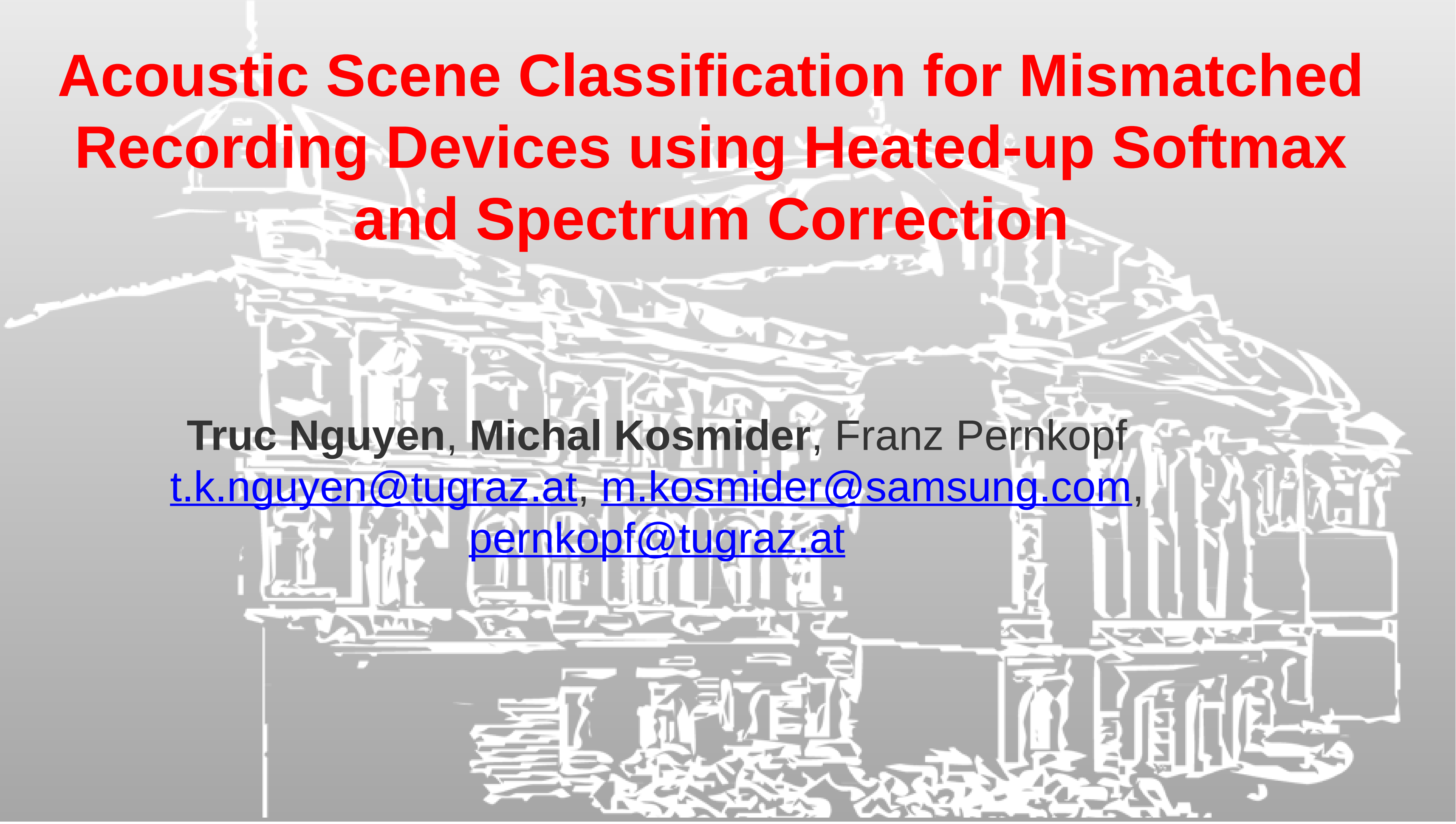


Acoustic Scene Classification for Mismatched Recording Devices using Heated-up Softmax and Spectrum Correction

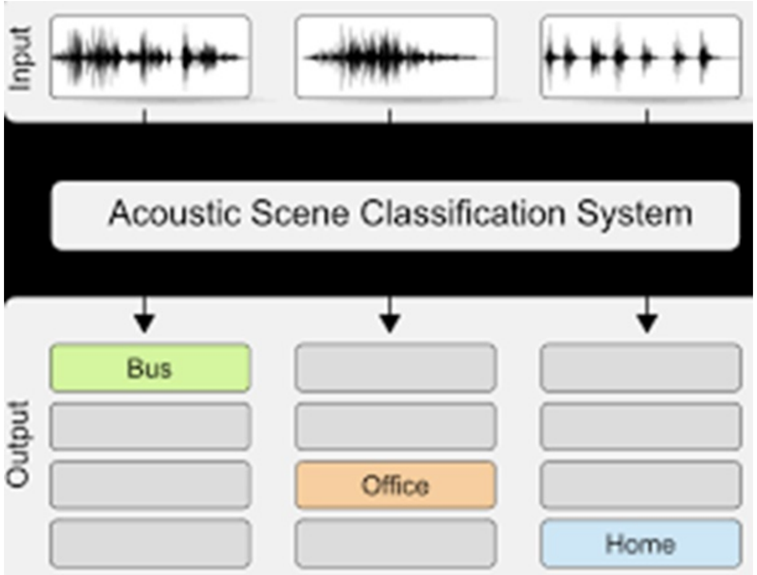
Truc Nguyen, **Michal Kosmider**, Franz Pernkopf
t.k.nguyen@tugraz.at, m.kosmider@samsung.com,
pernkopf@tugraz.at

Outline

- **A**coustic **S**cene **C**lassification (**ASC**) Introduction
- ASC Applications
- Data
- ASC System
- Experimental Results
- Conclusion

ASC Introduction

- **Acoustic Scene Classification (ASC):**
  - A multi-class classification
  - Classifying the **recoded environment sounds** as specific **acoustic scenes**.



http://dcase.community/challenge2016/task-acoustic-scene-classification

ASC Applications

- **Context-aware wearable devices:**
  - Hearing aids
  - Headphones
- **Smartphone**
- **Smart wear**
- **Smart home applications**

Data

- **DCASE 2019 Database:**
  - **Acoustic scenes for tasks (10 classes):**
    - **Outdoor:** Airport, Street pedestrian, Public square, Street traffic, Park
    - **Indoor:** Shopping mall, Metro station
    - **Vehicle:** Tram, Bus, Metro
  - **Recording locations:** 12 cities
  - **Recording devices:**

Binaural microphone
(**Device A**)

Samsung Galaxy S7
(Device B)

IPhone SE
(Device C)

GoPro Hero5 Session
(Device D)

http://dcase.community/challenge2019/task-acoustic-scene-classification

Data

- **DCASE 2019 Database:**
  - Development set:
    - **Device A**: 40 hours (14400 segments, resampled and single-channel)
    - **Device B**: 3 hours (1080 segments)
    - **Device C**: 3 hours (1080 segments)
  - **Training set:** 10265 segments (10s) (**540** segments for each Device B and C)
  - **Test set:** 5265 segment (10s) + **1030** segments recorded in a different city
  - Evaluation set: ~30 hours (Devices A, B, C + **Device D**)

DCASE2019 TASK 1B MISMATCH

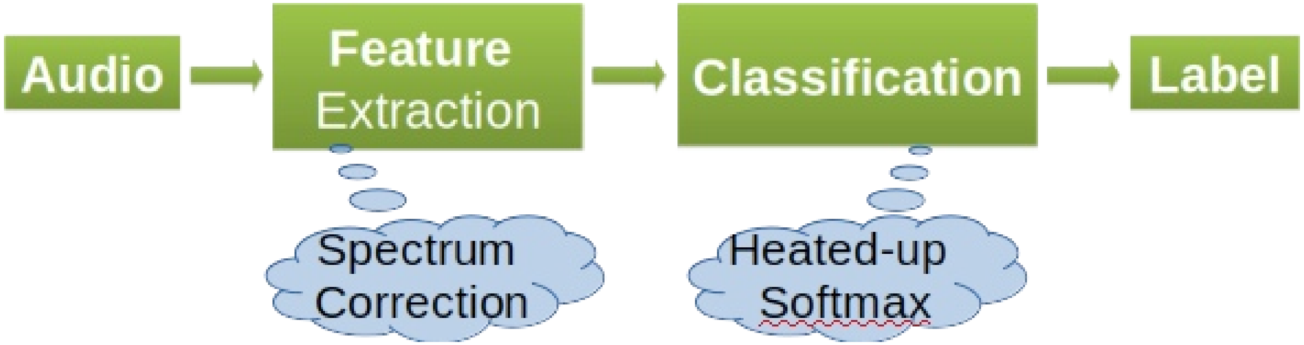
Acoustic Scene Classification with mismatched recording devices

http://dcase.community/challenge2019/task-acoustic-scene-classification

ASC System

- **Challenges:**
  - Differences in frequency responses of recording devices
  - A shift in data distribution between training set and test set

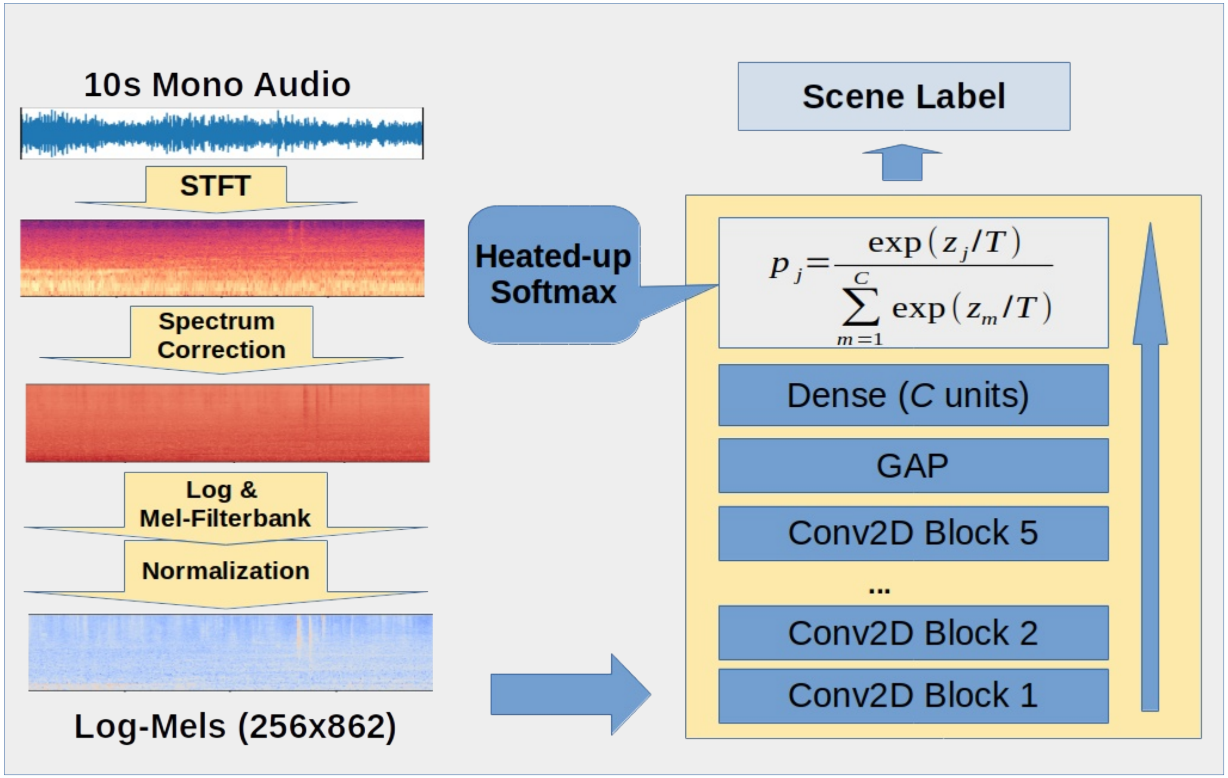
Framework



ASC System

▶ Proposed System

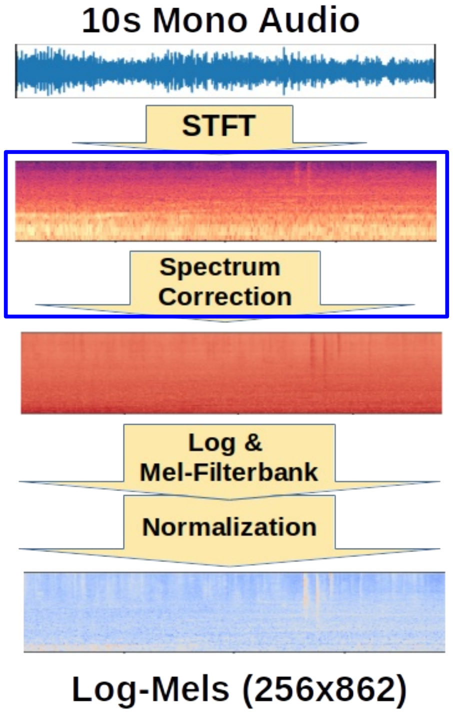
- **Pre-processing**
  - Sampling rate 44.1kHz
  - Hann window
  - Window size: 2048
  - Hop size: 512
  - STFT with **862**
  - **temporal frames**



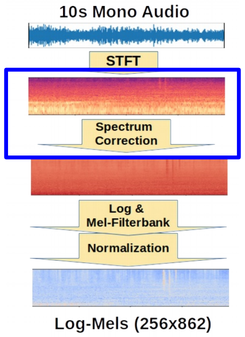
ASC System

Spectrum Correction

- Deal with **varying frequency response**
- of the **recording devices A, B and C**
- **Scaling Coefficients**
- Training set:
  - **Reference recordings: 540** segments
  - of the **same acoustic scenes** for each
  - device **A, B and C**
  - ***Spectrum***: average of a spectrogram
  - over temporal frames



ASC System



Spectrum Correction

- Deal with **varying frequency response** of the
- **recording devices A, B and C**
- Training set:
  - **Reference recordings: 540** segments of the
  - **same acoustic scenes** for each device A, B and C
  - ***Spectrum***: average of a spectrogram over temporal frames

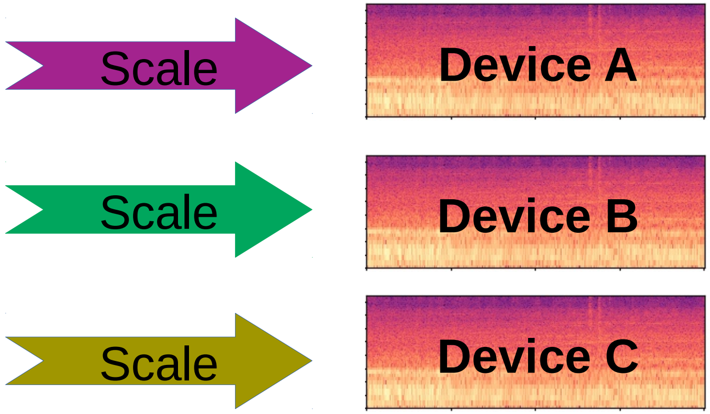
Scaling Coefficients using Reference Devices A, B and C (Ref.ABC)

$$\boxed{Spectrum_{ABC}} = \sum_{i=1}^{540} (Spectrum_{A_i} + Spectrum_{B_i} + Spectrum_{C_i})/3$$

$$Coefficients_{\underline{A}} = \frac{\boxed{Spectrum_{ABC}}}{\sum_{i=1}^{540} Spectrum_{\underline{A_i}}}$$

$$Coefficients_{\underline{B}} = \frac{Spectrum_{ABC}}{\sum_{i=1}^{540} Spectrum_{\underline{B_i}}}$$

$$Coefficients_{\underline{C}} = \frac{Spectrum_{ABC}}{\sum_{i=1}^{540} Spectrum_{\underline{C_i}}}$$



Scale

Device A

Scale

Device B

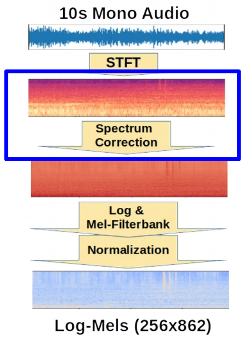
Scale

Device C

ASC System

Spectrum Correction

- Deal with **varying frequency response** of the
- **recording devices A, B and C**
- Training set:
  - **Reference recordings: 540** segments of the
  - **same acoustic scenes** for each device A, B and C
  - ***Spectrum***: average of a spectrogram over temporal frames



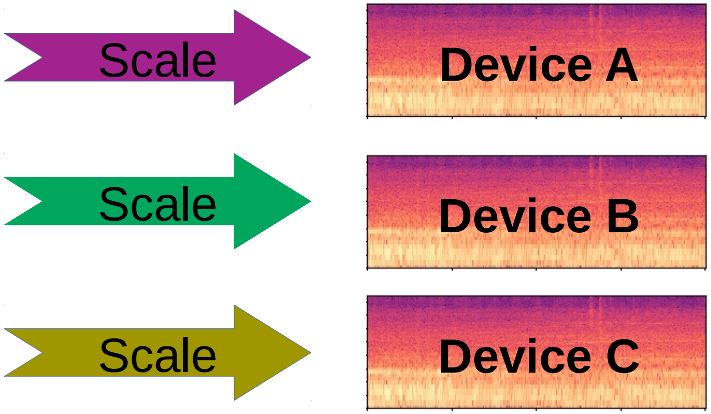
Scaling Coefficients using Reference Devices B and C (Ref.BC)

$$\boxed{Spectrum_{BC}} = \sum_{i=1}^{540} (Spectrum_{B_i} + Spectrum_{C_i})/2$$

$$Coefficients_{\underline{A}} = \frac{\boxed{Spectrum_{BC}}}{\sum_{i=1}^{540} Spectrum_{\underline{A_i}}}$$

$$Coefficients_{\underline{B}} = \frac{Spectrum_{BC}}{\sum_{i=1}^{540} Spectrum_{\underline{B_i}}}$$

$$Coefficients_{\underline{C}} = \frac{Spectrum_{BC}}{\sum_{i=1}^{540} Spectrum_{\underline{C_i}}}$$

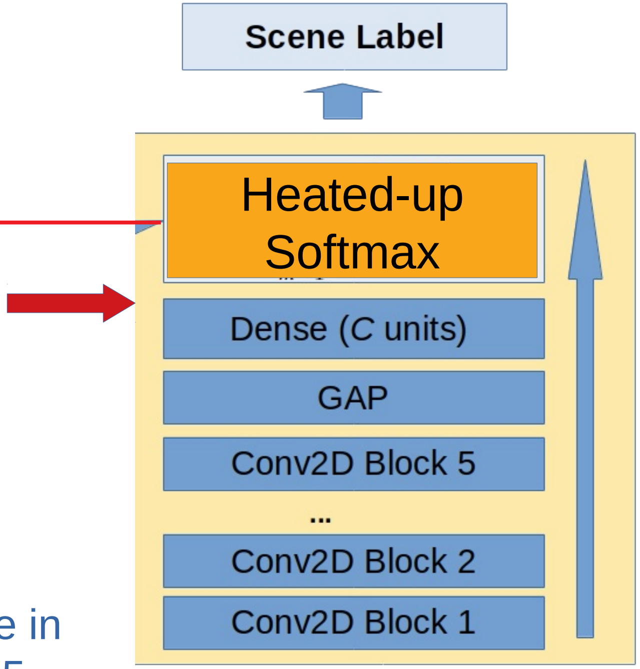


ASC System

Heated-up Softmax

- Temperature scaling*: **calibrating predictions**
- A **higher value** for **T** produces
- a **softer probability distribution** over classes
- Deal with the **shifted data distribution**

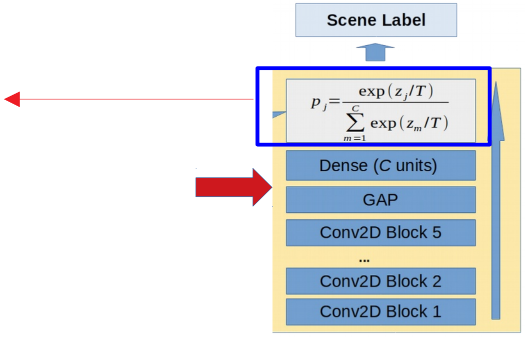
$$p_j = \frac{\exp\left(z_j/T\right)}{\sum_{m=1}^{C}\exp\left(z_m/T\right)}$$



* G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," NIPS 2014 Deep Learning Workshop, 2015.

ASC System

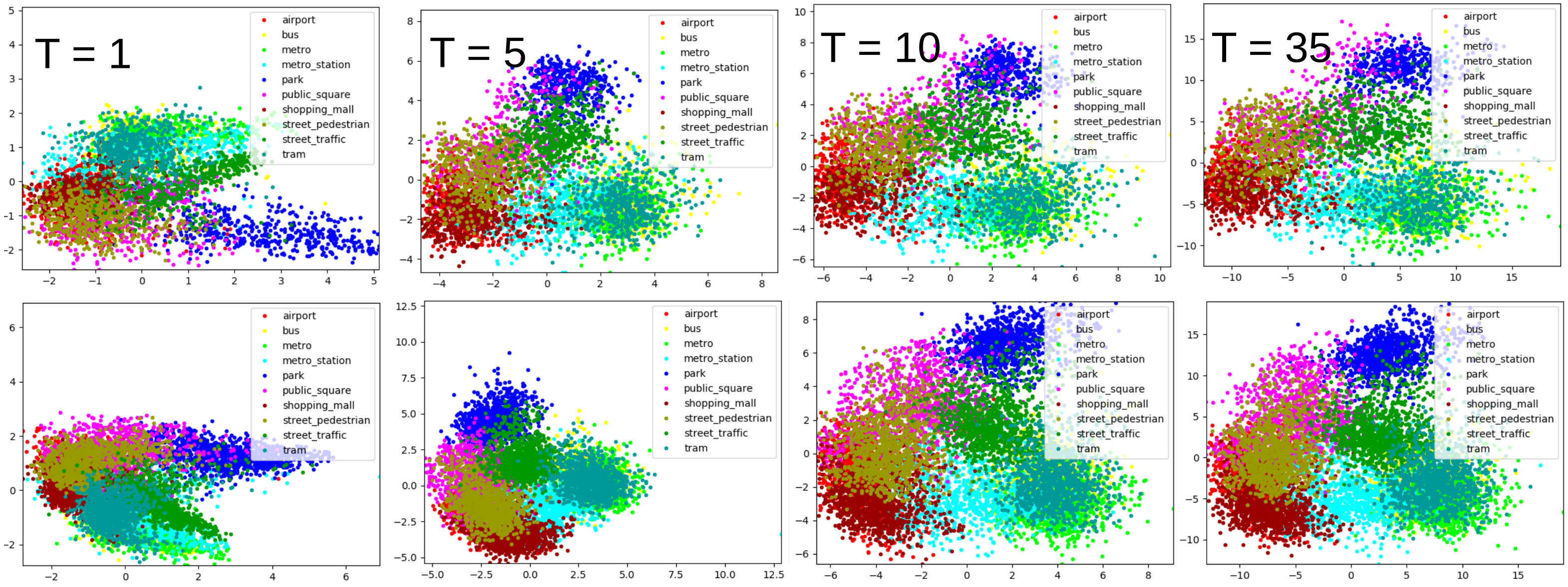
- **Heated-up Softmax**
  - Temperature scaling : **calibrating predictions**
  - A **higher value** for **T** produces
  - a **softer probability distribution** over classes
  - Deal with the **shifted data distribution**



Visualization of the GAP outputs by PCA

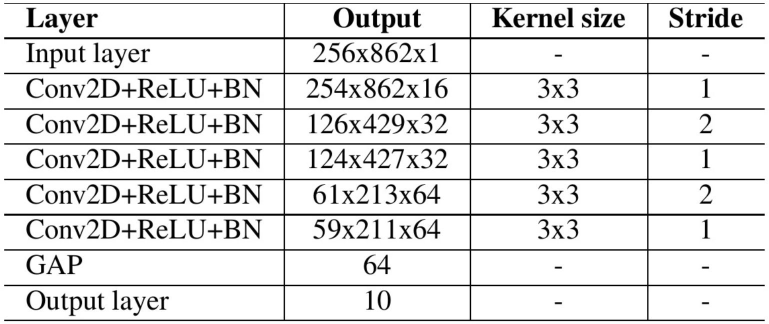
Training set

Test set



ASC System

- **Convolutional Neural Network**
  - The best model of DCASE
  - Challenge 2019 Task 1B
  - A modest number of 70K parameters

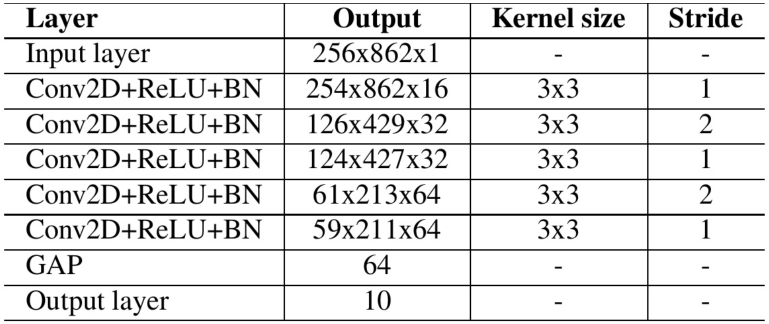


Layer	**Output**	**Kernel size**	**Stride**
Input layer	256x862x1	-	-
Conv2D+ReLU+BN	254x862x16	3x3	1
Conv2D+ReLU+BN	126x429x32	3x3	2
Conv2D+ReLU+BN	124x427x32	3x3	1
Conv2D+ReLU+BN	61x213x64	3x3	2
Conv2D+ReLU+BN	59x211x64	3x3	1
GAP	64	-	-
Output layer	10	-	-

ASC System

Convolutional Neural Network

- The best model of DCASE
- Challenge 2019 Task 1B
- A modest number of 70K parameters



Layer	Output	Kernel size	Stride
Input layer	256x862x1	-	-
Conv2D+ReLU+BN	254x862x16	3x3	1
Conv2D+ReLU+BN	126x429x32	3x3	2
Conv2D+ReLU+BN	124x427x32	3x3	1
Conv2D+ReLU+BN	61x213x64	3x3	2
Conv2D+ReLU+BN	59x211x64	3x3	1
GAP	64	-	-
Output layer	10	-	-

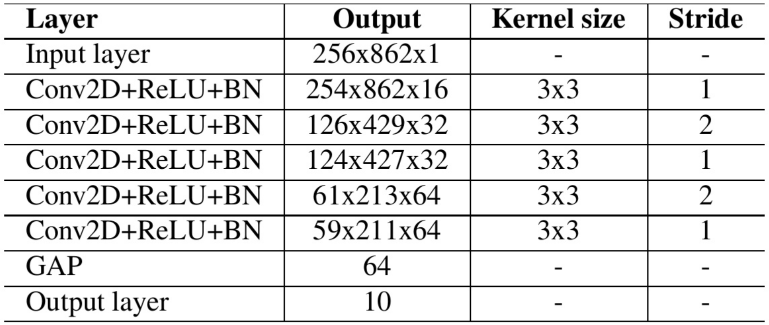
Focal Loss

- Suitable for the shifted data distribution
- Deal with difficult samples

ASC System

Convolutional Neural Network

- The best model of DCASE
- Challenge 2019 Task 1B
- A modest number of 70K parameters



Layer	Output	Kernel size	Stride
Input layer	256x862x1	-	-
Conv2D+ReLU+BN	254x862x16	3x3	1
Conv2D+ReLU+BN	126x429x32	3x3	2
Conv2D+ReLU+BN	124x427x32	3x3	1
Conv2D+ReLU+BN	61x213x64	3x3	2
Conv2D+ReLU+BN	59x211x64	3x3	1
GAP	64	-	-
Output layer	10	-	-

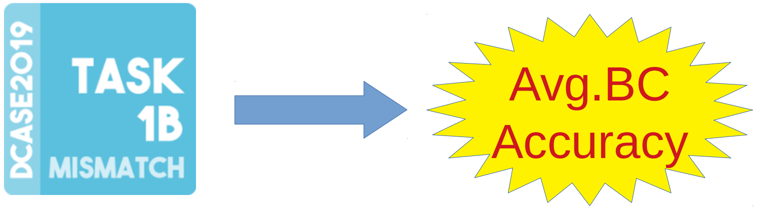
Focal Loss

- Suitable for the shifted data distribution
- Deal with difficult samples

Mixup Data Augmentation

- Suitable for the ASC dataset
- Enhance performance for ASC

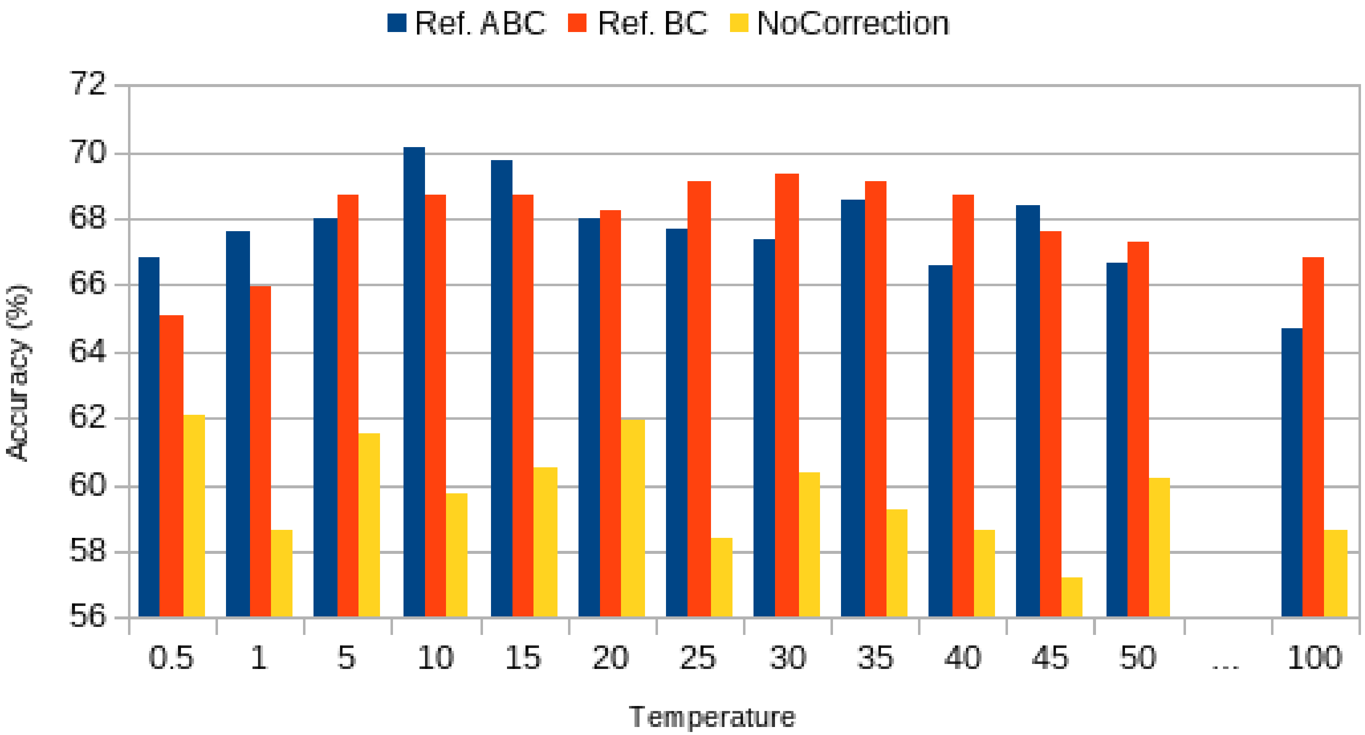
Experimental Results



DCASE2019 TASK 1B MISMATCH

Avg.BC Accuracy

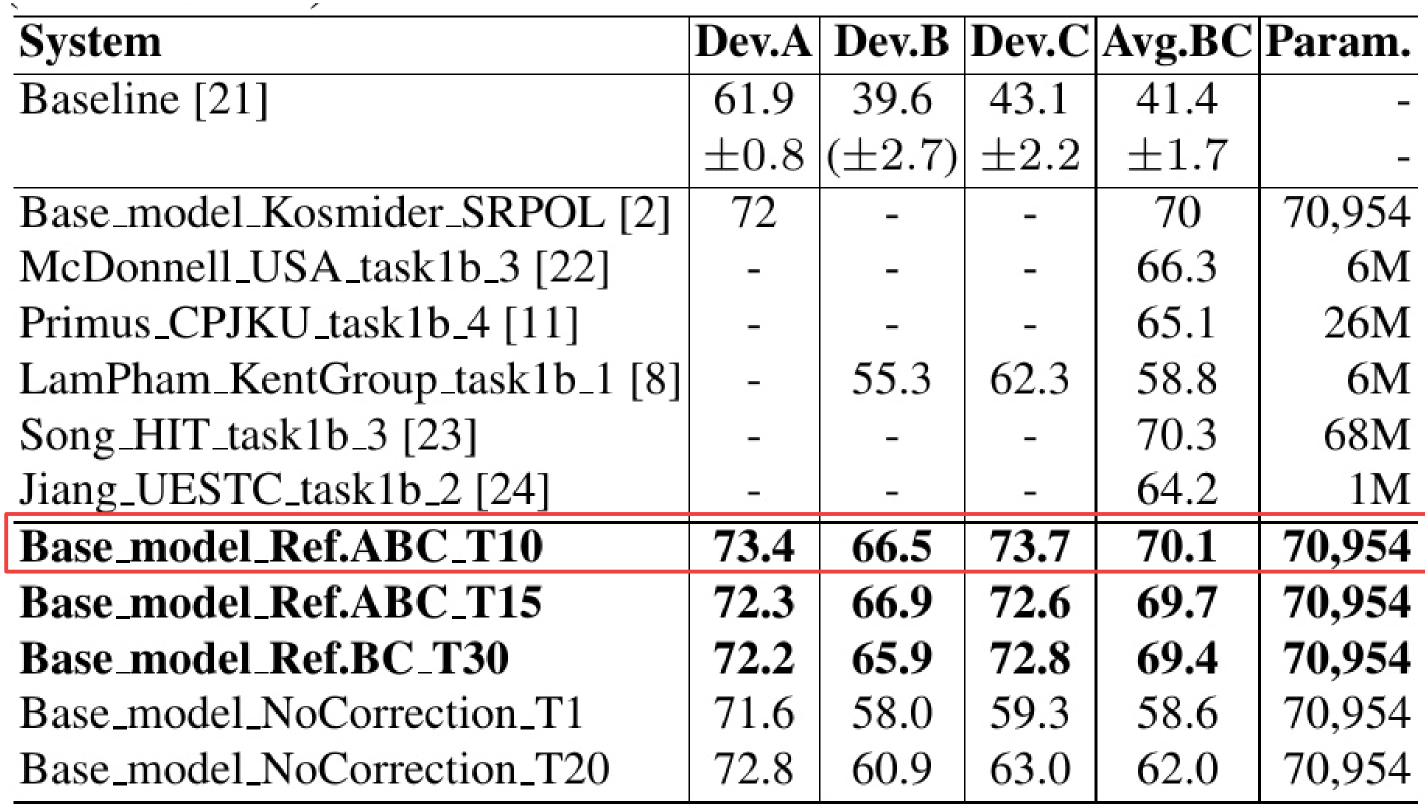
- **Observed spectrum correction and T values**



Experimental Results

Avg.BC Accuracy

- **Accuracy of the proposed system**
- **on test set of DCASE 2019 task 1B**



System	Dev.A	Dev.B	Dev.C	Avg.BC	Param.
Baseline [21]	61.9 ±0.8	39.6 (±2.7)	43.1 ±2.2	41.4 ±1.7	- -
Base_model_Kosmider_SRPOL [2]	72	-	-	70	70,954
McDonnell_USA_task1b_3 [22]	-	-	-	66.3	6M
Primus_CPJKU_task1b_4 [11]	-	-	-	65.1	26M
LamPham_KentGroup_task1b_1 [8]	-	55.3	62.3	58.8	6M
Song_HIT_task1b_3 [23]	-	-	-	70.3	68M
Jiang_UESTC_task1b_2 [24]	-	-	-	64.2	1M
Base_model_Ref.ABC_T10	**73.4**	**66.5**	**73.7**	**70.1**	**70,954**
Base_model_Ref.ABC_T15	**72.3**	**66.9**	**72.6**	**69.7**	**70,954**
Base_model_Ref.BC_T30	**72.2**	**65.9**	**72.8**	**69.4**	**70,954**
Base_model_NoCorrection_T1	71.6	58.0	59.3	58.6	70,954
Base_model_NoCorrection_T20	72.8	60.9	63.0	62.0	70,954

Spectrum Correction
Base CNN model

Conclusion

- Propose temperature scaling of the softmax activation
- function, namely **heated-up softmax** for ASC.
- Observe different versions of **spectrum correction.**
- Our system outperforms many state-of-the-art models of the DCASE 2019 challenge task 1B:
  - At **70.1% accuracy** and about **70 thousand parameters**.
  - At **28.7% of accuracy higher** than the **baseline model** of the DCASE 2019 challenge.

Thank you for your attention!