# DCASE2022 CHALLENGE

# DCASE Challenge 2013 - 2022

- Well-established challenge, based on open data and different tasks to solve
- Important channel for publishing new datasets and asking new research questions on environmental sound scene analysis
- Expanding in terms of research topics: audio classification, but also source localization/separation, video, language
- Diverse setups in terms of ML, from supervised classification to weak supervision, unsupervised learning, few-shot learning

# 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 |
| **2020** | 6 | 473 | 138 |
| **2021** | 6 | 394 | 127 |
| **2022** | 6 | 410 | 135 |

# DCASE2022 CHALLENGE


**Judges' award**

# DCASE 2022 Challenge


**Task 1**:  Low-Complexity Acoustic Scene Classification


**Task 2**: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques


**Task 3**: Sound Event Localization and Detection Evaluated in Real Spatial Sound Scenes


**Task 4**: Sound Event Detection in Domestic Environments


**Task 5**: Few-shot Bioacoustic Event Detection

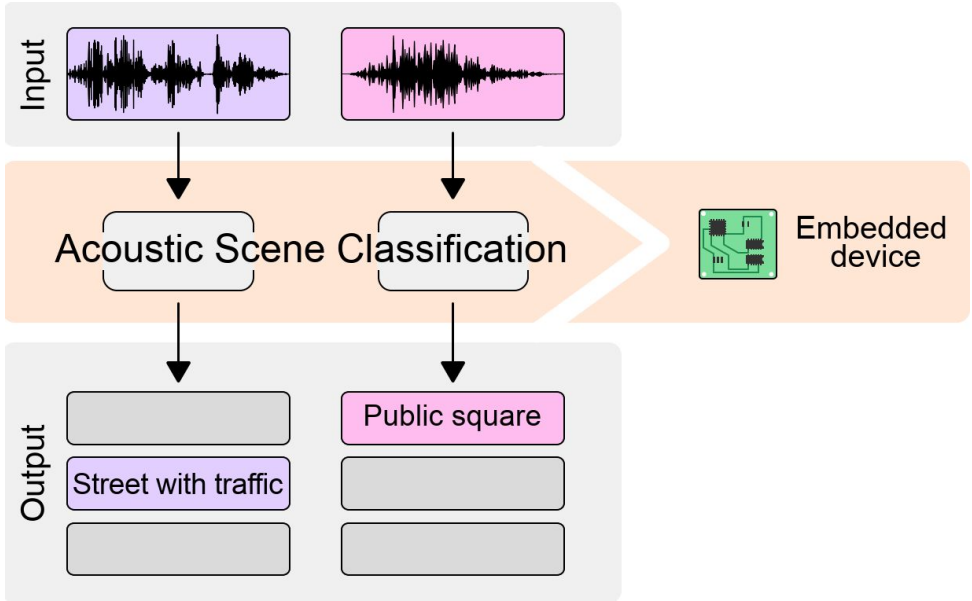

**Task 6**: Automated Audio Captioning and Language-Based Audio Retrieval

# DCASE2022 CHALLENGE


## Task 1: Low-Complexity Acoustic Scene Classification

# Task 1: Low-Complexity Acoustic Scene Classification

Classification of audio recordings into one of ten predefined scene classes




**New in 2022:**

- 128 K parameters (total, incl. zero) in INT8 representation
- 30 MMACs (suitable for Cortex-M4 devices)
- 1s audio clips (short inference time)

# Low-Complexity Acoustic Scene Classification

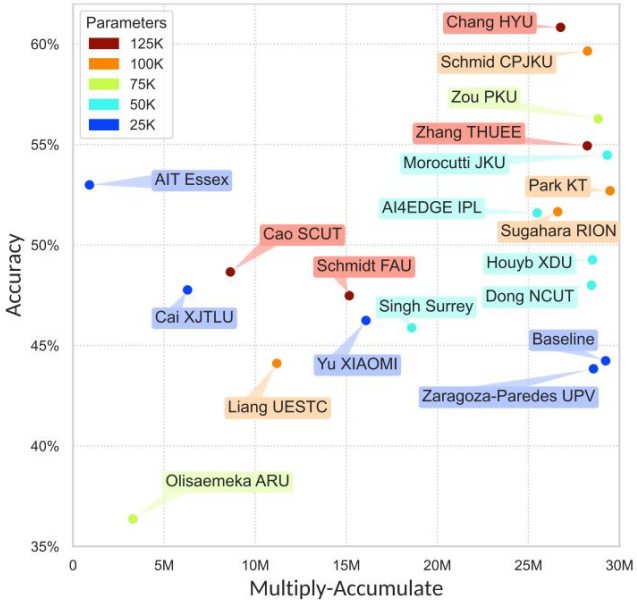
**Submissions**: 48 systems / 19 teams

# Low-Complexity Acoustic Scene Classification

**Submissions**: 48 systems / 19 teams

**Overall trends:**

- As always, mel energies and augmentation
- CNNs, some residual networks (not as much as in 2021)
- Most systems are close to the allowed limits for complexity
- Simple networks, but focus on training and augmentation: quantization-aware training, knowledge distillation



# Task 1: Summary

- Despite clever architectural designs, networks trained with optimized preprocessing and training strategies outperform the other approaches
- Small number of submissions compared to previous years.
  - Is the task too difficult?
  - Is ASC not of interest anymore?
- What kind of ASC task is still topical?

DCASE2022 CHALLENGE

Monitoring Task 2

# Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques

Monitoring Task 2

DCASE2022



Kota Dohi


Keusuke Imoto


Noboru Harada


Daisuke Niizumi


Yuma Koizumi

Tomoya Nishida  Harsh Purohit  Takashi Endo  Masaaki Yamamoto  Yohei Kawaguchi

HITACHI Inspire the Next

Doshisha University

Google

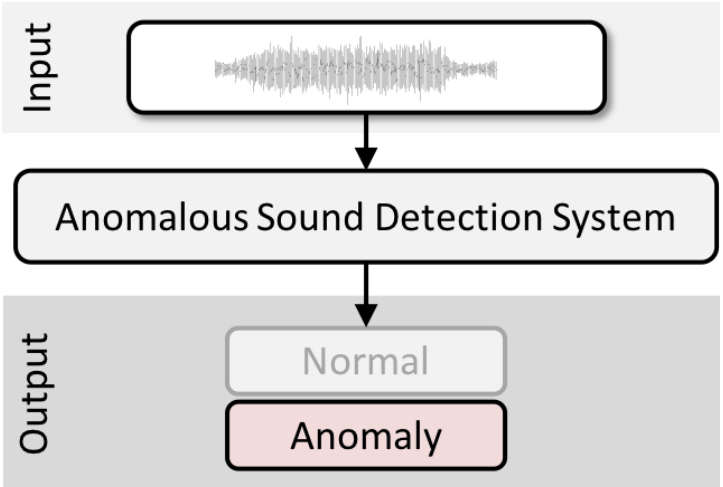
NTT

# Task scope

Monitoring
Task 2

## ◻ Anomalous Sound Detection (ASD)

Determine if a machine is **normal** or **anomalous** from sound




Background photo created by fanjianhua - www.freepik.com
https://www.freepik.com/photos/background

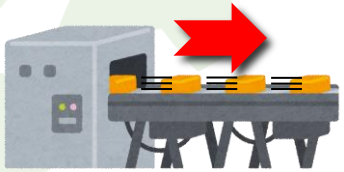
# Challenge

Monitoring
Task 2

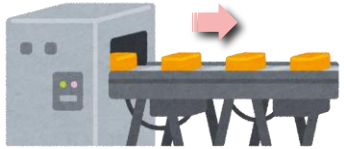
**How can we handle domain shifts?**

Domain shifts : Differences in machine's operational states or the environment

Winter
(source domain)



Summer
(target domain)



Domain shifts can significantly degrade the detection performance

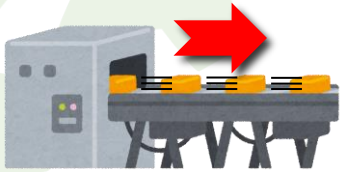
# Challenge

Monitoring
Task 2

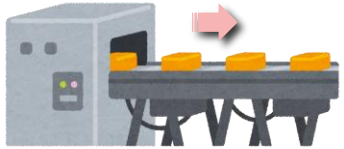
**How can we handle domain shifts?**

Domain shifts : Differences in machine's operational states or the environment

Winter
(source domain)



Summer
(target domain)



**adaptation**

Domain shifts can significantly degrade the detection performance

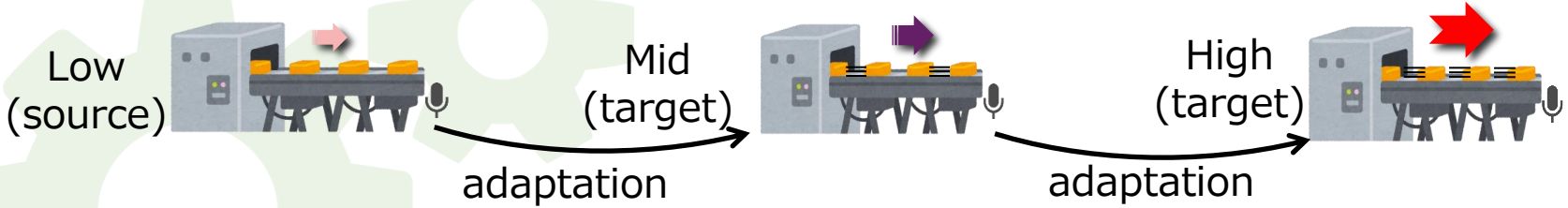
Adaptation of the model can be useful (DCASE2021 Task 2)

# Focus in 2022

Monitoring
Task 2

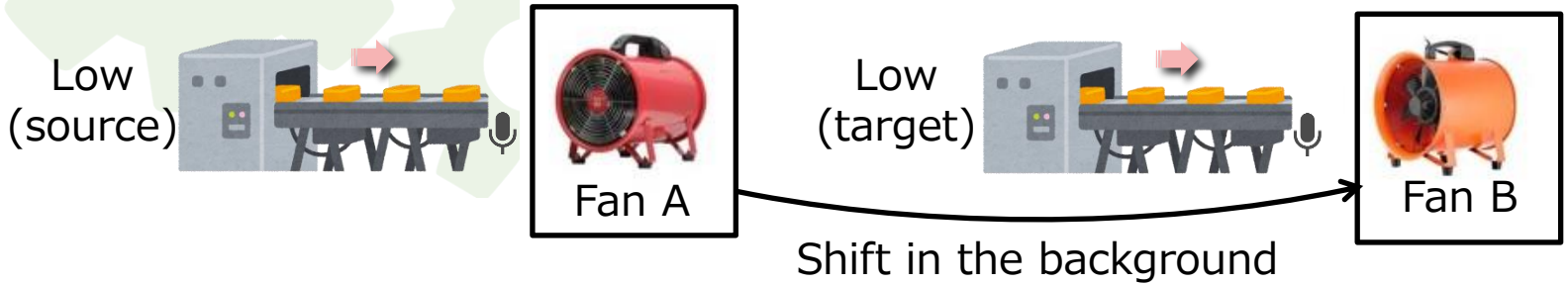
**Can we handle domain shifts without adaptation?**

Case1: Domain shifts can occur frequently




**Adaptation every time can be costly**

Case2: Domain shifts can be hard to notice

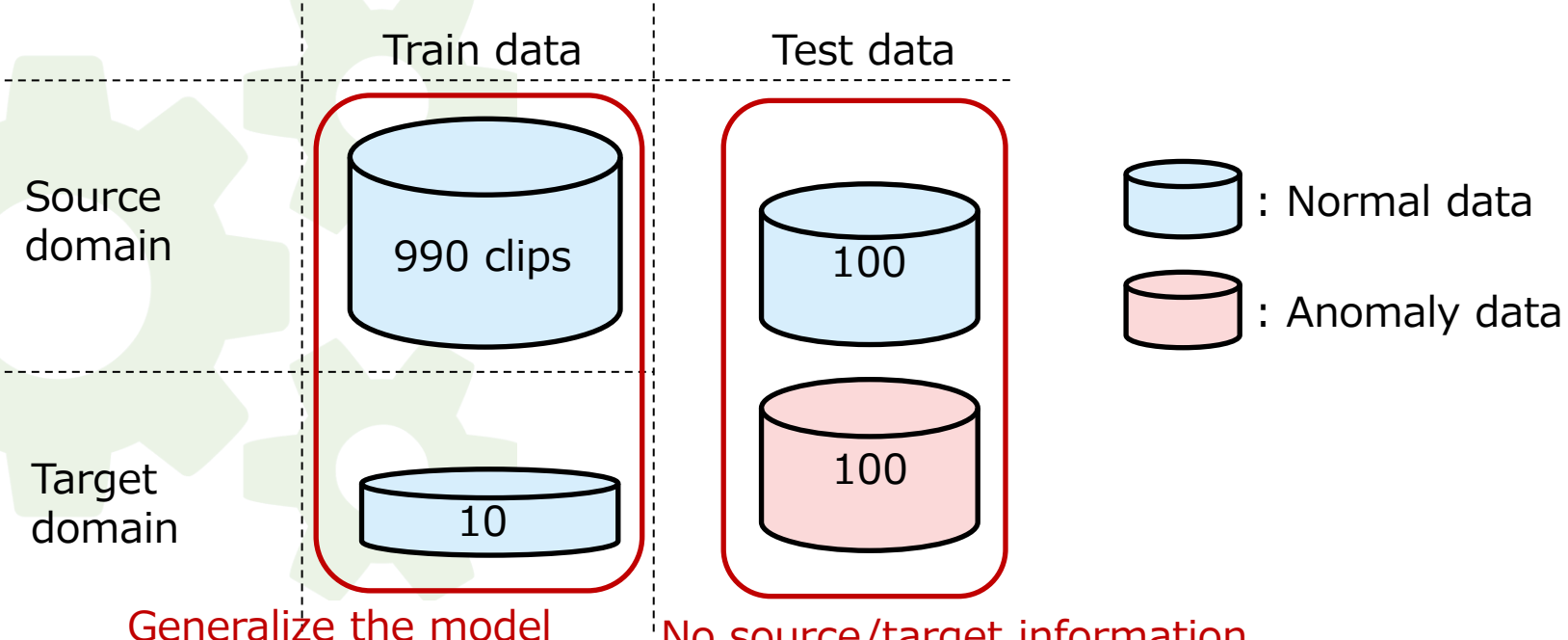



**Adaptation is difficult if shifts are hard to notice**

# Task in 2022: Domain generalization

Monitoring
Task 2

**Can we handle domain shifts by generalizing the model?**



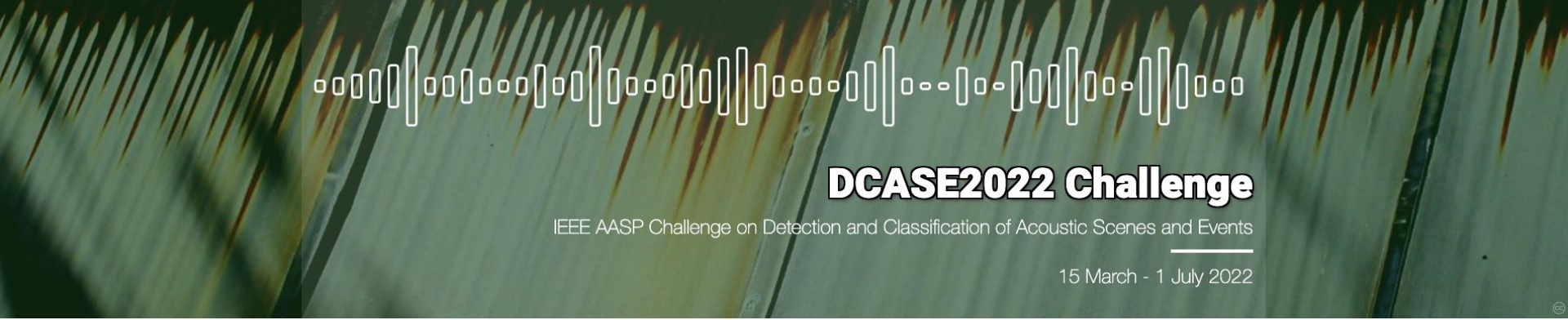

Generalize the model

No source/target information
Detect anomalies with a generalized model that works regardless of the domain

# Task2 related 6 papers will be presented in Workshop

## Enjoy Workshop!

- I. Nejjar+, “DG-MIX: Domain Generalization for Anomalous Sound Detection Based on Self-supervised Learning”
- L. Kai+, “Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Using Temporal Modulation Features on Gammatone Auditory Filterbank”
- K. Dohi+, “MIMII DG: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection for Domain Generalization Task”
- S. Venkatesh+, “Improved Domain Generalization via Disentangled Multi-task Learning in Unsupervised Anomalous Sound Detection”
- K. Mai+, “Explaining the Decisions of Anomalous Sound Detectors”
- Y. Deng+, “Ensemble of Multiple Anomalous Sound Detectors”




# Task 3

## Sound Event Localization and Detection evaluated in real spatial sound scenes

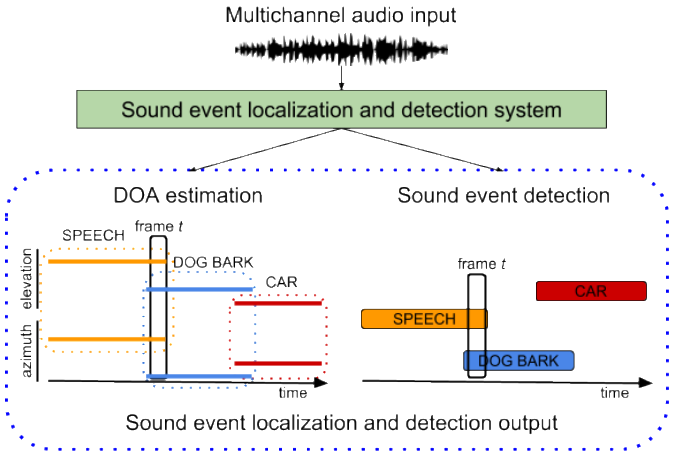

Archontis Politis, Parthasaarathy Sudarsanam, Daniel A. Krause, Sharath Adavanne, Tuomas Virtanen
Kazuki Shimada, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Yuki Mitsufuji

Tampere University

SONY

# Sound Event Localization and Detection

Joint **classification** of sound events, class-wise **activity detection**, and event **localization.**



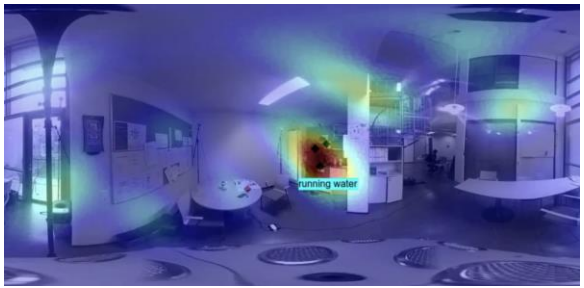
# Dataset

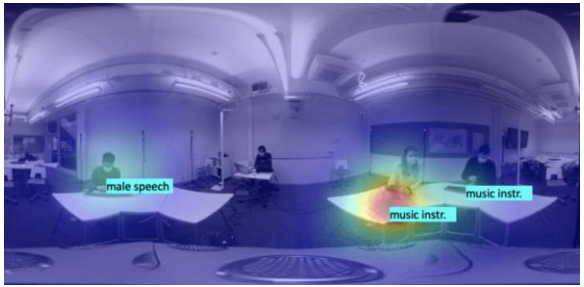
Recordings of naturally acted scenes with multiple human agents in rooms interacting between them and with the environment.

The recordings have been captured with multiple types of sensors and these have been used to annotate them spatiotemporally.

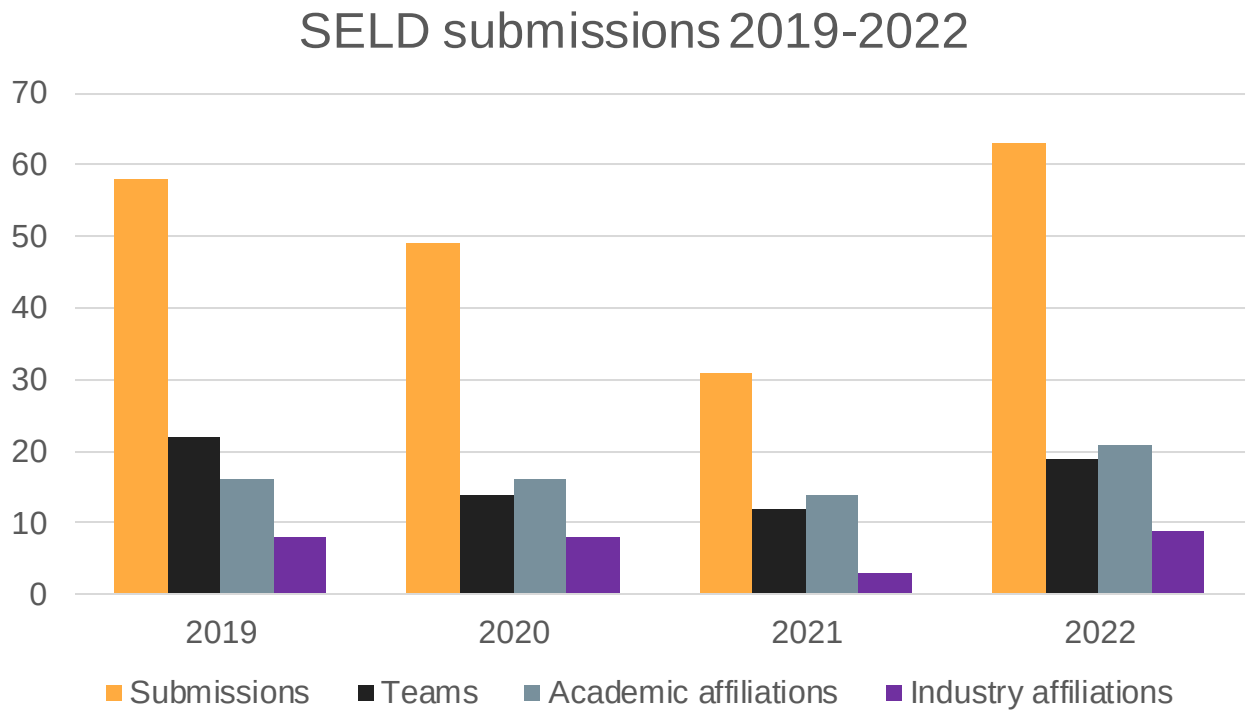
- ~7hrs of recordings captured in Tampere, FI, and Tokyo, JP
- semi improvised scenes of 1-4 actors
- 11 different rooms
- 13 annotated sound classes

- natural composition of classes, class presence, event occurences and co-occurences, and spatial distribution



# Submissions



# Results

| Systems | Format | Method | Features | $ER_{20^\circ}$ | $F_{20^\circ}$ | $LE$ | $LR$ |
|---|---|---|---|---|---|---|---|
| **Du_NERCSLIP** | FOA | CNN, Conformer | mel spectra, intensity vector | 0.35 | 58.3 | 14.6 | 73.7 |
| **Hu_IACAS** | FOA | EINV2, Conformer CNN | mel spectra, intensity vector | 0.39 | 55.8 | 16.2 | 72.4 |
| **Han_KU** | FOA | SE-ResNet34, GRU | mel spectra, intensity vector | 0.37 | 49.7 | 16.5 | 70.7 |
| **Xie_UESTC** | FOA | CRNN | mel spectra, intensity vector | 0.48 | 48.6 | 17.6 | 73.5 |
| **Bai_JLESS** | MIC | CNN, Conformer ensemble | mel spectra, SALSA-Lite | 0.47 | 49.3 | 16.9 | 67.9 |
| **Kang_KT** | BOTH | CRNN, ensemble | mel spectra, intensity vector, magnitude spectra, SALSA-Lite | 0.47 | 45.9 | 15.8 | 59.3 |
| **Ko_SKKU** | FOA | CRNN | magnitude spectra, eigenvector-based intensity vector | 0.49 | 39.9 | 17.3 | 54.6 |
| **Chun_Chosun** | FOA | CRNN, Transformer, ensemble | mel spectra, intensity vector | 0.59 | 31.0 | 19.8 | 50.7 |
| **Scheibler_LINE** | FOA | CNN, Conformer, SSAST, IVA | mel spectra, intensity vector | 0.62 | 30.4 | 16.7 | 49.2 |
| *Guo_XIAOMI | FOA | 3DCNN | mel spectra, intensity vector | 0.60 | 28.2 | 23.8 | 52.1 |
| *Wang_SJTU | BOTH | CRNN, Transformer, ensemble | mel spectra, intensity vector, GCC | 0.67 | 27.0 | 24.4 | 60.3 |
| Baseline | FOA | CRNN | mel spectra, intensity vector | 0.61 | 23.7 | 22.9 | 51.4 |

**These two entries had the same rank in the challenge*

- 12/19 systems did better than the baseline
- Top system *Du_NERCSLIP* had 145% improvement in spatial F-score and 36% improvement in localization error.

# Results: General trends

| Systems | Format | Method | Features | $ER_{20^\circ}$ | $F_{20^\circ}$ | $LE$ | $LR$ |
|---|---|---|---|---|---|---|---|
| **Du_NERCSLIP** | FOA | CNN, Conformer | mel spectra, intensity vector | 0.35 | 58.3 | 14.6 | 73.7 |
| **Hu_IACAS** | FOA | EINV2, Conformer CNN | mel spectra, intensity vector | 0.39 | 55.8 | 16.2 | 72.4 |
| **Han_KU** | FOA | SE-ResNet34, GRU | mel spectra, intensity vector | 0.37 | 49.7 | 16.5 | 70.7 |
| **Xie_UESTC** | FOA | CRNN | mel spectra, intensity vector | 0.48 | 48.6 | 17.6 | 73.5 |
| **Bai_JLESS** | MIC | CNN, Conformer ensemble | mel spectra, SALSA-Lite | 0.47 | 49.3 | 16.9 | 67.9 |
| **Kang_KT** | BOTH | CRNN, ensemble | mel spectra, intensity vector, magnitude spectra, SALSA-Lite | 0.47 | 45.9 | 15.8 | 59.3 |
| **Ko_SKKU** | FOA | CRNN | magnitude spectra, eigenvector-based intensity vector | 0.49 | 39.9 | 17.3 | 54.6 |
| **Chun_Chosun** | FOA | CRNN, Transformer, ensemble | mel spectra, intensity vector | 0.59 | 31.0 | 19.8 | 50.7 |
| **Scheibler_LINE** | FOA | CNN, Conformer, SSAST, IVA | mel spectra, intensity vector | 0.62 | 30.4 | 16.7 | 49.2 |
| *Guo_XIAOMI | FOA | 3DCNN | mel spectra, intensity vector | 0.60 | 28.2 | 23.8 | 52.1 |
| *Wang_SJTU | BOTH | CRNN, Transformer, ensemble | mel spectra, intensity vector, GCC | 0.67 | 27.0 | 24.4 | 60.3 |
| Baseline | FOA | CRNN | mel spectra, intensity vector | 0.61 | 23.7 | 22.9 | 51.4 |

*These two entries had the same rank in the challenge*

- Model: Baseline CRNN is widely used, and many teams upgrade the model with CNN, Transformer, or Conformer.
- Feature: Most teams keep the feature of the baseline, mel spectra and intensity vector, while a few teams take SALSA-Lite or others.
- SELD method: More than half teams follow the baseline to use Multi-ACCDOA while some teams use ACCDOA, EINV2, or others.
- Data augmentation:
  * Multichannel data simulation
  * Audio channel swapping (Rotation)
  * Mixup
  * SpecAugment
  * Band-pass filter
  * Perturbation of gain/frequency/frame/pitch
  * Angle noise to label

# Results: Comments on several systems

| Systems | Format | Method | Features | $ER_{20^\circ}$ | $F_{20^\circ}$ | $LE$ | $LR$ |
|---|---|---|---|---|---|---|---|
| **Du_NERCSLIP** | FOA | CNN, Conformer | mel spectra, intensity vector | 0.35 | 58.3 | 14.6 | 73.7 |
| **Hu_IACAS** | FOA | EINV2, Conformer CNN | mel spectra, intensity vector | 0.39 | 55.8 | 16.2 | 72.4 |
| **Han_KU** | FOA | SE-ResNet34, GRU | mel spectra, intensity vector | 0.37 | 49.7 | 16.5 | 70.7 |
| **Xie_UESTC** | FOA | CRNN | mel spectra, intensity vector | 0.48 | 48.6 | 17.6 | 73.5 |
| **Bai_JLESS** | MIC | CNN, Conformer ensemble | mel spectra, SALSA-Lite | 0.47 | 49.3 | 16.9 | 67.9 |
| **Kang_KT** | BOTH | CRNN, ensemble | mel spectra, intensity vector, magnitude spectra, SALSA-Lite | 0.47 | 45.9 | 15.8 | 59.3 |
| **Ko_SKKU** | FOA | CRNN | magnitude spectra, eigenvector-based intensity vector | 0.49 | 39.9 | 17.3 | 54.6 |
| **Chun_Chosun** | FOA | CRNN, Transformer, ensemble | mel spectra, intensity vector | 0.59 | 31.0 | 19.8 | 50.7 |
| **Scheibler_LINE** | FOA | CNN, Conformer, SSAST, IVA | mel spectra, intensity vector | 0.62 | 30.4 | 16.7 | 49.2 |
| *Guo_XIAOMI | FOA | 3DCNN | mel spectra, intensity vector | 0.60 | 28.2 | 23.8 | 52.1 |
| *Wang_SJTU | BOTH | CRNN, Transformer, ensemble | mel spectra, intensity vector, GCC | 0.67 | 27.0 | 24.4 | 60.3 |
| Baseline | FOA | CRNN | mel spectra, intensity vector | 0.61 | 23.7 | 22.9 | 51.4 |

**These two entries had the same rank in the challenge*

- Top 3 teams used external data and sophisticated data augmentation techniques.
- *Kang_KT* applied AD-PIT to multi-task SELDnet.
- *Ko_SKKU* modified original mixup for ACCDOA.
- *Scheibler_LINE* used IVA to separte sources, while Park_SU used ResUNet.
- *Guo_XIAOMI* proposed a network to consider time alignment.
- Many more SELD-specific innovations proposed (COLOC representation, Spatial Mixup a.o)

# Task 3 @ DCASE Workshop

Thursday 3 Nov

## Session I

1. STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events*
Archontis Politis, Kazuki Shimada, Parthasaarathy Ariyakulam Sudarsanam, Sharath Adavanne, Daniel A. Krause, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Yuki Mitsufuji, Tuomas Virtanen

Spotlight Talk

## Session II

9. Analyzing the effect of equal-angle spatial discretization on sound event localization and detection
Saksham Singh Kushwaha, Iran R. Roman, Juan P. Bello

Poster Spotlight Talk

Friday 4 Nov

## Session III

7. CoLoC: Conditioned Localizer and Classifier for Sound Event Localization and Detection
Sławomir Kapka, Jakub Tkaczuk

Poster Spotlight Talk

## Session IV

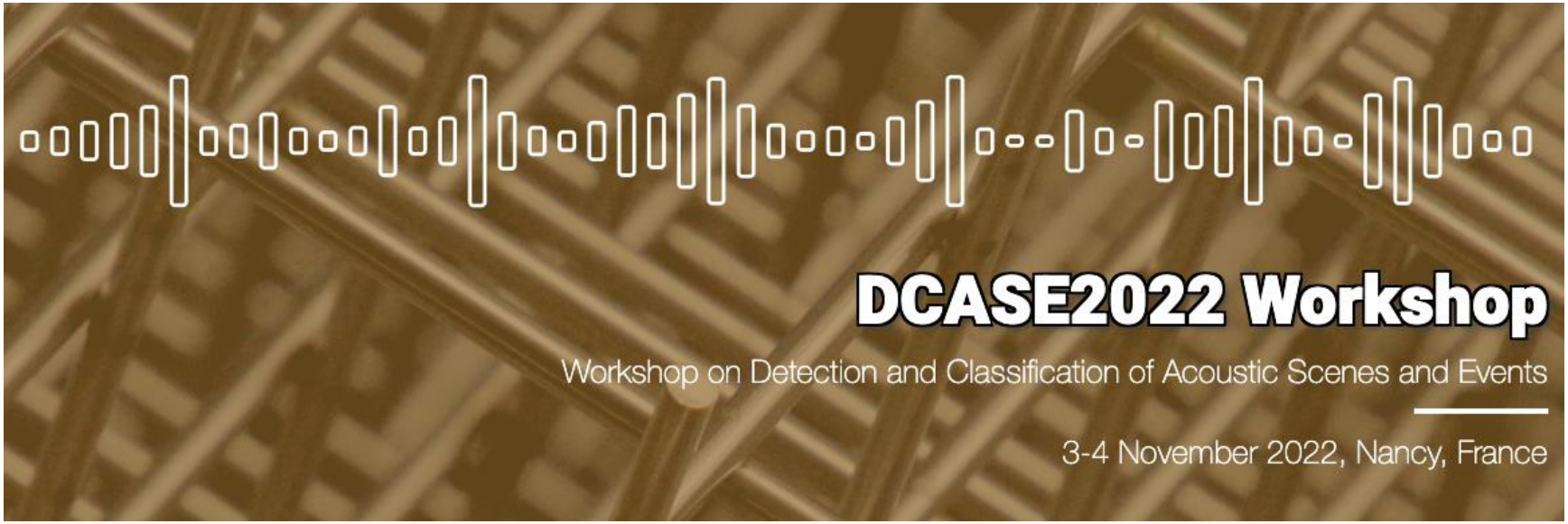
7. SOUND EVENT LOCALIZATION AND DETECTION WITH PRE-TRAINED AUDIO SPECTROGRAM TRANSFORMER AND MULTICHANNEL SEPARATION NETWORK
Robin Scheibler, Tatsuya Komatsu, Yusuke Fujita, Michael Hentschel

Poster Spotlight Talk

11. Sound event localization and detection for real spatial sound scenes: event-independent network and data augmentation chains
Jinbo Hu, Yin Cao, Ming Wu, Qiuqiang Kong, Feiran Yang, Mark D. Plumbley, Jun Yang

Poster Spotlight Talk

# Thank you!


DCASE2022 Workshop
Workshop on Detection and Classification of Acoustic Scenes and Events
3-4 November 2022, Nancy, France

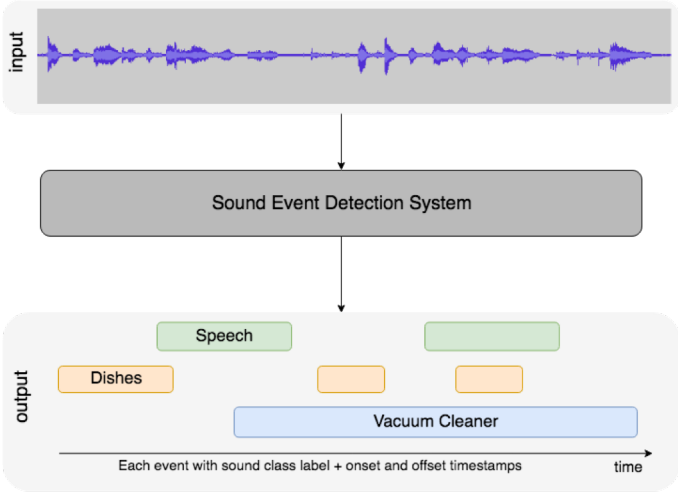
# Task 4 highlights

Romain Serizel[1], Francesca Ronchini[1], Nicolas Turpault[1] ,
Samuele Cornell[2], Eduardo Fonseca[3], Daniel P.W. Ellis[3]

[1]Université de Lorraine, CNRS, Inria, Loria [2]Università Politecnica delle Marche
[3]Google, Inc.

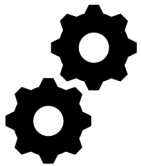
# Sound event detection in domestic environment

- Detecting sound events in 10 s domestic audio files
- **Motivation:** assisted living, surveillance, smart home applications
- **Examples:** Alarm the user when hazardous events appear



# 2022 Novelties

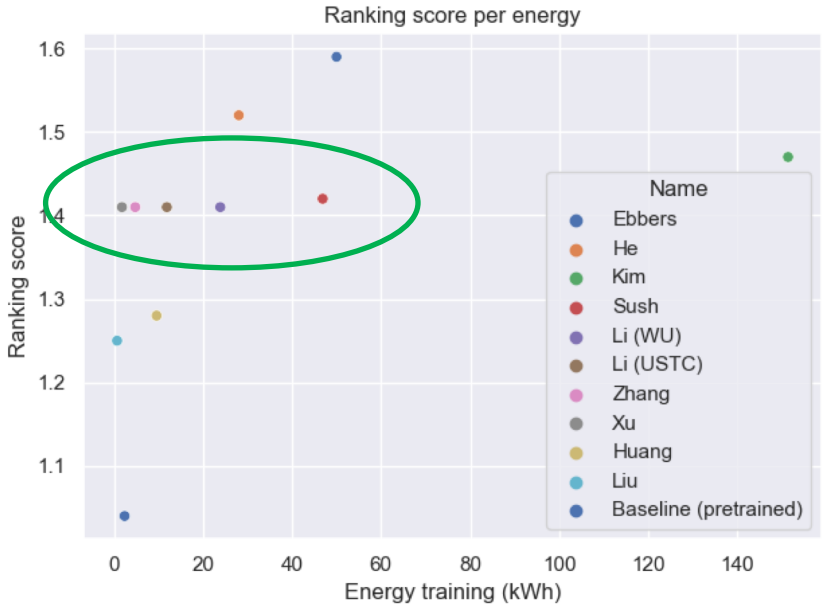
- Allowing some pre-trained models (from Ausioset, AST, ...)
- Introducing the use of Audioset strong
- Suggesting an energy based metric to sensitize on environmental impact



# Submissions

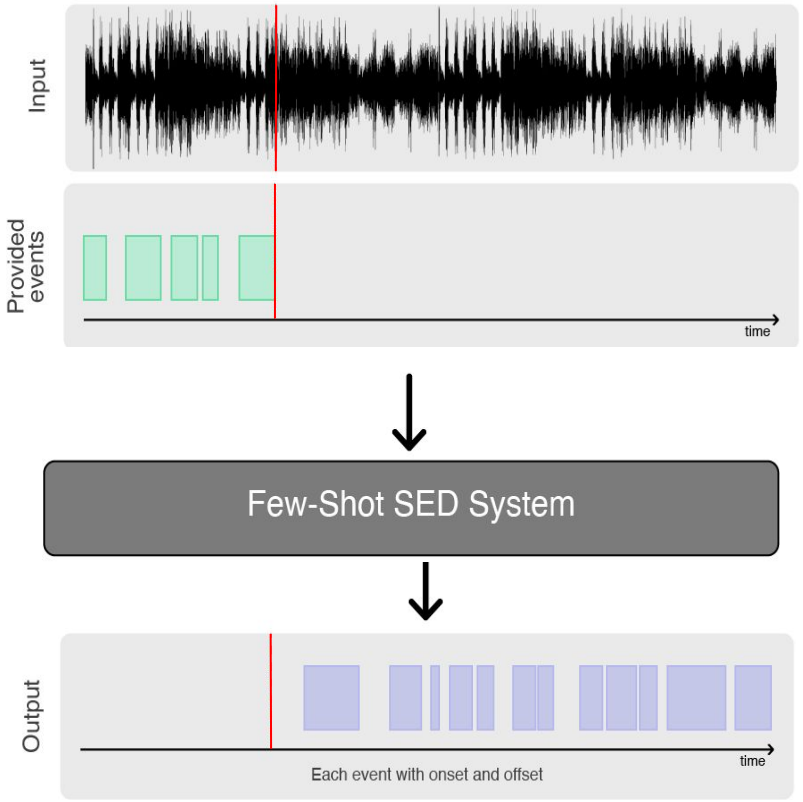
- 99 submissions :
  - >28 Teams / 110 Authors
  - > +25% compared to 2021

- Best system: 1.63 vs 1.4 in 2021:
  - > Do we update the metric including energy ?
  - > Complexity: 1.2M to 780M params (fewly correlated with score)




# FEW-SHOT BIOACOUSTIC EVENT DETECTION

Ines Nolasco, Shubhr Singh, Ester Vidana Vila, Vincent Lostanlen, Ariana Strandburg-Peshkin, Emily Grout, Lisa Gill, Hanna Pamula, Joe Morford, Michael Emmerson, Frants Jensen, Helen Whitehead, Ivan Kiskin, Veronica Morfi and Dan Stowell.

# Datasets








**Training Set**

- Multispecies flight calls (BV)
- Jackdaw calls (JD)
- Western Mediterranean Wetlands Bird calls (WMW)
- Hyena (HT)
- Meerkat (MT)

**Validation Set**

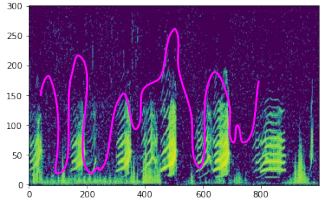
- Poland bird flight calls (PB)
- Meerkat stationary mics (ME)
- HumBug dataset (HB)

**Evaluation Set**

- Manx Shearwaters (MS)
- Chick calls (MGE)
- Biotopia dawn chorus birds (DC)
- Coati (south American mammal) (CT)
- Chernobyl TREE (chiffchaff, cuckoo ) (CHE)
- Dolphin calls (underwater) (QU)

# Baselines

- Template matching
- Prototypical networks




# Evaluation Metric

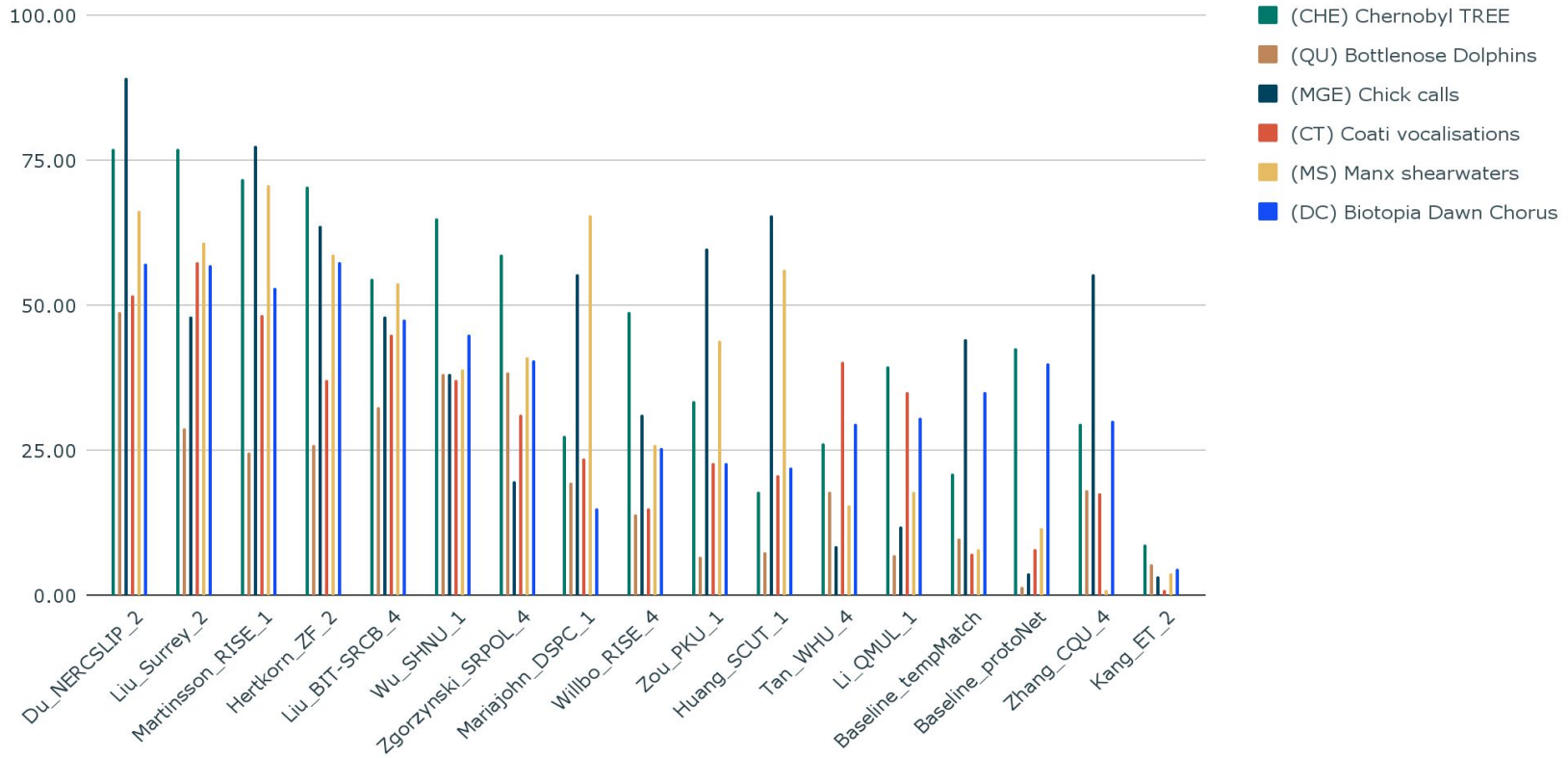
- Intersection over Union
- Bipartite graph matching
- Count TP, FP, FN
- Ranking given by F-score

# Results

15 Teams, 46 systems submitted

| Team code | Eval set: $F$-score % (95% CI) | Val set $F$-score % | Main characteristics |
|---|---|---|---|
| Du_NERCSLIP | 60.22 (59.66-60.70) | 74.4 | CNN+ProtoNet; Frame-level embeddings; PCEN; |
| Liu_Surrey | 48.52 (48.18-48.85) | 50.03 | CNN+ProtoNet; extra data; PCEN+$\triangle MFCC$; various post-process. |
| Martinsson_RISE | 47.97 (47.48-48.40) | 60 | ResNet+ProtoNet; Ensemble(15) based input size; logMel+PCEN |
| Hertkorn_ZF | 44.98 (44.44-45.42) | 61.76 | CNN; Frequency resolution preserving pooling; various post-process |
| Liu_BIT-SRCB | 44.26 (43.85-44.62) | 64.77 | CNN+ProtoNet; Transductive inference |
| Wu_SHNU | 40.93 (40.48-41.30) | 53.88 | ResNet+ProtoNet; Continual-learning; spectrogram |
| Zgorzynski_SRPOL | 33.24 (32.69-33.69) | 57.2 | CNN+Siamese Networks; Emsemble (3) average event-length; |
| Mariajohn_DSPC | 25.66 (25.40-25.91) | 43.89 | CNN+ProtoNet; logMel; augmentation with time-shifting and mirroring |
| Wilbo_RISE | 21.67 (21.32-21.97) | 47.94 | ResNet+ProtoNEt; Semi-supervised; Melspect+PCEN; various post-process |
| Zou_PKU | 19.20 (18.88-19.51) | 51.99 | CNN+protoNet; mutual information loss; time frequency masking + mixup |
| Huang_SCUT | 18.29 (18.01-18.56) | 54.63 | Transductive inference + Adapted central difference convolution |
| Tan_WHU | 17.22 (16.82-17.55) | 54.53 | CNN+ProtoNet pretrained; transductive inference; task adaptive features |
| Li_QMUL | 15.49 (15.16-15.77) | 47.88 | CNN+protoNet; PCEN; time, frequency masking + time warping |
| baseline-TempMatch | 12.35 (11.52-12.75) | 3.37 | Spectrogram Cross correlation |
| baseline-ProtoNet | 5.3 (5.1-5.2) | 28.45 | ResNet+ProtoNet |
| Zhang_CQU | 4.34 (3.74-4.56) | 44.17 | CNN+protoNet; Fine tunning with MIMI; PCEN |
| Kang_ET | 2.82 (2.76-2.87) | - | CNN+ProtoNEt; pretrained ECAPA-TDNN; Fine-tuning; Specaugment |

# F-score on best system per team and evaluation data subsets



Thank you to all the organizers and participants!



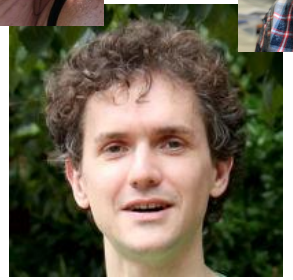










# DCASE2022 Challenge task 6

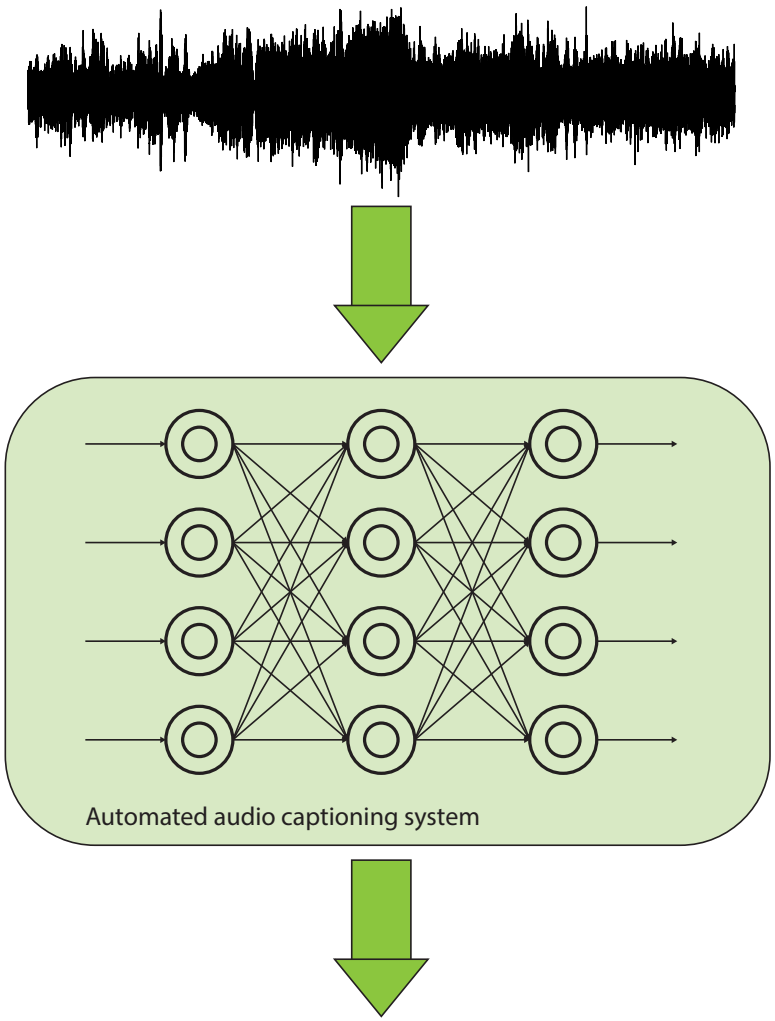
## Subtask A: Automated Audio Captioning

Félix Gontier, Huang Xie, Konstantinos Drossos
Samuel Lipping, Tuomas Virtanen, Romain Serizel

November 2, 2022




# Task presentation




- **Describe acoustic scenes with natural language**
- Clotho dataset
  - Total of 6972 acoustic scenes from 15s to 30s
  - Five reference captions for each scene, 8-20 words

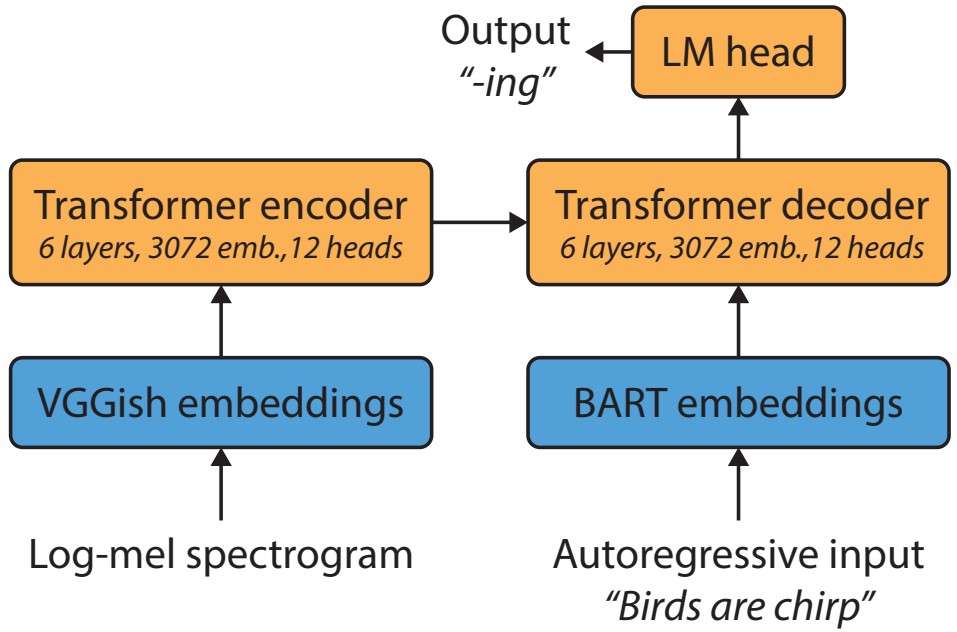
| Subset | Split | # Files | Duration (h) |
|---|---|---|---|
| Development | Training | 3839 | 24.0 |
| | Validation | 1045 | 6.6 |
| | Evaluation | 1045 | 6.5 |
| Testing | - | 1043 | 6.6 |

  - **New analysis subset** based on the Development-Evaluation split for robustness assessment (**microphone**, **scene mixing**, **environmental noise**)

# Task presentation

## Baseline system

- Transformer encoder-decoder
- Conditioned on frozen VGGish embeddings
- Pre-trained BART text embeddings
- Cross-entropy loss with teacher forcing




## Evaluation

- Ranking metric: **SPIDEr** (combination of CIDEr and SPICE)
- Contrastive metrics: **Sentence-BERT** and **FENSE**

# Task results

- Participation: 9 teams, 36 systems, 37 authors
- All teams managed to outperform the baseline
- Top 3 submissions used reinforcement learning to optimize CIDEr

| Author | Audio encoder | Decoder | SPIDEr | Sentence-BERT | FENSE |
|---|---|---|---|---|---|
| Xu | PANNs (CNN14) + BiGRU | Transformer, GRU | **31.9** | 50.8 | 22.7 |
| Zou | PANNs (ResNet38) | LSTM | **31.8** | 49.4 | 22.5 |
| Mei | PANNs (CNN14) | Transformer | **30.9** | 49.6 | 34.8 |
| Primus | PANNs (CNN10) +Transformer | Transformer | **29.6** | 47.6 | 44.0 |
| Kouzelis | PaSST | Transformer | **29.3** | 51.7 | 51.1 |
| Guan | PANNs (CNN10) | Transformer (GraphAC, LocalAFT) | **29.1** | 49.2 | 48.4 |
| Kiciński | PANNs (CNN14) | Transformer | **27.0** | 48.1 | 47.3 |
| Pan | PANNs (CNN10) | Transformer | **25.5** | 47.4 | 45.6 |
| Labbé | PANNs (CNN10) | Transformer | **24.1** | 47.5 | 45.2 |
| Baseline | VGGish + Transformer | Transformer | **22.4** | 45.4 | 44.6 |

# System characteristics

## General trends

- Log-mel spectrogram and pre-trained AudioSet model: **PANNs**
- Transformer architecture for language modeling
- Use of external data: **AudioCaps**, **MACS**
- Data augmentation: **SpecAugment**, **MixUp**
- Fitting evaluation metrics:
  - Validation monitoring
  - **Reinforcement learning on CIDEr**

## Specific approaches

- Recent or novel architectures: GraphAC, LocalAFT
- Text conditioning with keyword retrieval or estimation
- Transfer learning from audio retrieval

# Task 6b
# Language-based Audio Retrieval

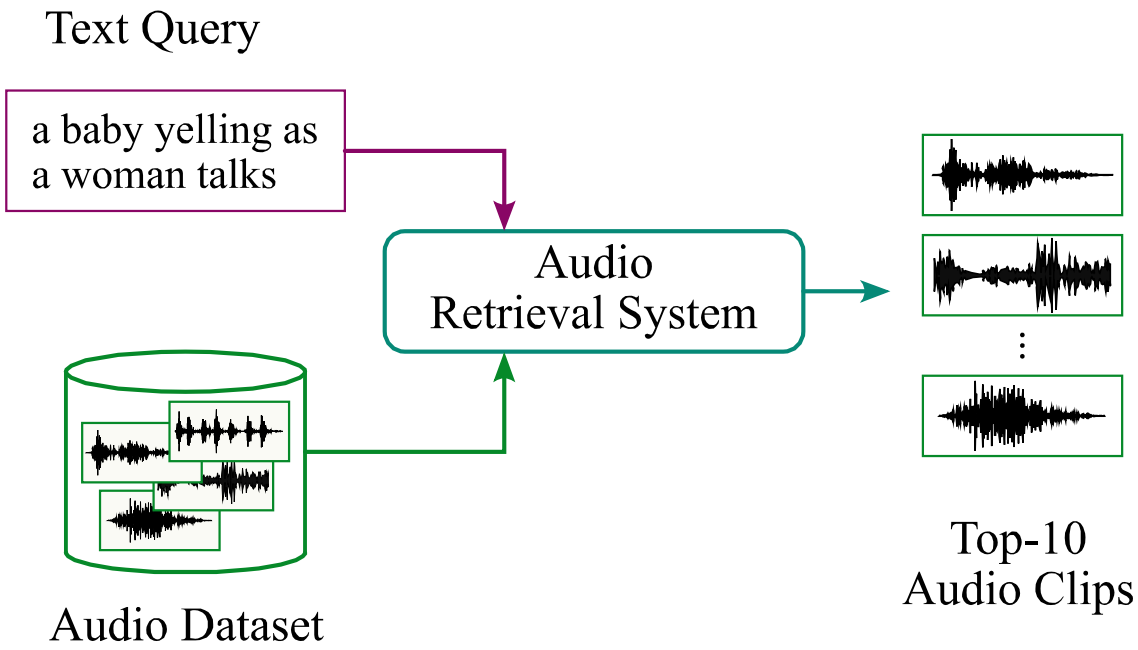
Huang Xie, Félix Gontier, Samuel Lipping,
Konstantinos Drossos, Tuomas Virtanen, Romain Serizel

# Motivation

- The “Big Data Era”
  - Explosive growth of audio data on the web.
  - Great demand for **content-based audio search** tools.
- Search Queries
  - Natural language vs. Keywords & Phrases
  - For example, “a baby yelling as a woman talks followed by a dog barking”.
- Language-based Audio Retrieval
  - Retrieving desired audio with free-form text.
- Real-world Applications
  - Search engines, multimedia databases, human-computer interactions, etc.

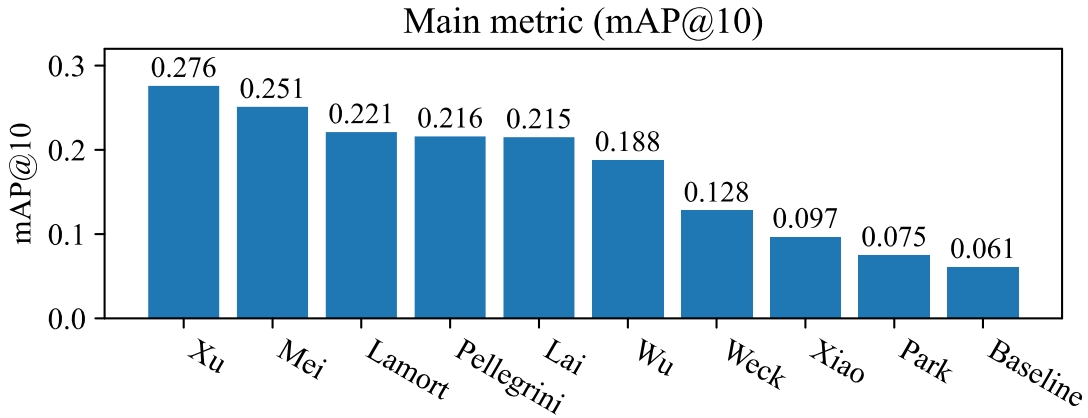
# Task Setup

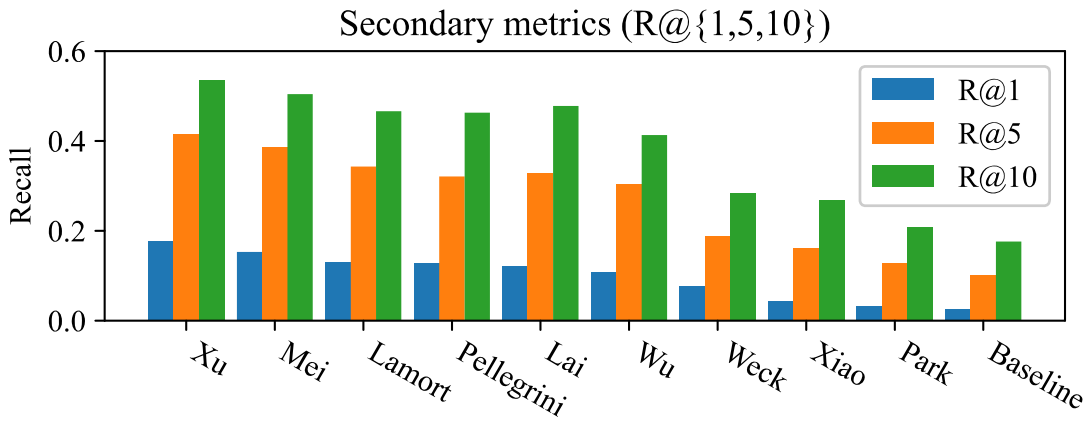
- Task Description
  - To retrieve 10 audio files from a given dataset.
  - To sort them according to their semantic relevance to the query.
- Audio Dataset
  - 1,000 audio-caption pairs.



# Submissions

- Total 31 systems from 9 teams.
  - Teams ranking (mAP@10)
  - Secondary metrics: R@{1, 5, 10}.
- System Summary
  - **Bi-encoder architecture** adopted in all systems.
  - **Pretrained audio** and **NLP models** as encoders.
  - **Contrastive learning** approaches for training.
- Further information present in our poster.



# Thank You!