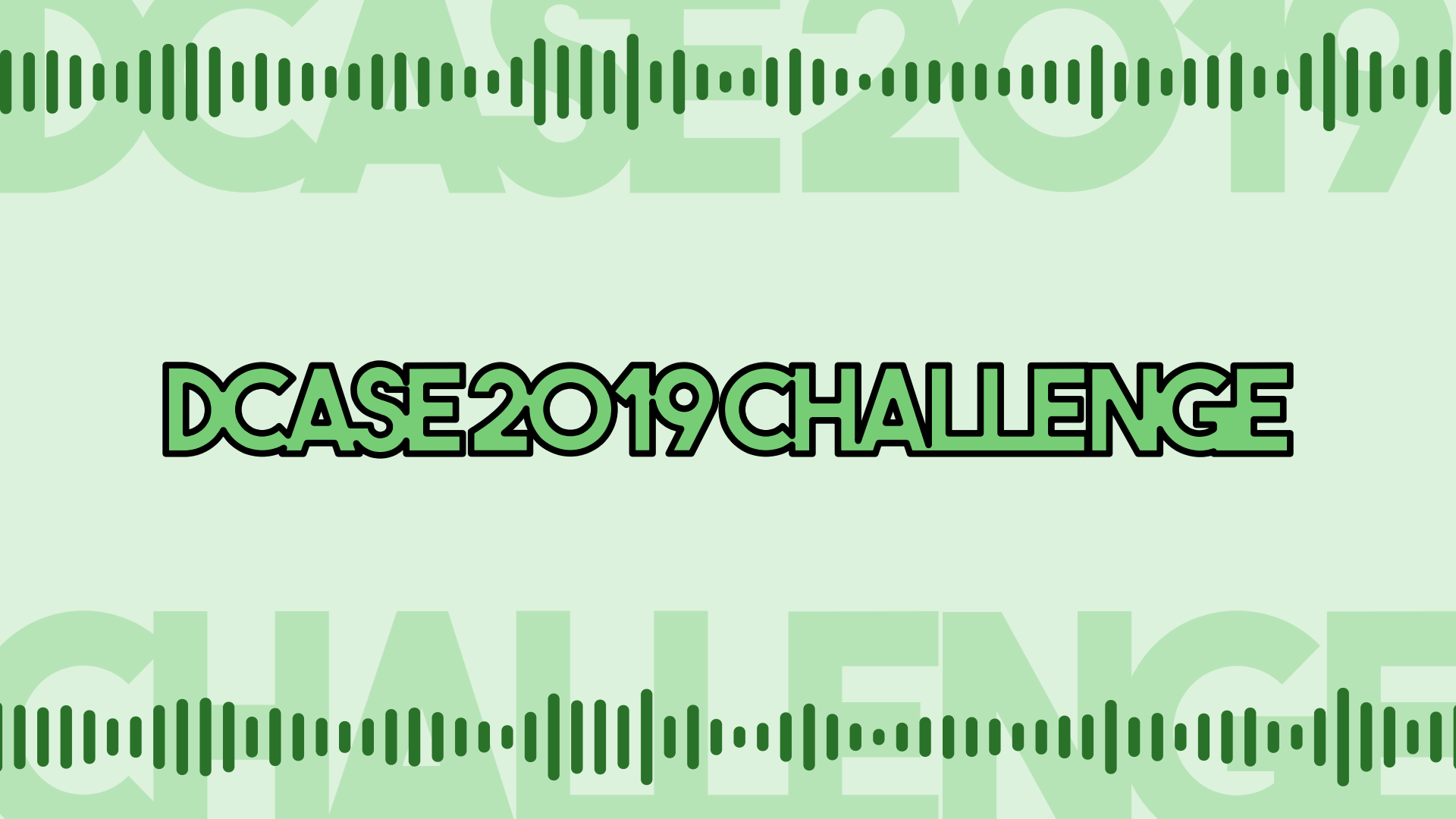

DCASE2019CHALLENGE

DCASE Challenge

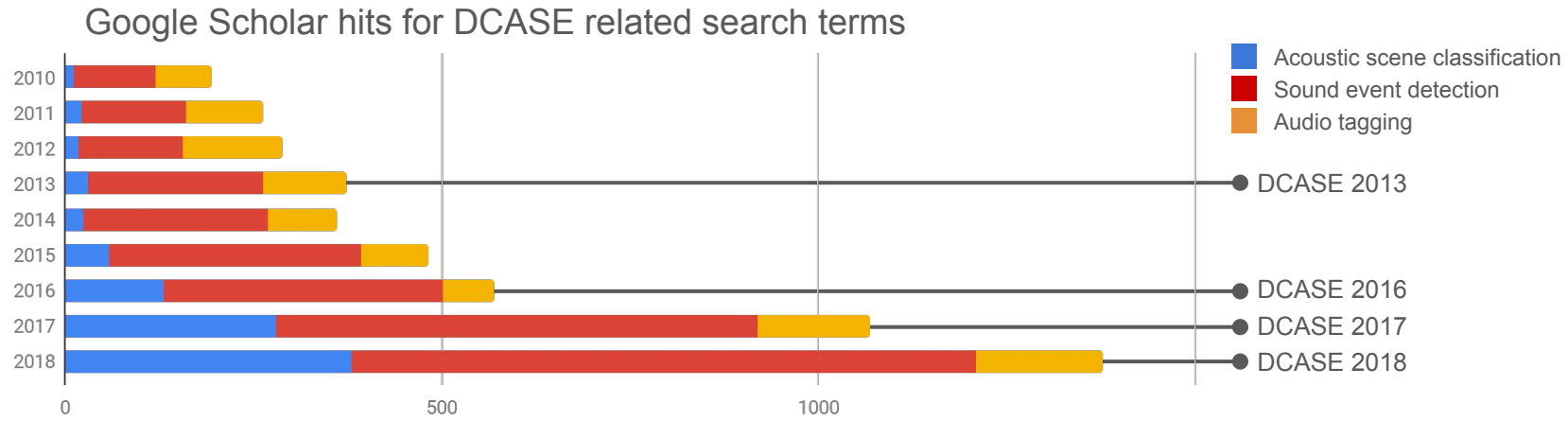
- Aim to provide open data for researchers to use in their work
- Encourage reproducible research
- Attract new researchers into the field
- Create reference points for performance comparison

Participation statistics

Edition	Tasks	Entries	Teams
2013	3	31	21
2016	4	84	67
2017	4	200	74
2018	5	223	81
2019	5	311	109

Outcome

- Development of state of the art methods
- Many new open datasets
- Rapidly growing community of researchers



Challenge tasks 2013 - 2019

Classical tasks:

- **Acoustic scene classification** – textbook example of supervised classification (2013-2019) with increasing amount of data and acoustic variability; mismatched devices (2018, 2019); open set classification (2019)
- **Sound event detection** – synthetic audio (2013-2016), real-life audio (2013-2017), rare events (2017), weakly labeled training data (2017-2019)
- **Audio tagging** – domestic audio, smart cars, Freesound, urban (2016-2019)

Novel openings:

- **Bird detection** (2018) – mismatched training and test data, generalization
- **Multichannel** audio classification (2018)
- Sound event **localization** and detection (2019)

DCASE2019 CHALLENGE

Reproducible system award

Judges' award

Awards sponsored by

SONOS

HARMAN
A SAMSUNG COMPANY

cochl.

oticon
PEOPLE FIRST

sound intelligence

DCASE 2019 Challenge

Task 1: Acoustic Scene Classification

Task 2: Audio Tagging with Noisy Labels and Minimal Supervision

Task 3: Sound Event Localization and Detection

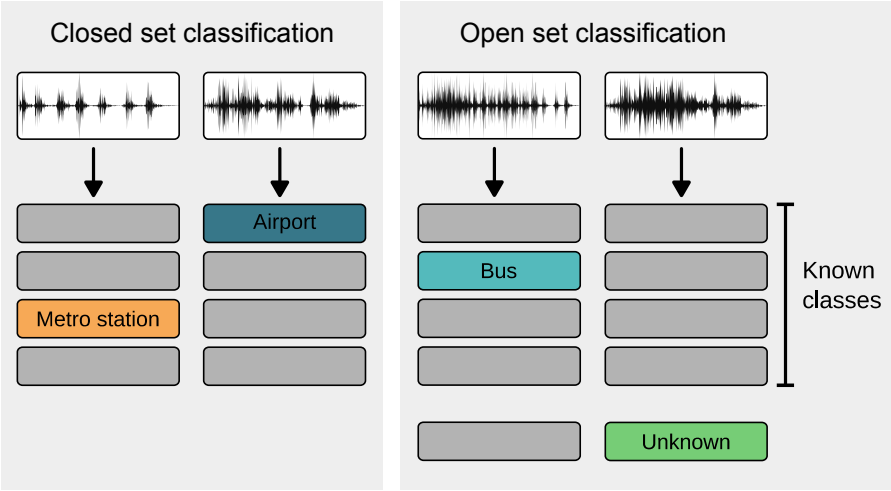
Task 4: Sound Event Detection in Domestic Environments

Task 5: Urban Sound Tagging

Task 1: Acoustic Scene Classification

Classification of audio recordings into one of 10 predefined acoustic scene classes:

- Subtask A: Acoustic Scene Classification
- Subtask B: Acoustic Scene Classification with Mismatched Devices
- Subtask C: Open Set Acoustic Scene Classification



Data: TAU Urban Acoustic Scenes 2019

- 10 classes, 12 cities, 4 devices
- Some parallel data available for Subtask B
- Some “unknown” scenes data available for Subtask C

Task 1: Submissions and results

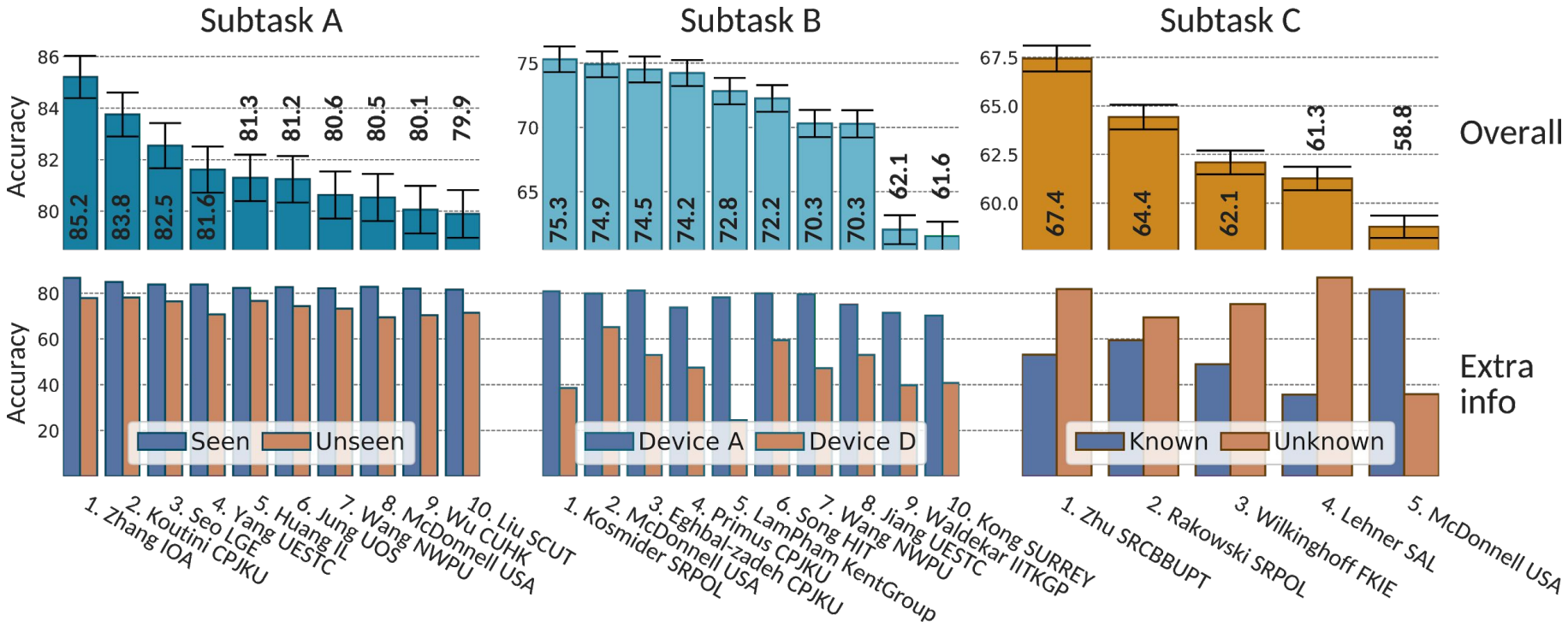
Most popular task throughout the years: 146 submissions this year (98, 29, 19)

All systems easily outperformed the baseline system (small exceptions)

State of the art performance:

- 85% in matching conditions
- 75% with mismatched devices
- 67% in open set scenario

Task 1: Results



Task 1: Summary

- Solution is dominated by ensemble classifiers, most of them being CNNs
- Augmentation by mixup became common/default pre-processing method
- Mel energies still rule the feature domain
- External data usage was minimal

- Subtask A attracted most participants, as a textbook classification problem
- Specific methods emerged for Subtask B compared to DCASE 2018
- Subtask C as the novelty item gathered least interest

Task 2: Audio tagging with noisy labels and minimal supervision

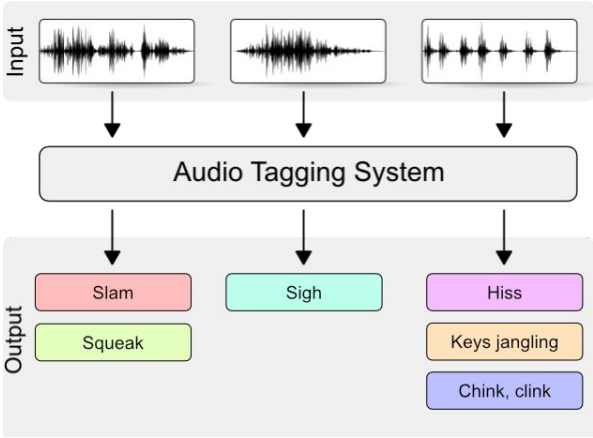
General purpose sound event recognition

Follow-up of last year's edition

- 2x number of classes
- more data
- multi-class → multi-label

Goal: multi-label audio tagging

- a small set of manually-labeled data
- a larger set of noisy-labeled data
- 80 classes of everyday sounds

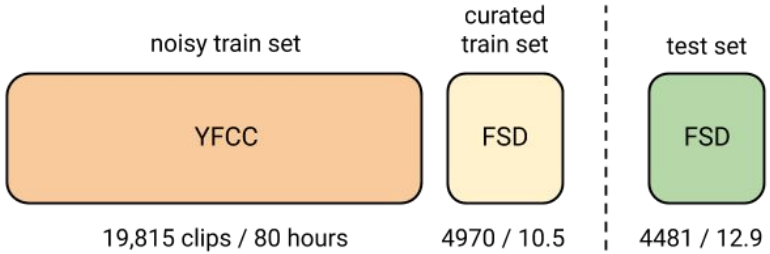


Task 2 Dataset: FSDKaggle2019

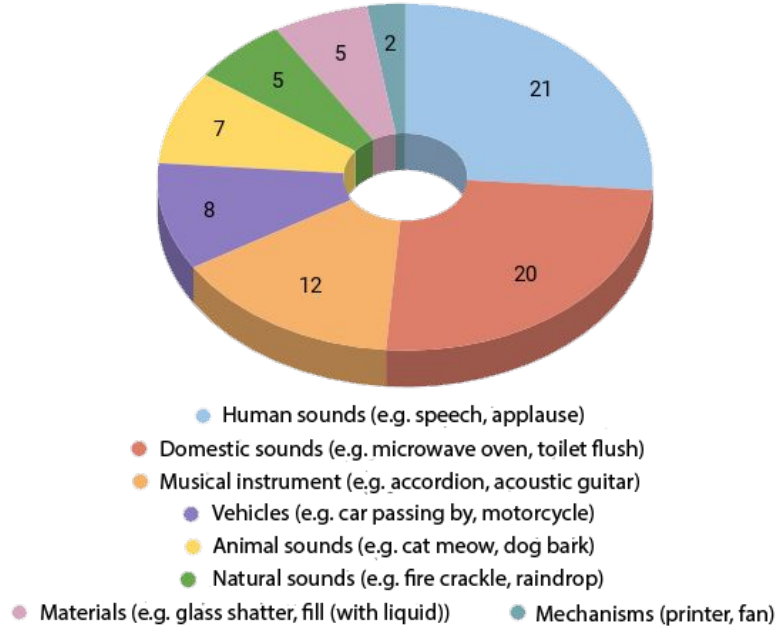
freesound → FSD

flickr → YFCC100M

- 80 classes of everyday sounds / 100+ hours
- Three types of labels
  - test set: exhaustive
  - curated train set: correct but potentially incomplete
  - noisy train set: noisy (machine-generated)

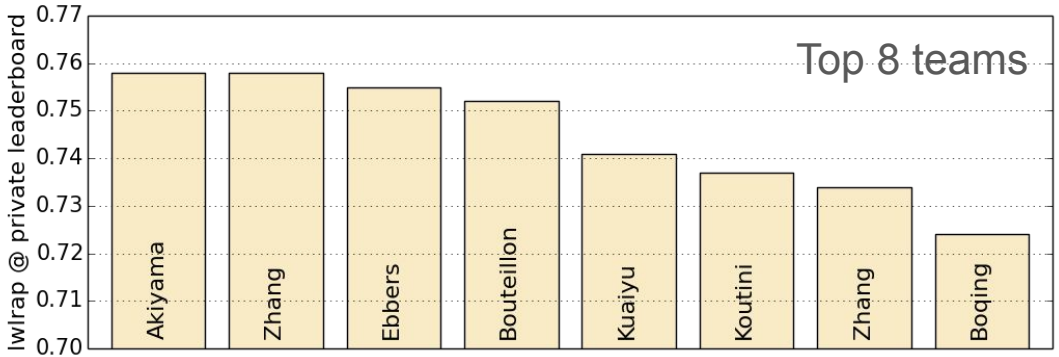


- Potential acoustic mismatch
  - Freesound - Flickr



Task 2 Numbers

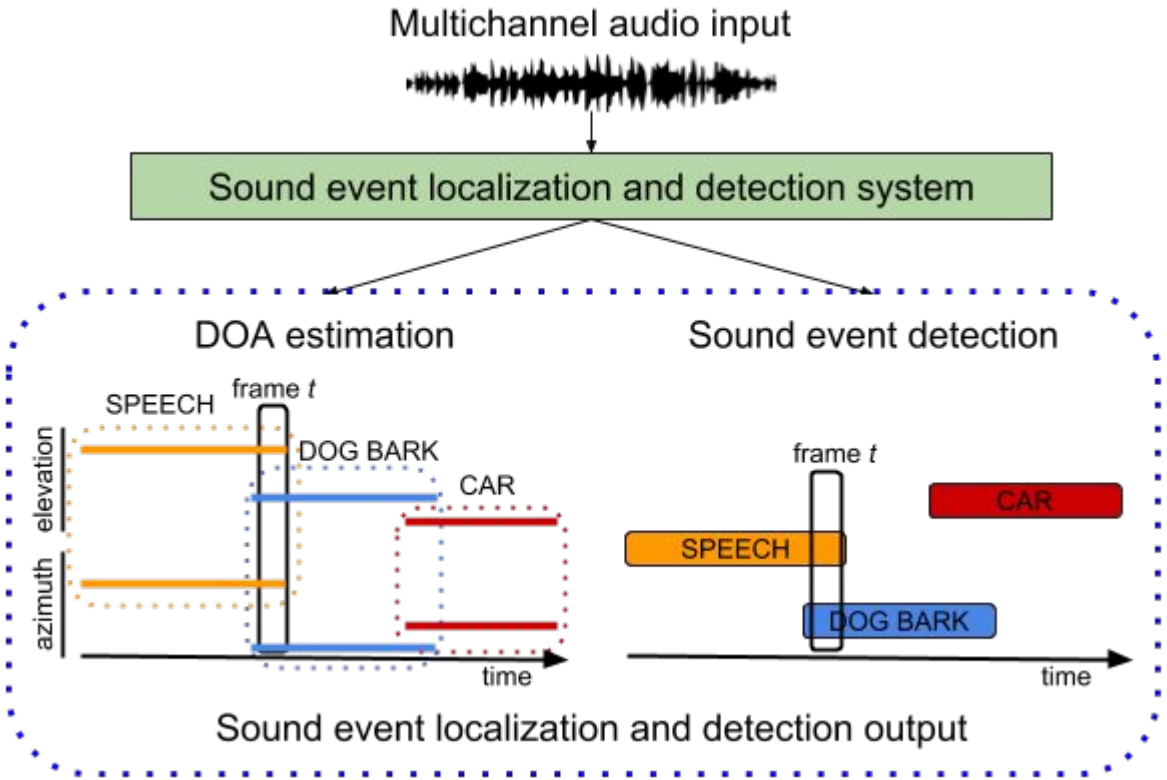
- Run on kaggle
- 880 teams / 8618 entries:
  - some teams only made few entries
  - 14 teams submitting 28 systems to DCASE
- Lots of knowledge spread in the discussion forum
- Evaluation: label-weighted **label-ranking** average precision (lwlrap)



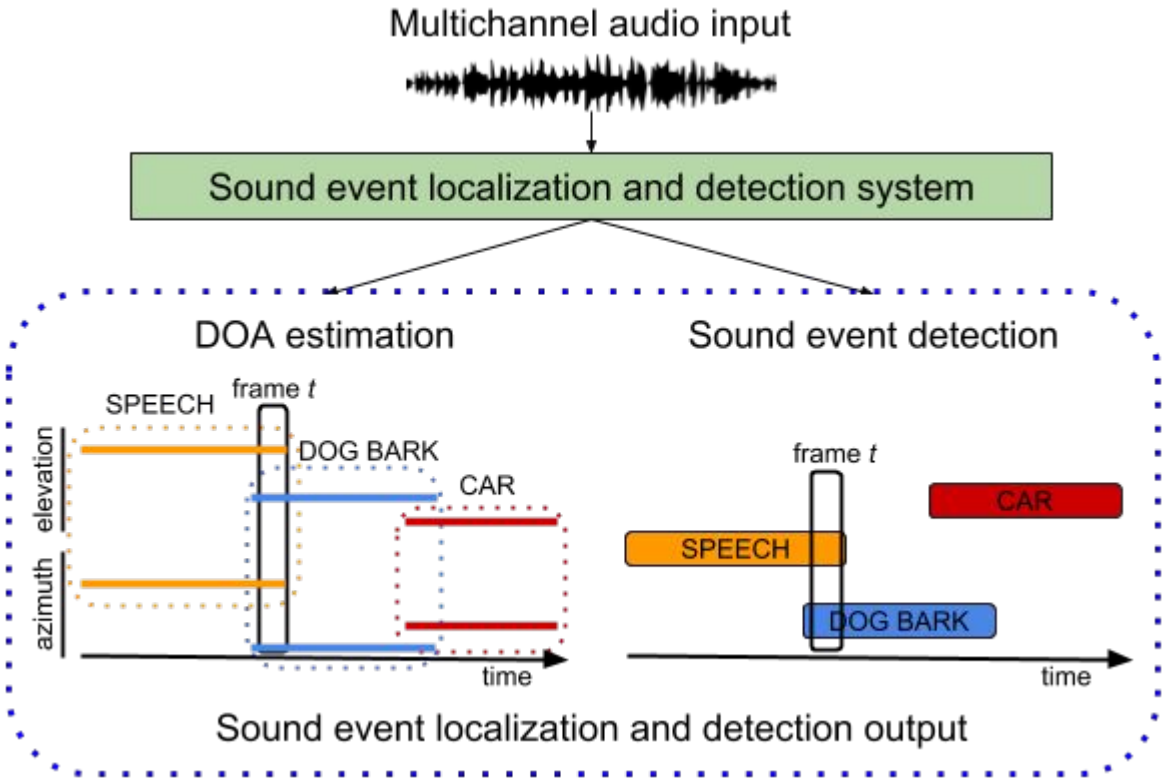
Task 2 Takeaways

- **Log-mel energies**, waveform, CQT
- Mainly **CNN**/CRNN: VGG, DenseNet, ResNe(X)t, Shake-Shake, Frequency-Aware CNNs, Squeeze-and-Excitation, EnvNet, MobileNet
- Heavy usage of **ensembles** (2 → 170)
- **Augmenting** curated train set: mix-up, SpecAugment, SpecMix, TTA
- Label noise: **variety** of approaches rather than common trend
  - semi-supervised learning
  - multi-task learning
  - robust loss functions

Task 3: Sound Event Localization and Detection

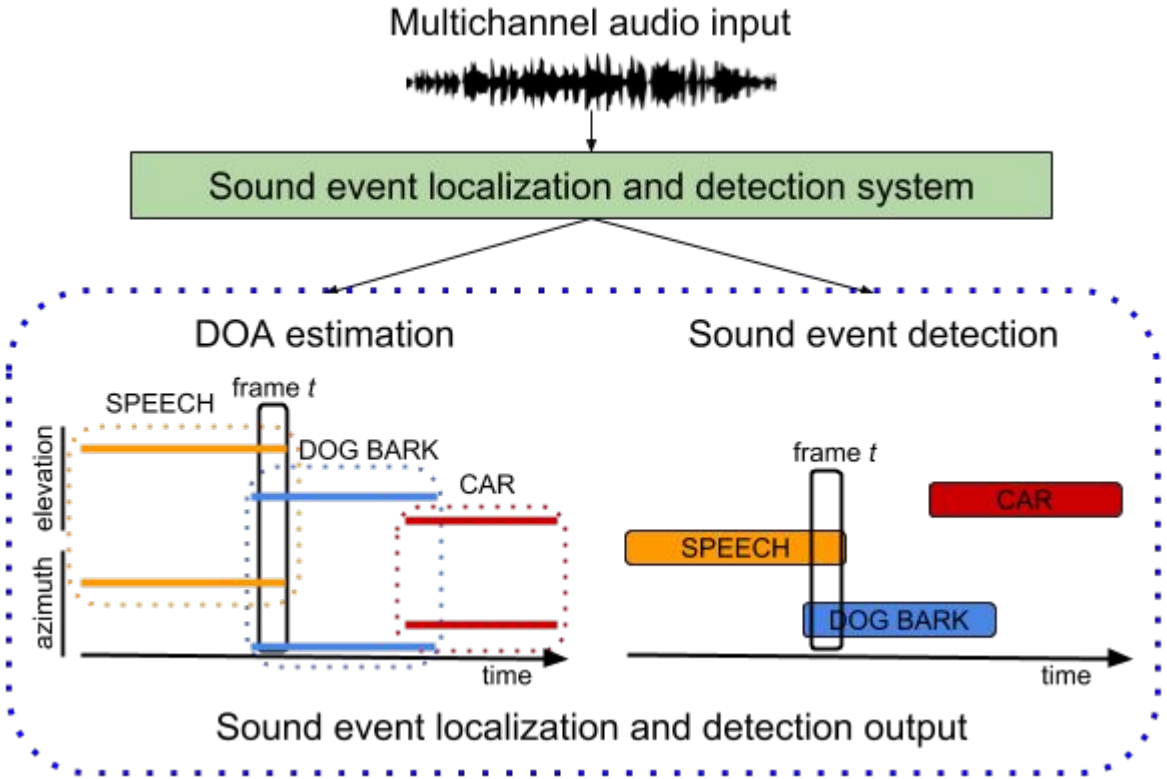


Task 3: Sound Event Localization and Detection



Input: Multichannel audio

Task 3: Sound Event Localization and Detection



Input: Multichannel audio

Output:

- Identify known set of sound classes
- their temporal onset-offset
- spatial location in 2D (azimuth and elevation angles)

Task 3: Dataset

- Two (four-channel) audio formats - Ambisonic and microphone array signals
  - Identical sound scene, captured with different microphone-configurations
  - Participants allowed to choose either or both formats

Task 3: Dataset

- Two (four-channel) audio formats - Ambisonic and microphone array signals
  - Identical sound scene, captured with different microphone-configurations
  - Participants allowed to choose either or both formats
- Train methods on development set (400 mins), and test on unseen evaluation set (100 mins)

Task 3: Dataset

- Two (four-channel) audio formats - Ambisonic and microphone array signals
  - Identical sound scene, captured with different microphone-configurations
  - Participants allowed to choose either or both formats
- Train methods on development set (400 mins), and test on unseen evaluation set (100 mins)
- The recording consisted of sound events from 11 classes, each associated with azimuth and elevation angles sampled at 10-degree resolution.
  - complete azimuth
  - elevation from -40 to 40 degrees

Task 3: Dataset

- Two (four-channel) audio formats - Ambisonic and microphone array signals
  - Identical sound scene, captured with different microphone-configurations
  - Participants allowed to choose either or both formats
- Train methods on development set (400 mins), and test on unseen evaluation set (100 mins)
- The recording consisted of sound events from 11 classes, each associated with azimuth and elevation angles sampled at 10-degree resolution.
  - complete azimuth
  - elevation from -40 to 40 degrees
- The dataset has equal distribution of
  - two-polyphonies (single and upto two overlapping sound events) and,
  - impulse responses from five different indoor environments

Task 3: Top 10 team results

Systems	Format	Method	Features	ER	F (%)	DE (°)	FR (%)
Kapka	FOA	CRNN	Phase and magnitude spectrogram	0.08	94.7	3.7	96.8
Cao	BOTH	CRNN	Log-mel and intensity vectors	0.08	95.5	5.5	92.2
Xue	MIC	CRNN	Log-mel, Q-transform, cross-power spectrum, and phase spectrogram	0.06	96.3	9.7	92.3
He	FOA	CRNN	Log-mel, phase, and magnitude spectrogram	0.06	96.7	22.4	94.1
Jee	MIC	CRNN	Log-mel spectrogram and GCC-PHAT	0.12	93.7	4.2	91.8
Nguyen	FOA	CRNN, DOA Par.	Log-mel, phase, and magnitude spectrogram	0.11	93.4	5.4	88.8
Mazzon	BOTH	CRNN, ResNet	Log-mel spectrogram and GCC-PHAT	0.1	94.2	6.4	88.8
Chang	MIC	CRNN, CNN	Log-mel spectrogram, cochleagram, and GCC-PHAT	0.14	91.9	2.7	90.8
Ranjan	MIC	ResNet-RNN	Log-mel and phase spectrogram	0.16	90.9	5.7	91.8
Park	BOTH	CRNN, TrellisNet	Log-mel and intensity vectors	0.15	91.9	5.1	87.4
Baseline	FOA	CRNN	Phase and Magnitude Spectrogram	0.28	85.4	24.6	85.7
Baseline	MIC	CRNN	Phase and Magnitude Spectrogram	0.3	83.2	38.1	83.4

Task 3: Results

- Submissions**:** 58 Systems - 22 Teams, 65 Authors from 24 Affiliations (8 Industry). **Second popular DCASE task**.

Task 3: Results

- Submissions**:** 58 Systems - 22 Teams, 65 Authors from 24 Affiliations (8 Industry). **Second popular DCASE task**.
- Method**:** Except for one team which employed CNN, **all teams** used **CRNN (21/22)** as one of their classifiers.

Task 3: Results

- Submissions**:** 58 Systems - 22 Teams, 65 Authors from 24 Affiliations (8 Industry). **Second popular DCASE task**.
- Method**:** Except for one team which employed CNN, **all teams** used **CRNN (21/22)** as one of their classifiers.
- Joint learning: About **half the systems (10/22)** employed **multi-task learning**. Remaining systems, including the **top system**, performed different kinds of **engineering for data association** of detection and localization.

Task 3: Results

- Submissions**:** 58 Systems - 22 Teams, 65 Authors from 24 Affiliations (8 Industry). **Second popular DCASE task**.
- Method**:** Except for one team which employed CNN, **all teams** used **CRNN (21/22)** as one of their classifiers.
- Joint learning: About **half the systems (10/22)** employed **multi-task learning**. Remaining systems, including the **top system**, performed different kinds of **engineering for data association** of detection and localization.
- Parametric DOA estimation: Few systems (3/22) experimented using parametric DOA estimation in association with deep-learning based SED. **Best parametric system** achieved **17th position**.

Task 3: Results

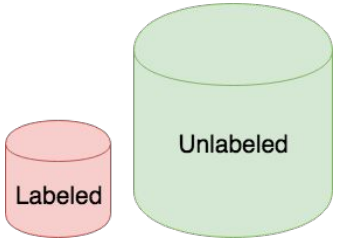
- Submissions**:** 58 Systems - 22 Teams, 65 Authors from 24 Affiliations (8 Industry). **Second popular DCASE task**.
- Method**:** Except for one team which employed CNN, **all teams** used **CRNN (21/22)** as one of their classifiers.
- Joint learning: About **half the systems (10/22)** employed **multi-task learning**. Remaining systems, including the **top system**, performed different kinds of **engineering for data association** of detection and localization.
- Parametric DOA estimation: Few systems (3/22) experimented using parametric DOA estimation in association with deep-learning based SED. **Best parametric system** achieved **17th position**.
- Audio format: Methods proposed in both formats performed comparably. **No obvious choice.**

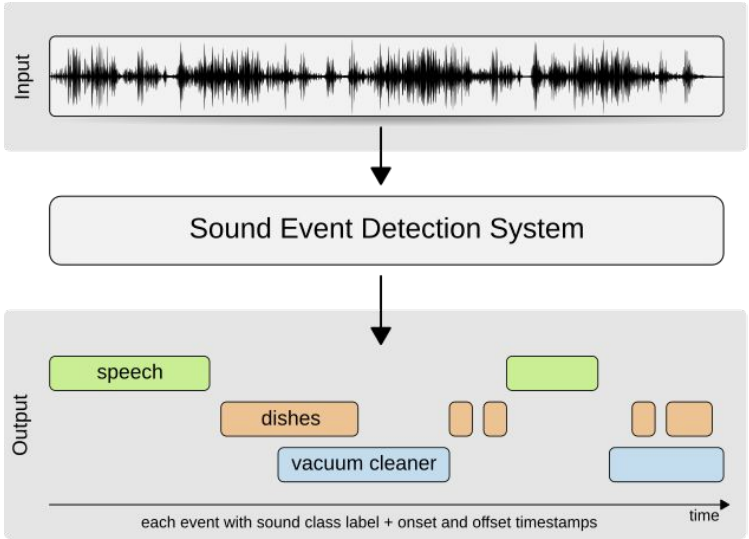
Task 4: Sound event detection in domestic environments

Dataset: 10 s audio clips from audioset, 10 sound event classes

- Weak labels

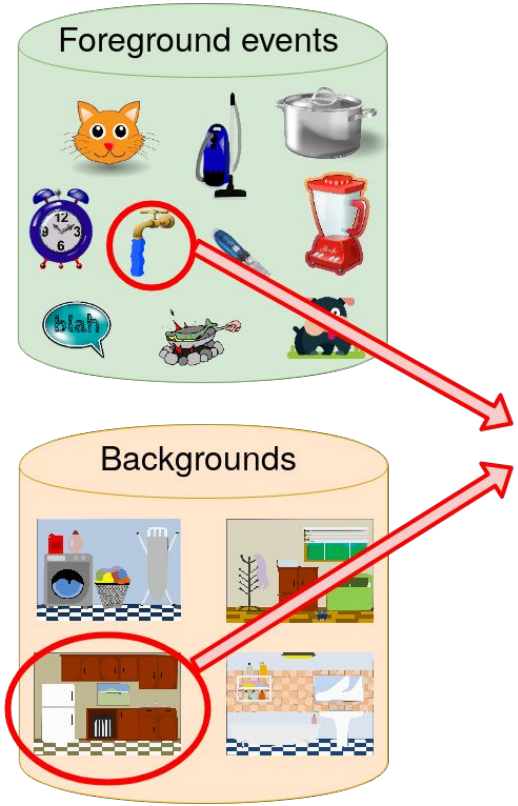
- Small labeled set



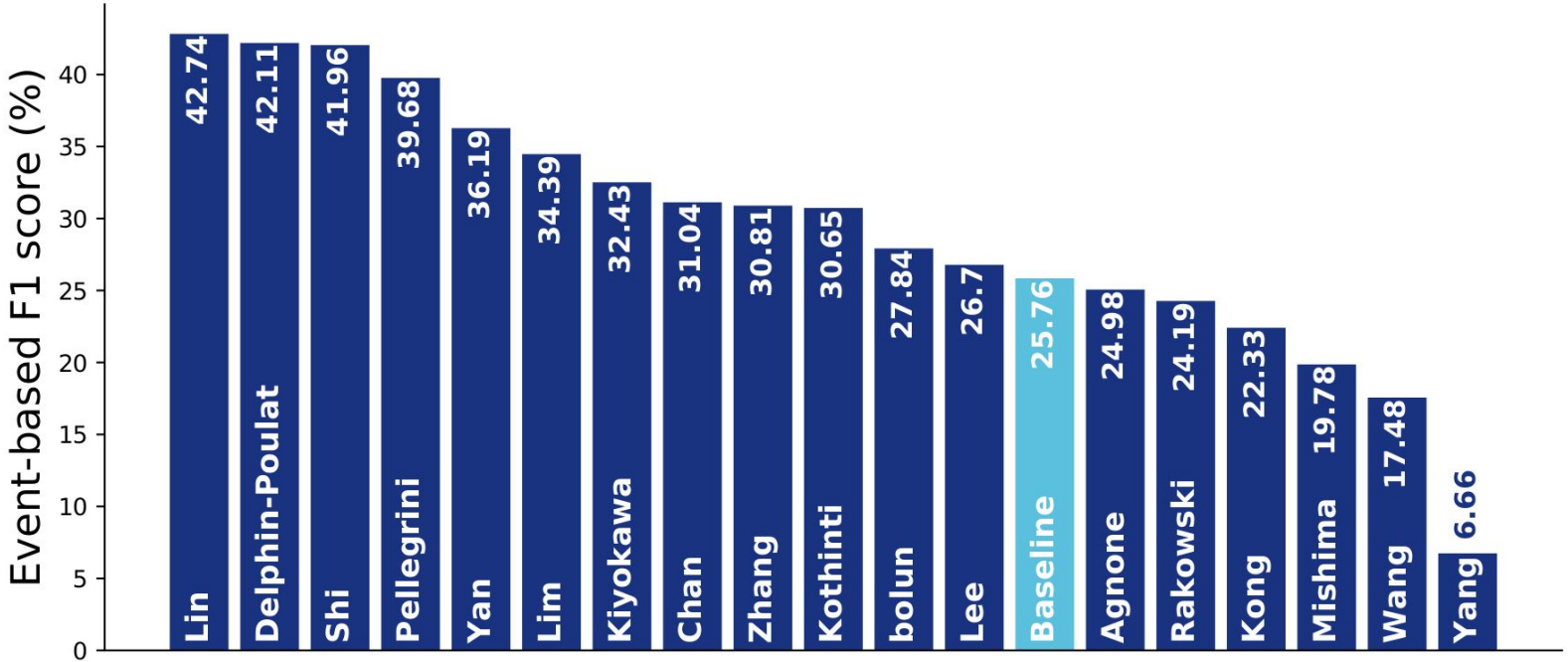


Task 4: Synthetic soundscapes

- Isolated events from the Freesound dataset
- Backgrounds from SINS and MUSAN dataset and youtube videos.
- Distribution similar to the real data.



Task 4: Results



Task 4: Summary

Task 4 overview

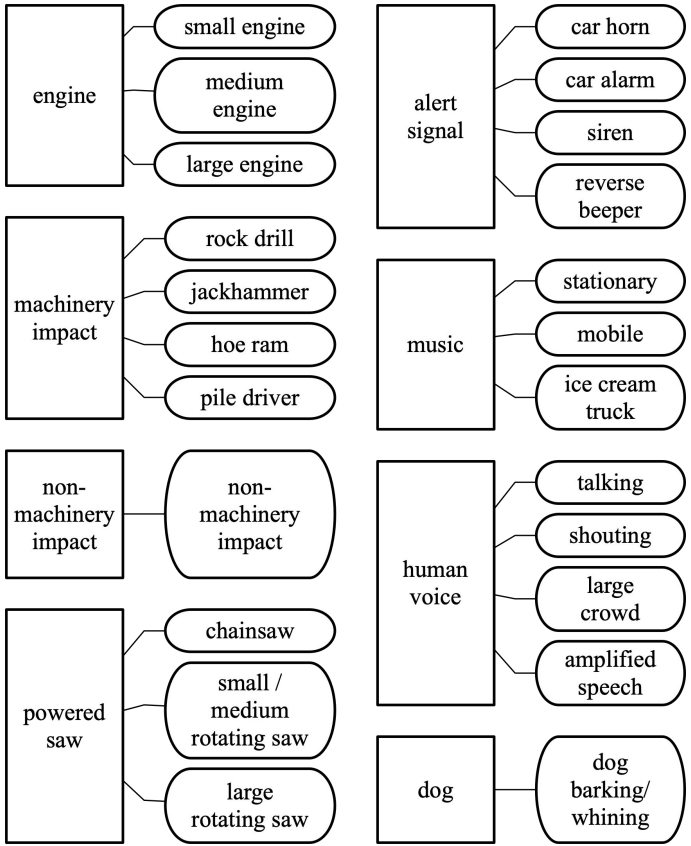
- Steady number of participants
- Last year's top performing system: outperformed by more than 10%

Task 4 in the workshop

- **Friday 13.40 (Posters I)** - Wootaek Lim: SpecAugment for sound event detection in domestic environments using ensemble of convolutional recurrent neural networks
- **Friday 16.40 (L08)** - Liwei Lin, Xiangdong Wang, Hong Liu, Yueliang Qian: Guided learning convolution system for DCASE 2019 task 4 **(top performing system)**
- **Saturday 13.40 (Posters II)** - Chan Teck Kai, Chan Teck Kai, Chin Cheng Siong, Li Ye: Non-negative matrix factorization-convolutional neural network (NMF-CNN) for sound event detection

Task 5: Urban Sound Tagging

- Multilabel tagging 10s urban sensor recordings on coarse and fine categories

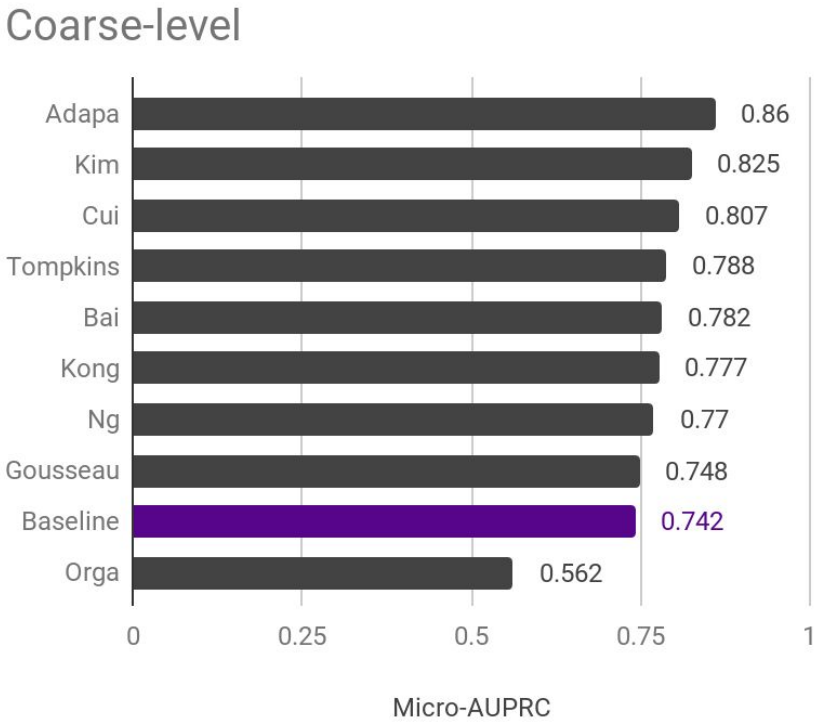


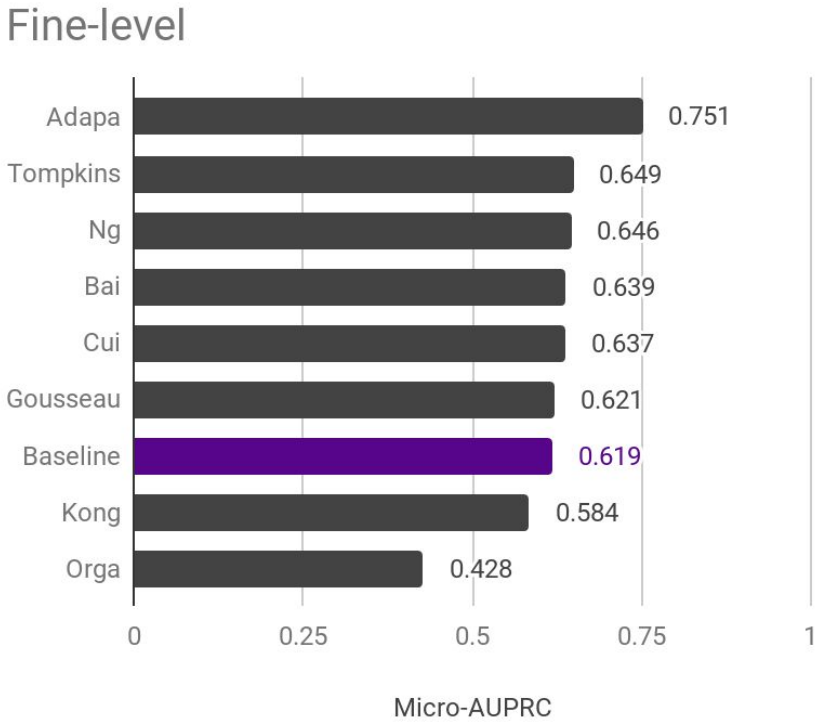
Task 5: SONYC Urban Sound Tagging Dataset

Recorded from 44 acoustic sensors in New York City

- Labels:
  - 23 fine-level classes
  - 8 coarse-level classes
- Splits:
  - 2351 recordings in *train*, each annotated by 3 Zooniverse volunteers
  - 443 recordings in *validate*, annotated by the SONYC research team
  - 274 recordings in *test*, annotated by the SONYC research team
- Additional metadata:
  - Sensor ID
  - Annotator ID
  - Proximity of each class (*near/far/unsure*)

Task 5: Results





DCASE 2020 Challenge

Call for task proposals is now open

- Review process: Steering Committee reviews and selects the tasks
- Proposal: maximum 2 pages, given structure
- Deadline : 1 Dec 2019
- Planned challenge opening: 1 March 2020
- Challenge coordinators will provide support and guidance during the challenge
- **New**: collaborative tasks are encouraged, aiming to minimize task overlap