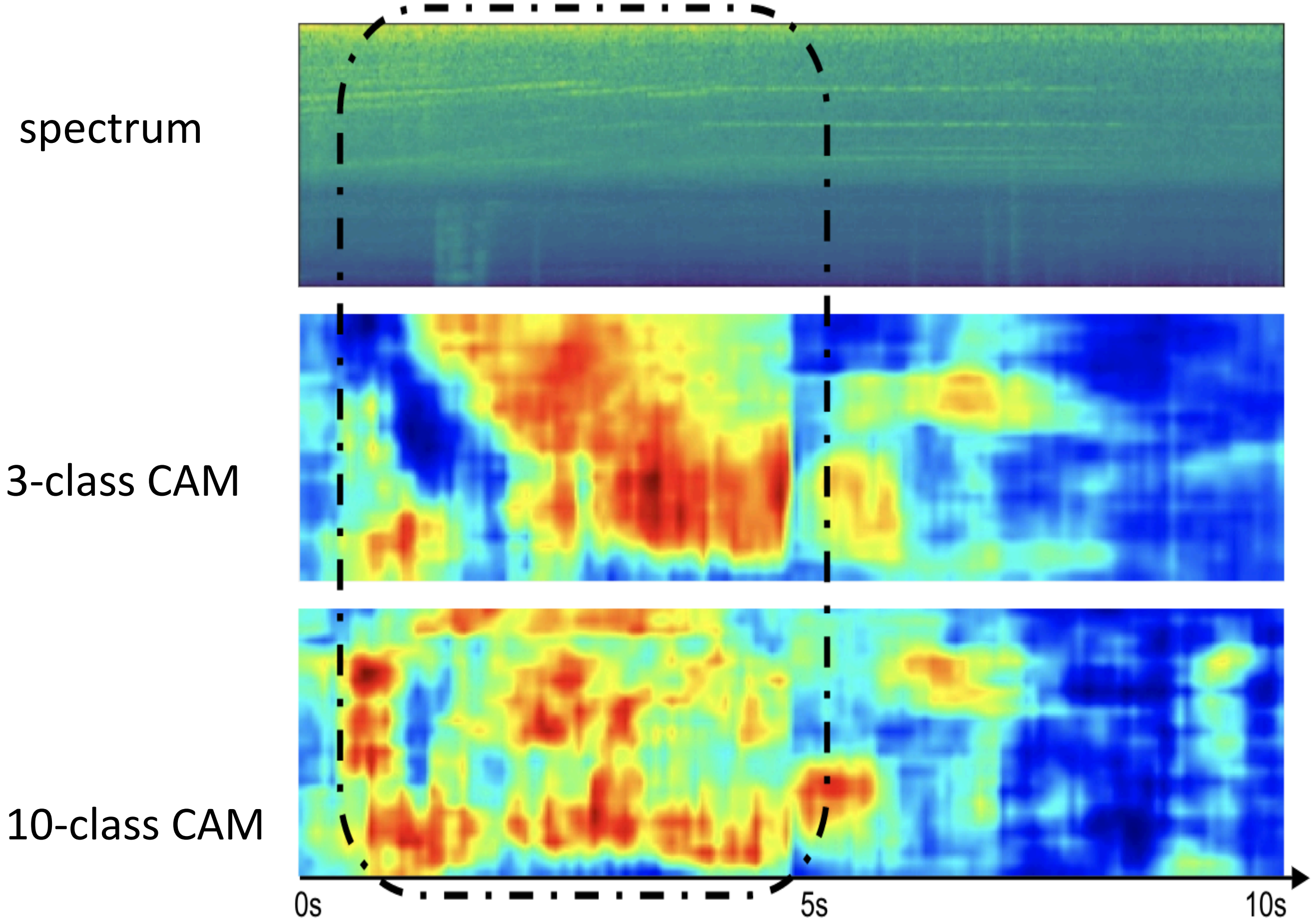


System	A %	B&C %	s1-s3 %	s4-s6 %	Avg. %
Official Baseline	70.6	61.6	53.3	44.3	54.1
Resnet	83.0	76.1	73.6	71.0	74.6
FCNN	87.3	79.5	75.7	73.0	76.9
fsFCNN	83.9	78.6	75.4	72.8	76.2
Ensemble	87.0	81.5	78.0	76.9	79.4
2-stage Resnet	84.5	78.6	76.2	76.4	77.7
2-stage FCNN	89.1	82.9	78.5	76.9	80.1
2-stage fsFCNN	83.9	81.2	78.6	76.4	79.0
2-stage Ensemble	**87.9**	**84.1**	**80.4**	**79.9**	**81.9**

- CNN classifiers show good robustness on different test devices. Ensemble can further improve the performance.
- Two-stage systems can boost the performance a lot. 2.5% to 3.2% absolute accuracy gains are obtained.
- The best system, 2-stage ensemble can leverage the gap between simulated seen data and simulated unseen data a lot.
- The results on different devices confirm the effectiveness of the proposed two-stage system with respect to improving device robustness.

Neural Saliency Analysis

- CAM[1] analysis of ASC systems: example 1



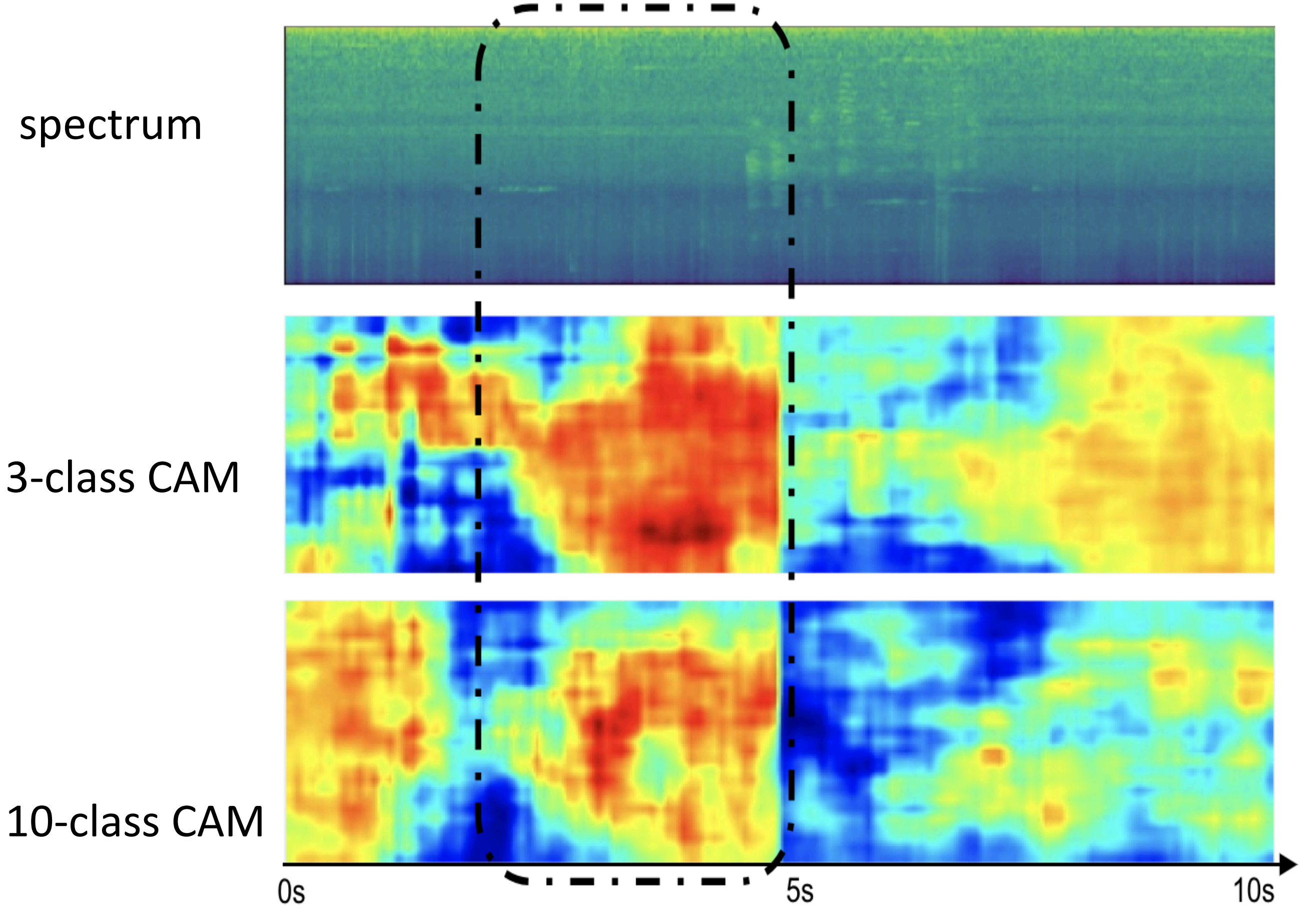
- Utterance: metro_station-vienna-87-2389-a
- Class: metro station
- **Brake and horn sound starts from 0s to around 8s. After 5s, only reverberation remains.**

- Class Probability:
  - 3-class: 0.626, indoor
  - 10-class: 0.707, metro_station
  - 2-stage: 0.850 metro_station

[1] B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Tor-ralba, “Learning Deep Features for Discriminative Localiza-tion.,”CVPR, 2016.

Neural Saliency Analysis (cont'd)

- CAM analysis of ASC systems: example 2



- Utterance:bus-prague-1102-42431-a
- Class: bus
- **brake sound starts from around 2s and stops at around 5s, human talks starts from around 5s**

- Class Probability:
  - 3-class: 0.783, transportation
  - 10-class: 0.794, bus
  - 2-stage: 0.919 bus

Conclusions

- We propose a novel two-stage ASC framework based on CNNs to leverage the device robustness issue. A general 3-class classifier and a specific 10-class classifier are combined through score fusion.

- Three different CNN models and 9 different data augmentation strategies are investigated to improve device robustness.

- Experiments on the DCASE 2020 task1a development set show the effectiveness of our solution. Specifically, our best system, a two-stage fusion of a CNN-based ensemble, obtains a state-of-the-art 81.9% average ASC accuracy.

- We perform CAM-based neural saliency analysis to demonstrate that CNNs pay particular attention to audio segments strictly related acoustic events.

Thank you~