Master M2 - DataScience

Audio and music information retrieval

Lecture on
Machine Listening, DCASE

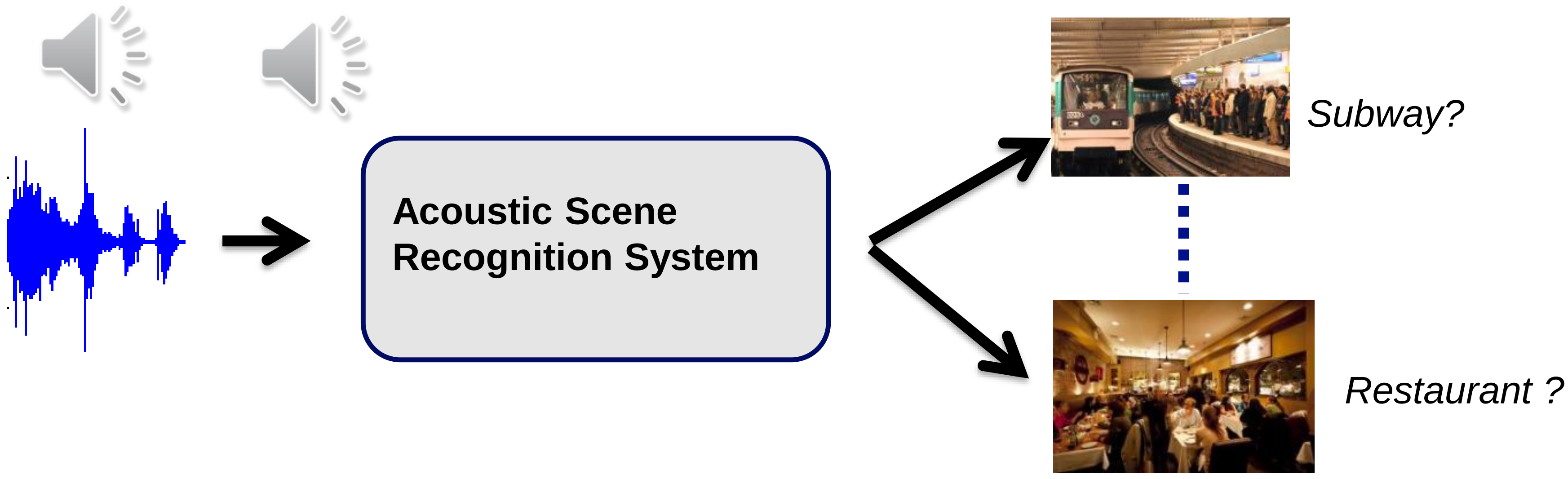
Gaël RICHARD
Télécom Paris
March 2022

Content

- **Introduction**
  - What is Machine listening / audio recognition ?
  - Some applications
- **Machine listening: DCASE**
- **Signal decomposition models**
  - Sinusoidal models
  - Decomposition models (matching pursuit, NMF)
  - Exploitation of such models in scene analysis
- **Audiofingerprint or Music recognition**

Acoustic scene and sound event recognition

- **Acoustic scene recognition:**
  - « associating a semantic label to an audio stream that identifies the environment in which it has been produced »

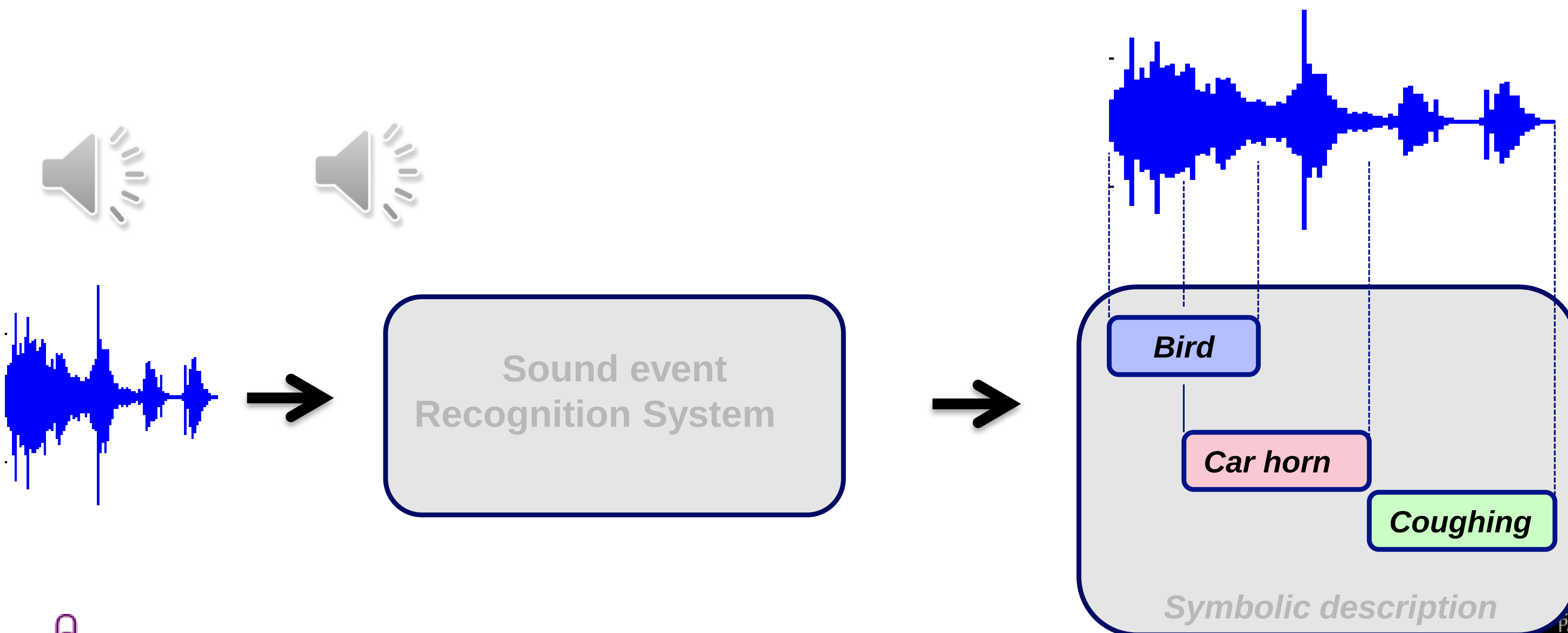


  - Related to CASA (*Computational* Auditory Scene Recognition) and SoundScape cognition (*psychoacoustics*)

D. Barchiesi, D. Giannoulis, D. Stowell and M. Plumbley, « Acoustic Scene Classification », IEEE Signal Process Magazine [16], May 2015

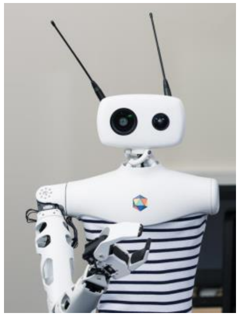
Acoustic scene and sound event recognition

- **Sound event recognition**
  - “aims at transcribing an audio signal into a symbolic description of the corresponding sound events present in an auditory scene”.

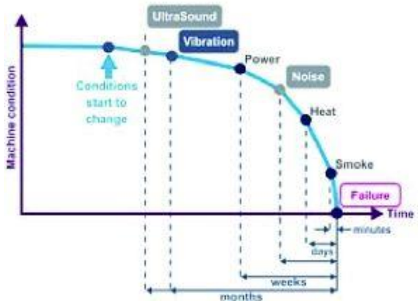


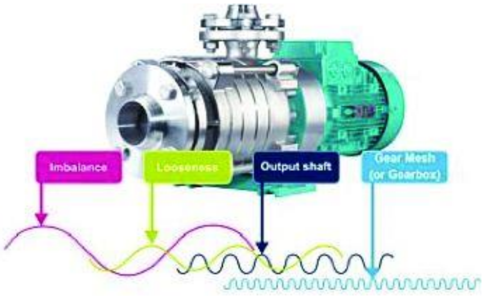
Applications of scene and events recognition

- **Smart hearing aids (Context recognition for adaptive hearing-aids, Robot audition,..)**
- **Security**
- **indexing,**
- **sound retrieval,**
- **predictive maintenance,**
- **bioacoustics,**
- **environment robust speech recognition,**
- **ederly assistance, smart homes**
- **…..**



The Rowe Wildlife Acoustic lab





From ST Microelectronics

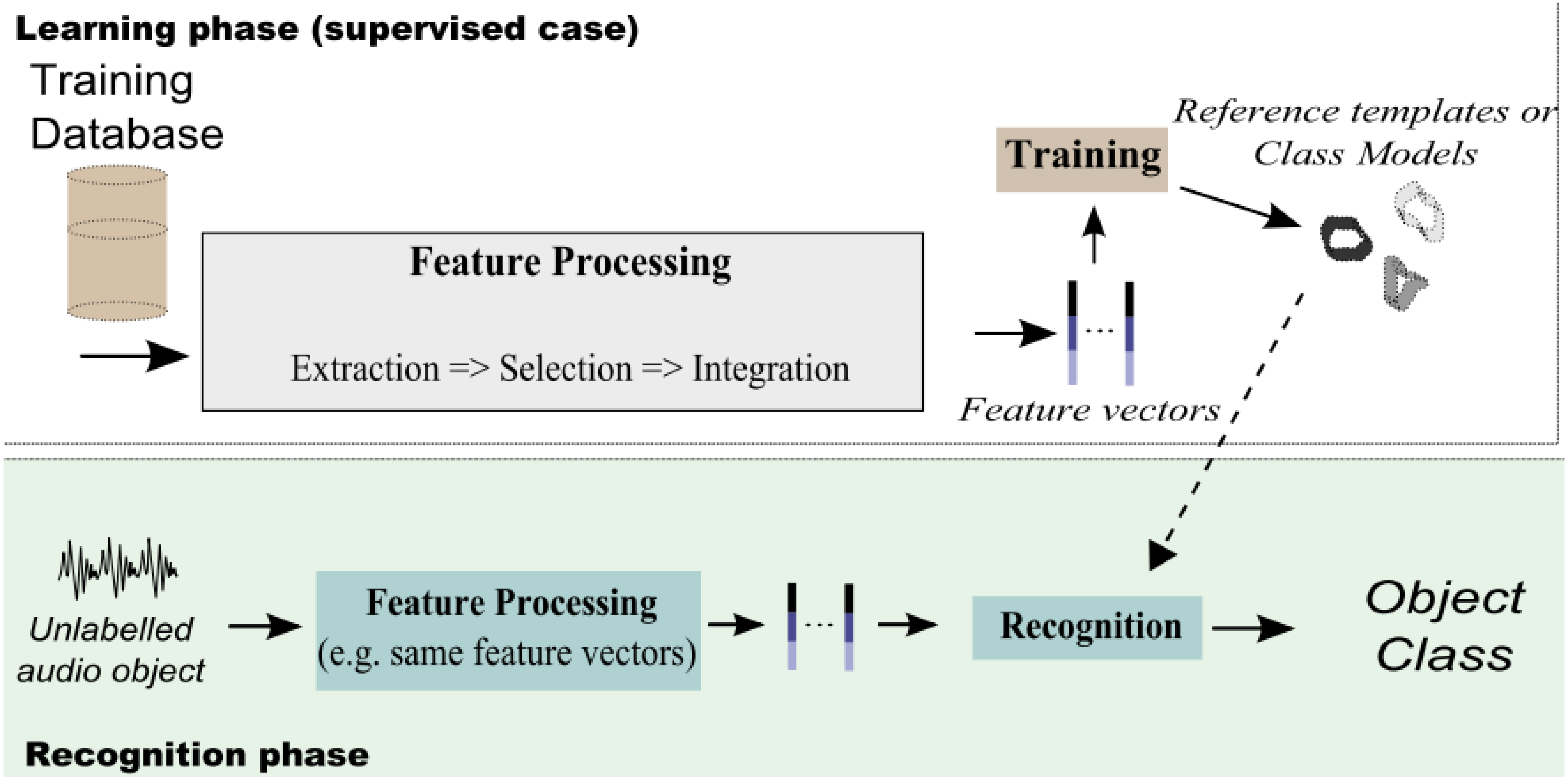
Classification systems

- **Several problems, a similar approach**
  - Speaker identification/recognition
  - Automatic musical genre recognition
  - Automatic music instruments recognition.
  - Acoustic scene recognition
  - Sound samples classification.
  - Sound track labeling (speech, music, special effects etc…).
  - Automatically generated Play list
  - Hit predictor...

Some challenges in Audio listening

- **Huge databases of recordings and sounds**
- **But …. few recordings are precisely annotated**
  - Ex. *label is « bird song » while the bird song last 2s in a 1 mn recording*
- ***The individual sources composing the scene are rarely available.***
  - *Complexifies the learning paradigm*
- ***In Predictive maintenance, the abnormal event is very rare (sometimes never observed)***
  - *Importance of the few-shot learning paradigms, weakly supervised schemes.*

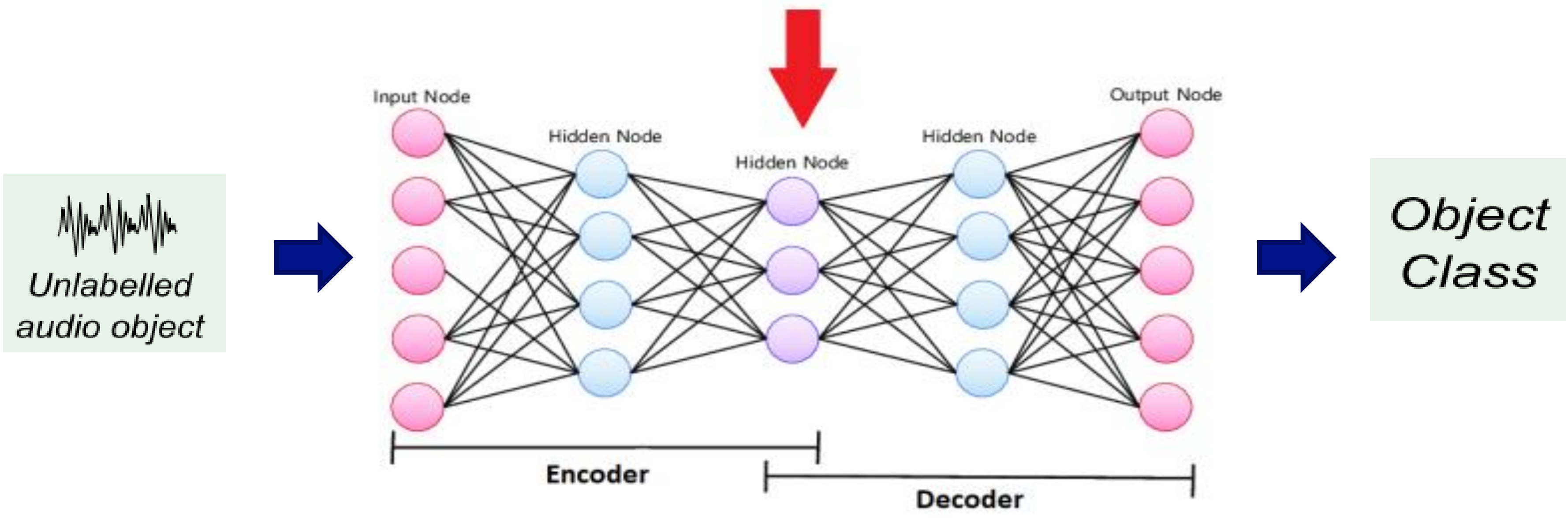
Traditional Classification system



From G. Richard, S. Sundaram, S. Narayanan, "Perceptually-motivated audio indexing and classification", Proc. of the IEEE, 2013

Current trends in audio classification

- **Deep learning now widely adopted**
  - For example under the form of encoder/decoder for representation learning

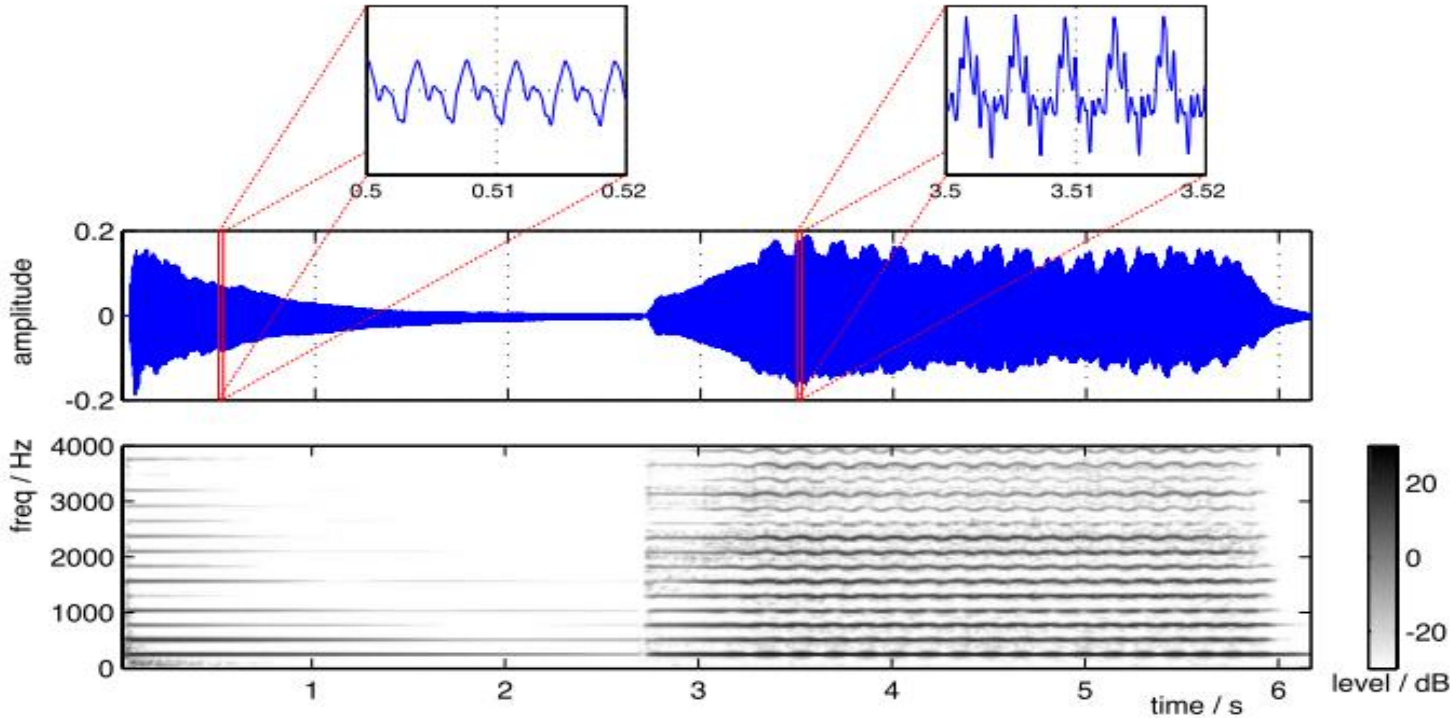


Audio signal representations

- **Example on a music signal: note C (262 Hz) produced by a piano and a violin.**

Temporal Signal

Spectrogram

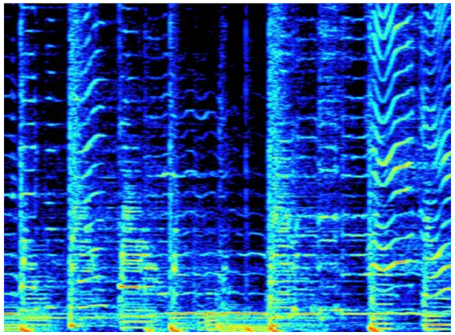


From M. Mueller & al. « Signal Processing for Music Analysis, IEEE Trans. On Selected topics of Signal Processing, oct. 2011

Deep learning for audio

Differences between an image and audio representation

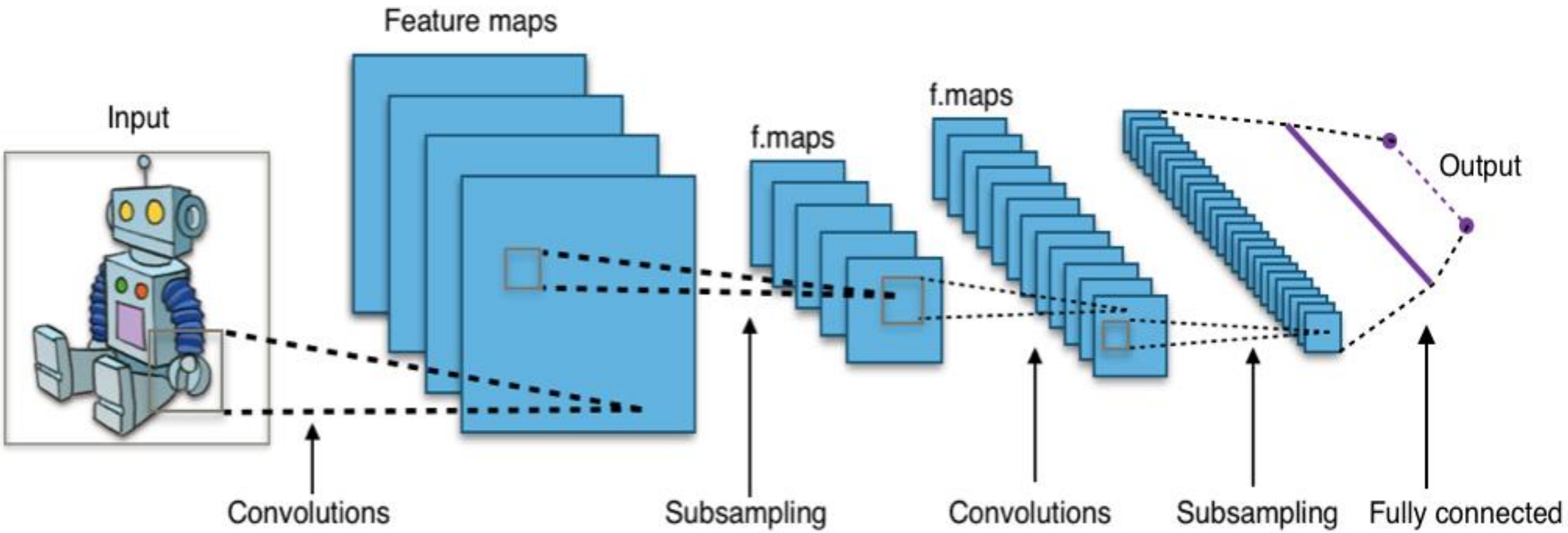
- x and y axes: **same concept** (spatial position).
- Image elements (cat's ear) : **same meaning** independently of their positions over x and y.
- **Neighbouring pixels** : often correlated, often belong to the same object
- **CNN are appropriate :**
  - Hidden neurons locally connected to the input image,
  - Shared parameters between various hidden neurons of a same feature map
  - Max pooling allows spatial invariance



- x and y axes: **different concepts** (time and frequency).
- Spectrogram elements (e.g. a time-frequency area representing a sound source): **same meaning** independently in time **but not over frequency**.
- No invariance over y (even with log-frequency representations): neighboring pixels of a spectrogram are not necessarily correlated since an harmonic sound can be distributed overt he whole frequency in a sparse way
- **CNN not as appropriate than it is for natural images**

G. Peeters, G. Richard, « Deep learning for audio» , Multi-faceted Deep Learning: Models and Data, Edited by Jenny Benois-Pineau, Akka Zemmari, Springer-Verlag, 2021 (to appear)

A typical CNN



From https://en.wikipedia.org/wiki/Convolutional_neural_network

DCASE:Detection and Classification of Acoustic Scenes and Events

- **A recent domain:**
  - A (very) brief historical view of
    - speech recognition
    - Music instrument recognition
    - DCASE

An overview of speech recognition

1952: Analog Digit Recognition, 1 speaker
Features: ZCR in 2 bands
Davis, Biddulph, Balashek

1956: Analog 10 syllable recognition
1 speaker
Features: Filterbank (10 filt.)

1962: Digital vowel Recognition, N speakers
Taxonomy consonant/ vowel
Features: Filterbank (40 filt.)
Schotlz, Bakis

1971: Isolated word Recognition,
Few speakers, DTW
Features: Filterbank
Vintsjuk,…

1975-1985: Rule-based Expert systems
1000 words, few speakers
Features: Many…Filterbanks, LPC, V/U detection, Formant center frequencies, energy, « frication » ….
Decision trees, probabilistic labelling
Woods, Zue, Lamel,…

1980: MFCC
Davis, Mermelstein

1980 - : HMM, GMM,
Baker, Jelinek, Rabiner ,…

2009 - :
Mel spectrogram
DNN
Hilton , Dahl…

An overview of music genre/instrument recognition

1964 - : musical timbre perception
Clarke, Fletcher, Kendall…..

1995 - : Music instrument recognition on isolated notes
Kaminskyj, Martin, Peeters ,..

2000 - : First use of MFCC for music modelling
Logan

2001 - : **Genre recognition**
Multiple musically motivated features + GMM
Tzanetakis,…

2004 - : **Instrument recognition (polyphonic music)**
Multiple timbre features + GMM, SVM, …
Eggink, Essid,…

2007 - : **Instrument recognition : exploiting source separation, dictionary learning**
NMF, Matching pursuit,…
Cont, Kitahara,Heittola, Leveau, Gillet, …

2009 - : instrument recognition
DNN, …
Hamel, Lee …

An overview of Acoustic scene/Events recognition

1980 - : HMM, GMM in speech/speaker recognition,
Baker, Jelinek, Rabiner ,…

1983,1990 Auditory Sound Analysis (Perception/Psychology):
Scheffer, Bregman, …

1993 Computational ASA
(Audio stream segregation)
Use of auditory periphery model
Blackboard model (‘IA)
M. Cook & al.

1997 Acoustic scenes recognition
5 classes of sound
PLP + filter bank features,
RNN or K-NN
Sahwney & al.

1998 Acoustic scene recognition
Use of HMM
Clarksson &al.

2003: Acoustic scene recognition
MFCC+HMM+GMM
Eronen & al.

2005: Event recognition
MFCC+ other feat.
Feature reduction by PCA
GMM
Clavel & al.

From 2009: Scene/Event recognition
More specific methods exploiting sparsity, NMF, image features …
Chu & al, Cauchy & al,…

2014 - :
DNN for acoustic event recognition
Gencoglu & al, ...

DCASE:Detection and Classification of Acoustic Scenes and Events

- **A domain of growing interest: https://dcase.community/**

DCASE2022 WORKSHOP

November 2022, Nancy, France

- A yearly workshop

DCASE2022 CHALLENGE

Tasks

- Low-Complexity Acoustic Scene Classification
- Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques
- Sound Event Localization and Detection Evaluated in Real Spatial Sound Scenes
- Sound Event Detection in Domestic Environments
- Few-shot Bioacoustic Event Detection
- Automated Audio Captioning and Language-Based Audio Retrieval

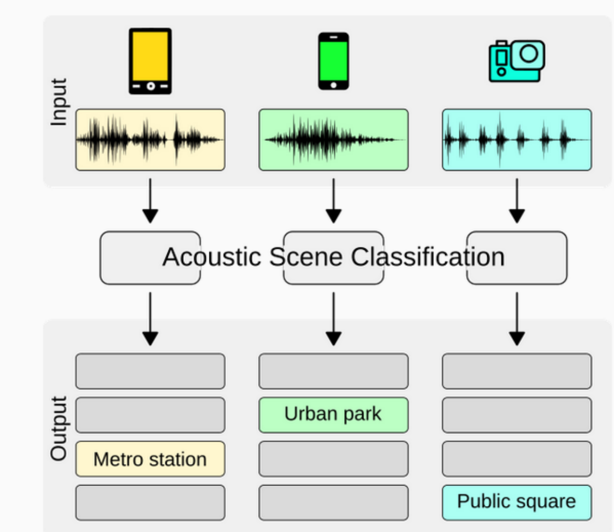
DCASE
Acoustic scene classification (ASC)

- **Goal**: to classify a test recording into one of the provided predefined classes that characterizes the recording environment

- **Two subtasks in the challence DCASE 2021 (1/2)**

Devices A Task 1

ASC with Multiple Devices (10 classes)
Classification of data from multiple devices (real and simulated)



Dataset : TAU Urban Acoustic Scenes 2020 Mobile.
- recordings from 12 cities
- 10 different acoustic scenes
- 4 different devices.

+ synthetic data for 11 mobile devices was created based on the original recordings.

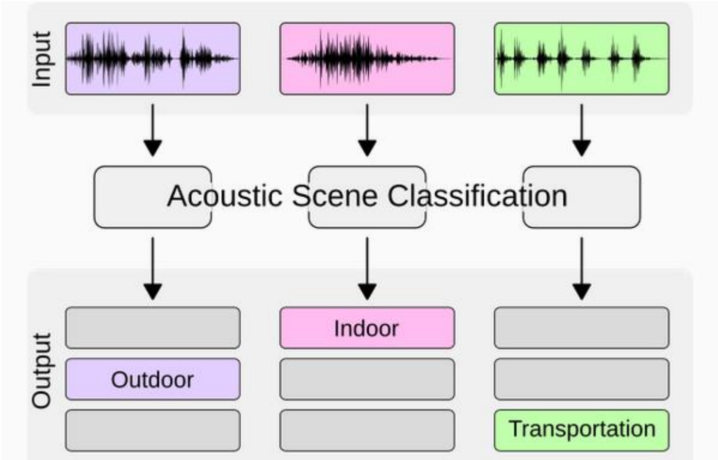
DCASE
Acoustic scene classification (ASC)

- **Goal**: to classify a test recording into one of the provided predefined classes that characterizes the recording environment

- **Two subtasks in the challence DCASE 2021 (2/2)**

Complexity B Task 1

low complexity ASC into three major classes: indoor, outdoor, and transportation.



Dataset : TAU Urban Acoustic Scenes 2020 3Class
- recordings from 12 cities
- 10 different acoustic scenes (*but 3 meta classes*)
- 1 device.

+ synthetic data for 11 mobile devices was created based on the original recordings.

DCASE: *Acoustic scene classification (ASC)*
Task 1.B: low complexity

System complexity requirements

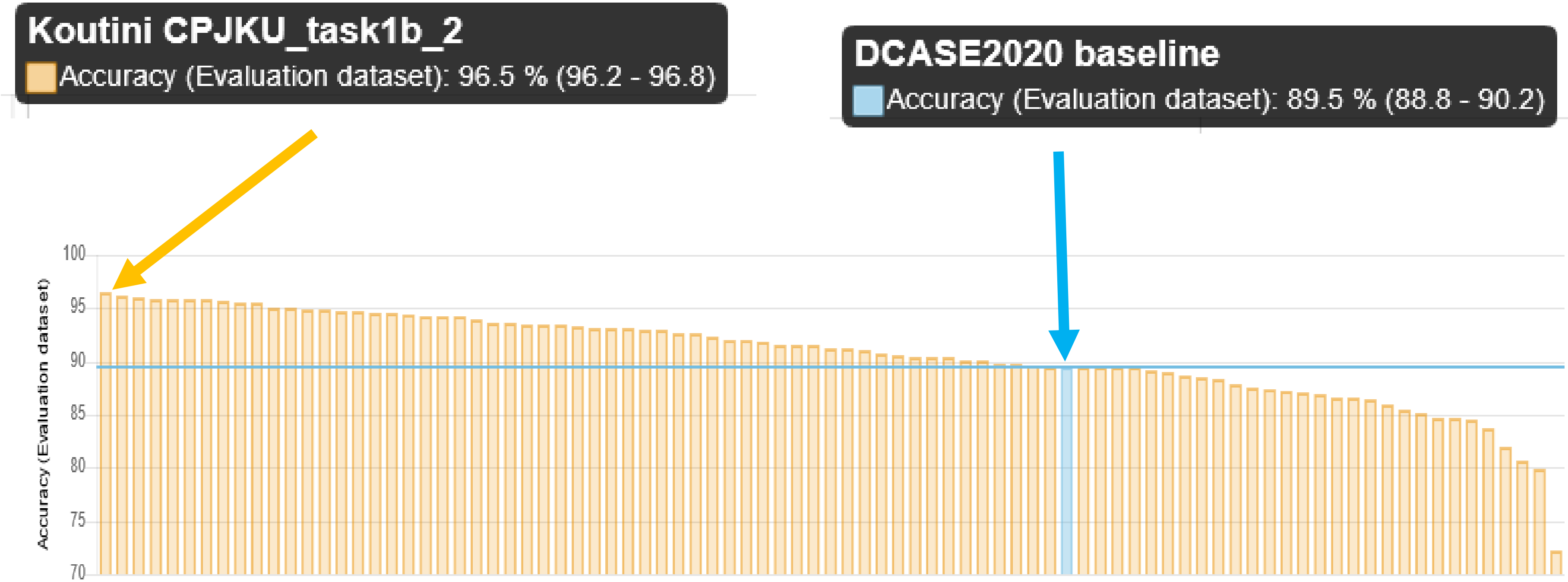
- Classifier complexity limited to :
    - **500KB** size for the **non-zero parameters**
    (excluding layer 1 if it is a feature extraction layer, and batch normalization layers).
    but including the parameters of the network generating the embeddings if used (e.g VGGish, OpenL3, or EdgeL3),

Evaluation:

- **macro-average accuracy (average of the class-wise accuracies)**

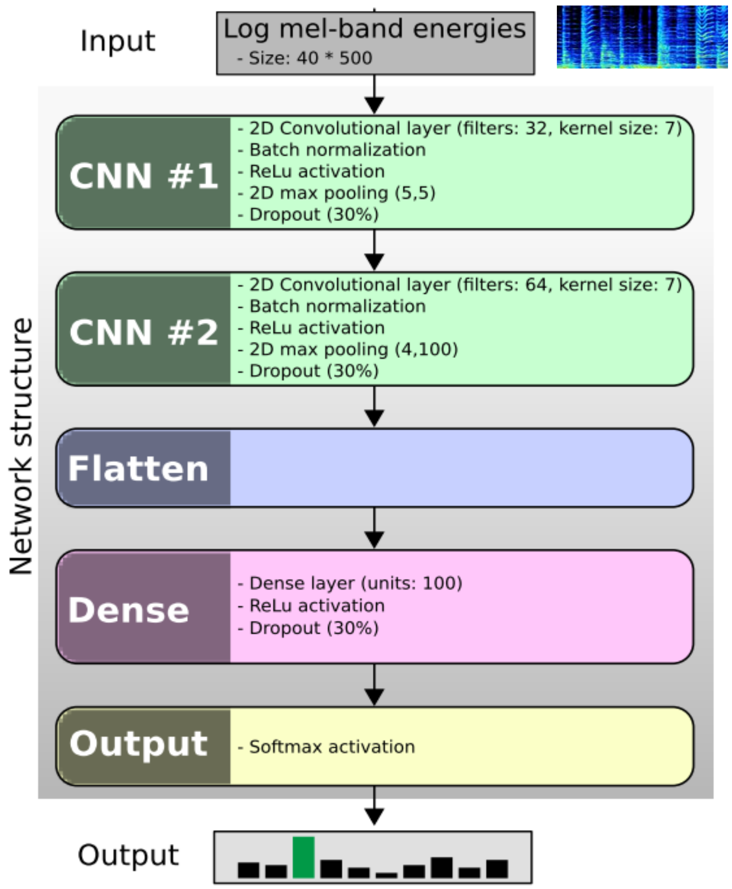
DCASE: *Acoustic scene classification (ASC) Task 1.B: low complexity*

- **Performances (DCASE 2020)**



DCASE: *Task 1.B: low complexity Baseline 2020 system*

- **Parameters (model size = 450 kB)**
- **Audio features:**
  - Log mel-band energies (40 bands), analysis frame 40 ms (50% hop size)
- **Neural network:**
  - Input shape: 40 * 500 (10 seconds)
  - Architecture:
    - CNN layer #1
      - 2D Convolutional layer (filters: 32, kernel size: 7) + Batch normalization + ReLu activation
      - 2D max pooling (pool size: (5, 5)) + Dropout (rate: 30%)
    - CNN layer #2
      - 2D Convolutional layer (filters: 64, kernel size: 7) + Batch normalization + ReLu activation
      - 2D max pooling (pool size: (4, 100)) + Dropout (rate: 30%)
    - Flatten
    - Dense layer #1
      - Dense layer (units: 100, activation: ReLu)
      - Dropout (rate: 30%)
    - Output layer (activation: softmax)
  - Learning: 200 epochs (batch size 16), data shuffling between epochs
  - Optimizer: Adam (learning rate 0.001)



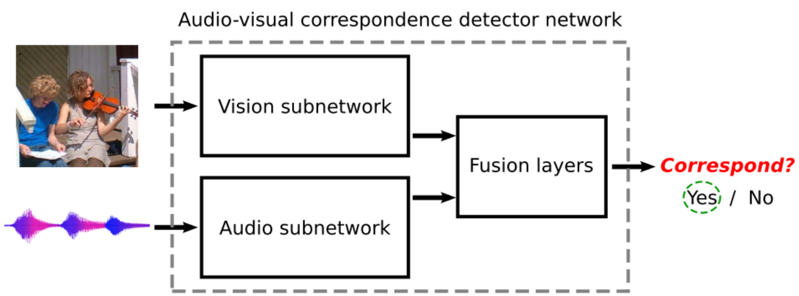
A. Mesaros, T. Heittola, and T. Virtanen. *A multi-device dataset for urban acoustic scene classification.* In Proc. of DCASE 2018.
T. Heittola & al. *Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions.* In Proc. of the DCASE 2020 Workshop

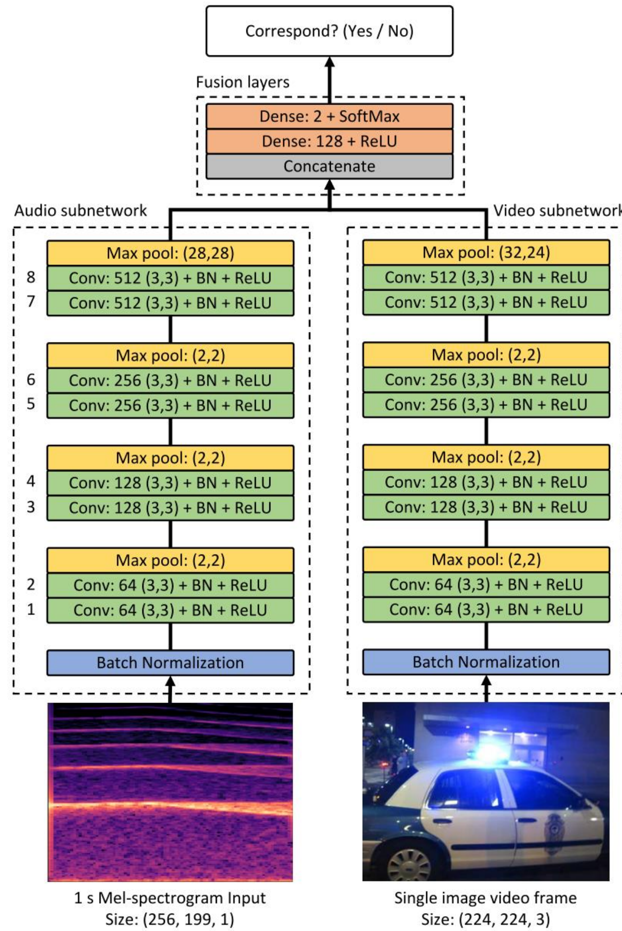
Comparasion with other baselines

System	Accuracy	Log loss	Audio embedding	Acoustic model	Total size
DCASE2020 Task 1 Baseline, Subtask A *OpenL3 + MLP (2 layers, 512 and 128 units)*	89.8 % (± 0.3)	0.266 (± 0.006)	17.87 MB	145.2 KB	19.12 MB
Modified DCASE2020 Task 1 Baseline, Subtask A *EdgeL3 + MLP (2 layers, 64 units each)*	88.9 % (± 0.3)	0.298 (± 0.003)	840.6 KB	145.2 KB	985.8 KB
DCASE2020 Task 1 Baseline, Subtask B *Log mel-band energies + CNN (2 CNN layers and 1 fully-connected)*	87.3 % (± 0.7)	0.437 (± 0.045)	-	450.1 KB	450 KB

DCASE: Audio Scene classification

DCASE2020 Task 1 Baseline, Subtask A *OpenL3 + MLP (2 layers, 512 and 128 units)*





R. Arandjelovi´c and A. Zisserman, *“Look, listen and learn,”* in IEEE ICCV, 2017, pp. 609–617.
S. Kumari, D. Roy, M. Cartwright, J. P. Bello, and A. Arora. *EdgeI^3: compressing l^3-net for mote scale urban noise monitoring.* In 2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW),

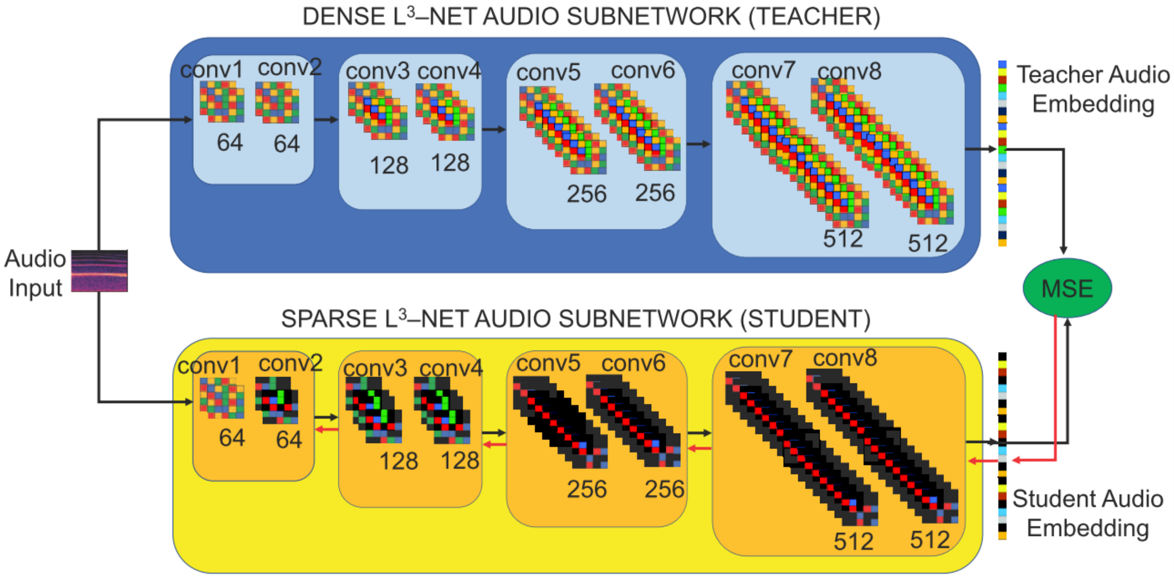
DCASE: Audio Scene classification

Modified DCASE2020 Task 1 Baseline, Subtask A

EdgeL3 + MLP (2 layers, 64 units each)

- **Sparsity**
  - Teacher-student
  - Different level of sparsit

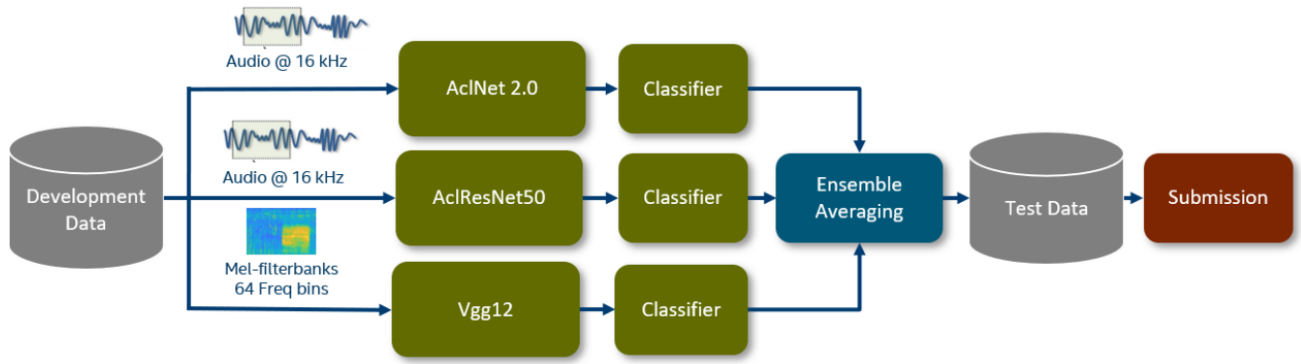
  For each layer



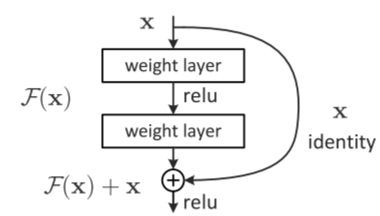
S. Kumari, D. Roy, M. Cartwright, J. P. Bello, and A. Arora. *Edgel^3: compressing l^3-net for mote scale urban noise monitoring.* In 2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW),

Acoustic scene recognition:
How to improve ?

- **Some trends and tricks**
  - Use ensemble techniques



  - Use Data augmentation (*mix up, random cropping, channel confusion, Spectrum augmentation, spectrum correction, reverberation, pitch shift, speed change, random noise, mix audios, ...*)



  - Use large networks (> 17 layers), Resnets
  - Use signal or audio models (NMF, ..)

P. Lopez & al. "Ensemble of Convolutional Neural Networks", in DCASE 2020 Acoustic Scene Classification Challenge

Acoustic scene recognition:
Why using signal or perceptual models

- **Using perceptual models**
  - Example: Mel specrogram, MFCC, CQT,..
  - The classifier does not learn what is not audible

- **Using signal models**
  - Example: Harmonic + noise, Source filter, NMF, …
  - *e.g The classifier does not learn what is not typical of an audio signal*

- **With such models**
  - The training may be simpler (faster convergence)
  - The need for data may be far less (frugality in data)
  - The need for complex architecture may be lower (frugality in computing power)

Non-negative Matrix Factorization (NMF)

- Use of non-supervised decomposition methods (for example Non-Negative Factorization methods or NMF)

- **Principle of NMF :**

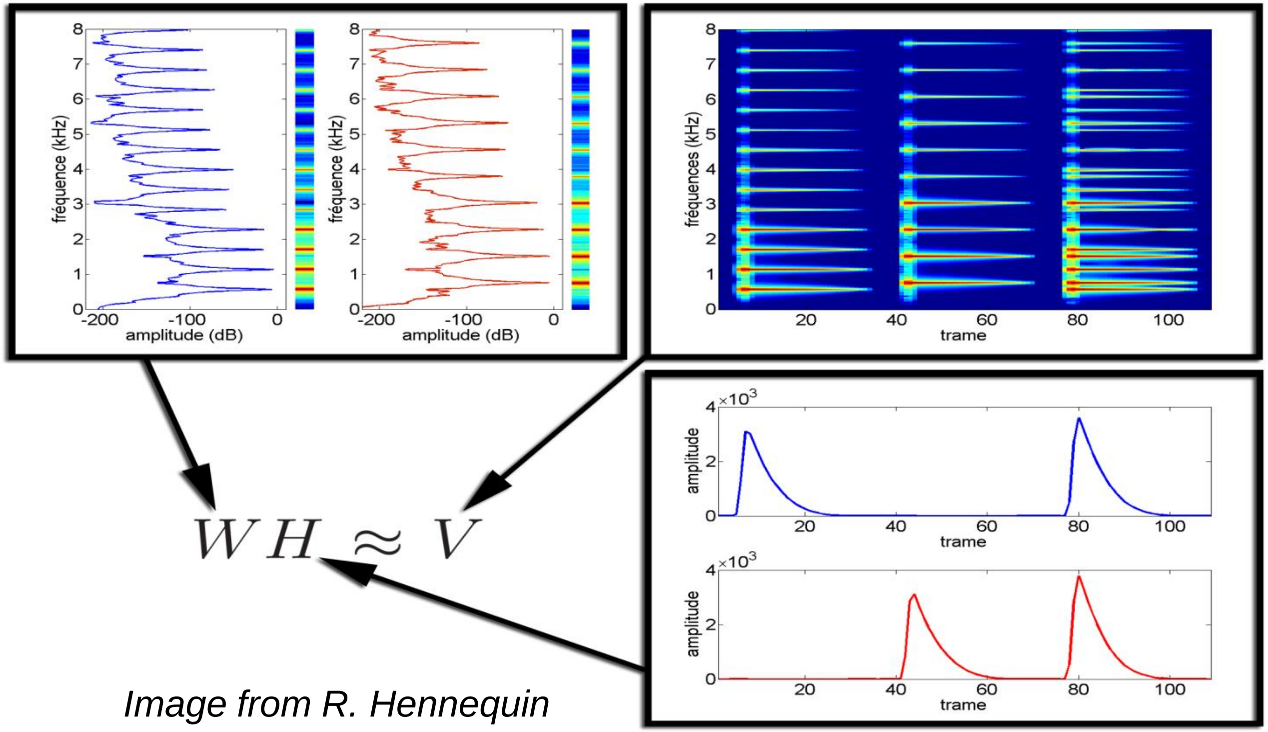
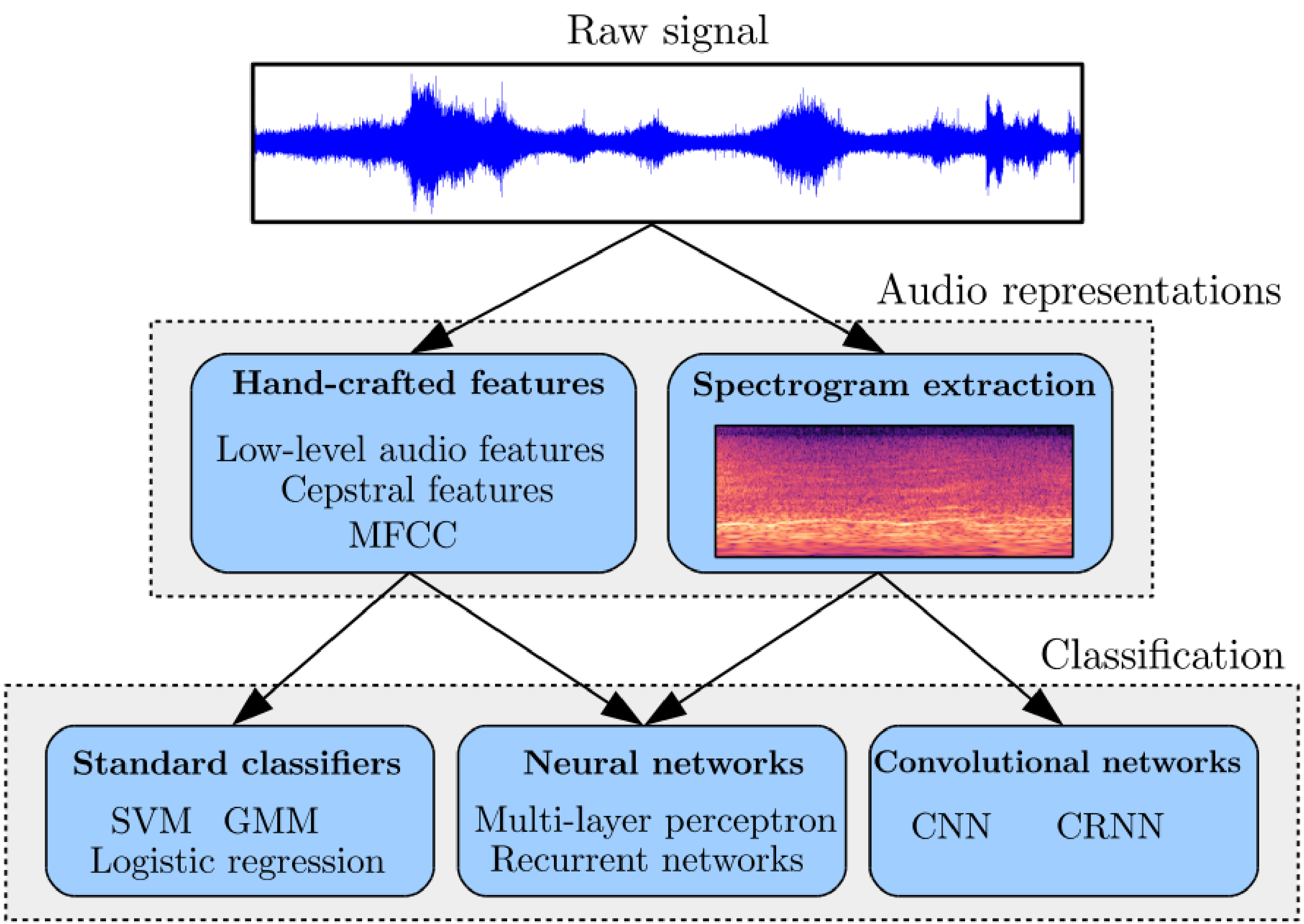
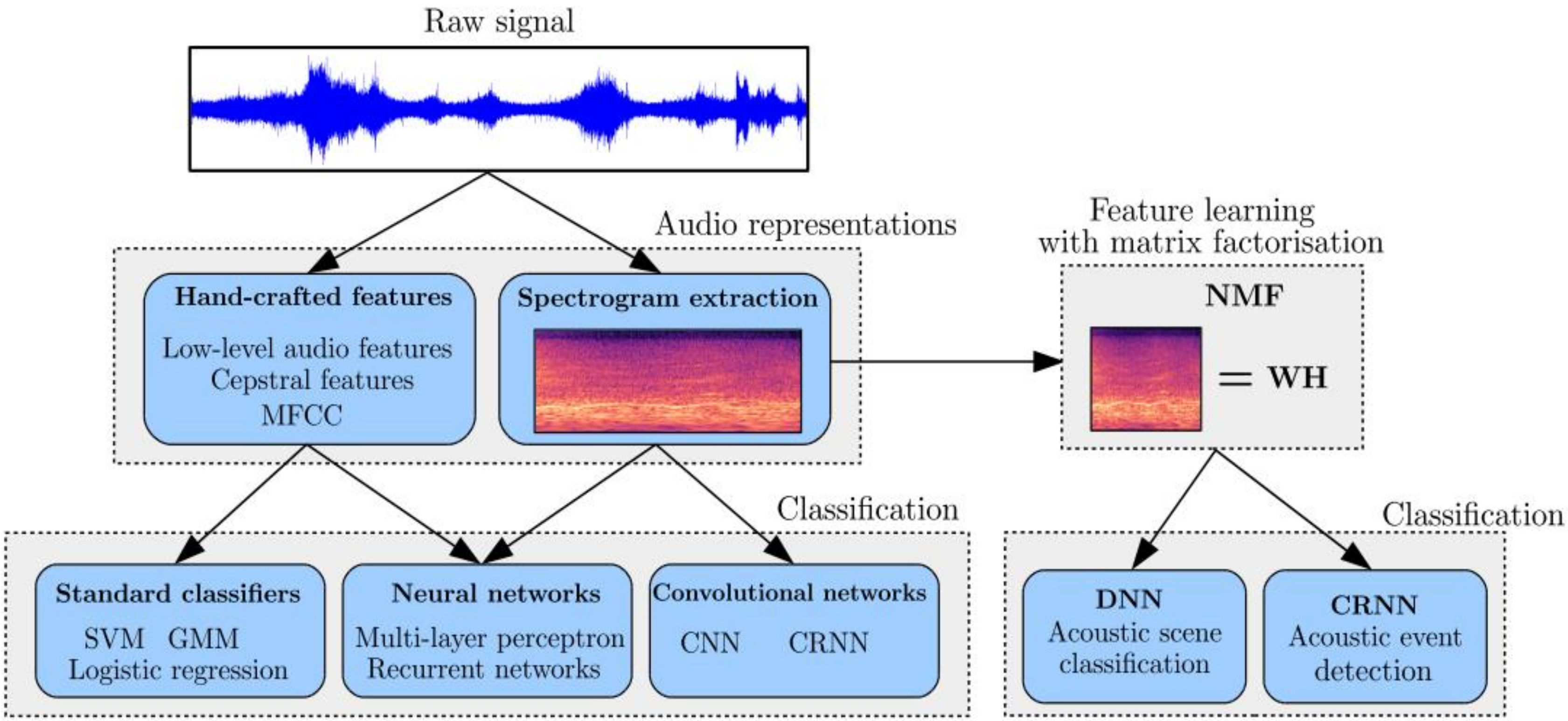


Image from R. Hennequin

Recent approaches for Audio scene and event recognition



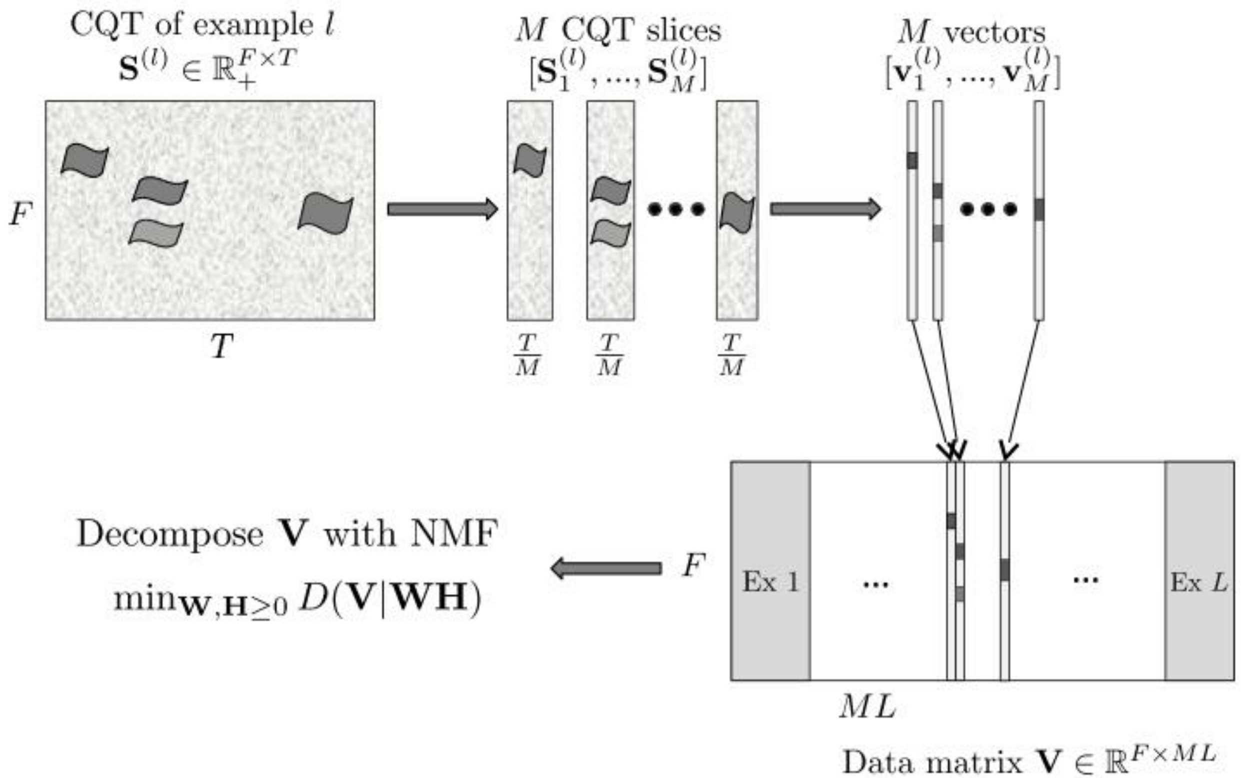
A recent framework for Audio scene and event recognition (Bisot & al. 2017)



V. Bisot & al., "Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, (2017),
V. Bisot & al., Leveraging deep neural networks with nonnegative representations for improved environmental sound classification IEEE International Workshop on Machine Learning for Signal Processing MLSP, Sep 2017, Tokyo,

Example for scene classification

From time-frequency representations to dictionary learning

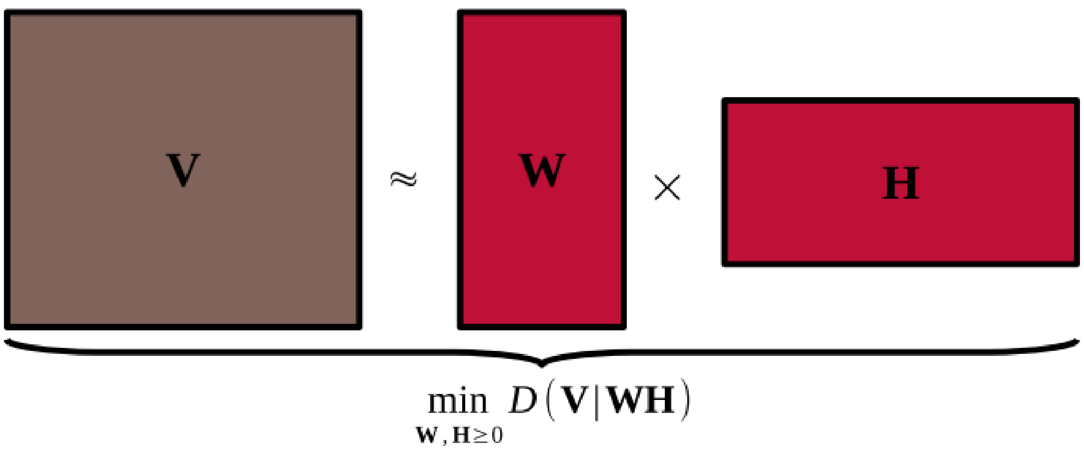


Unsupervised NMF for acoustic scene recognition

Nonnegative matrix factorization

$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH})$ with $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$

Dictionary learning with NMF



Unsupervised NMF for acoustic scene recognition

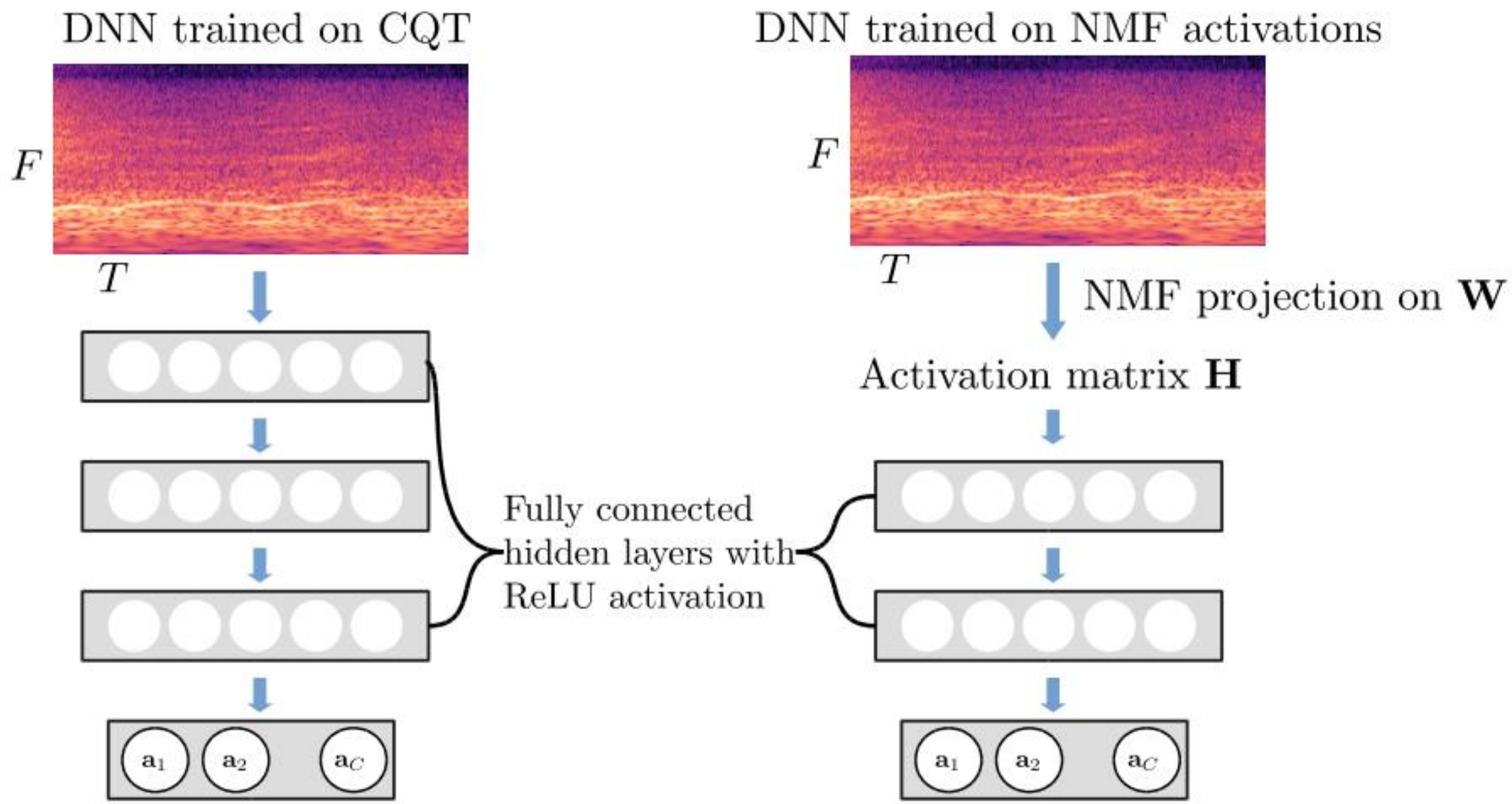
Nonnegative matrix factorization

$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH})$ with $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$

Feature extraction $\longrightarrow$ project on learned dictionary

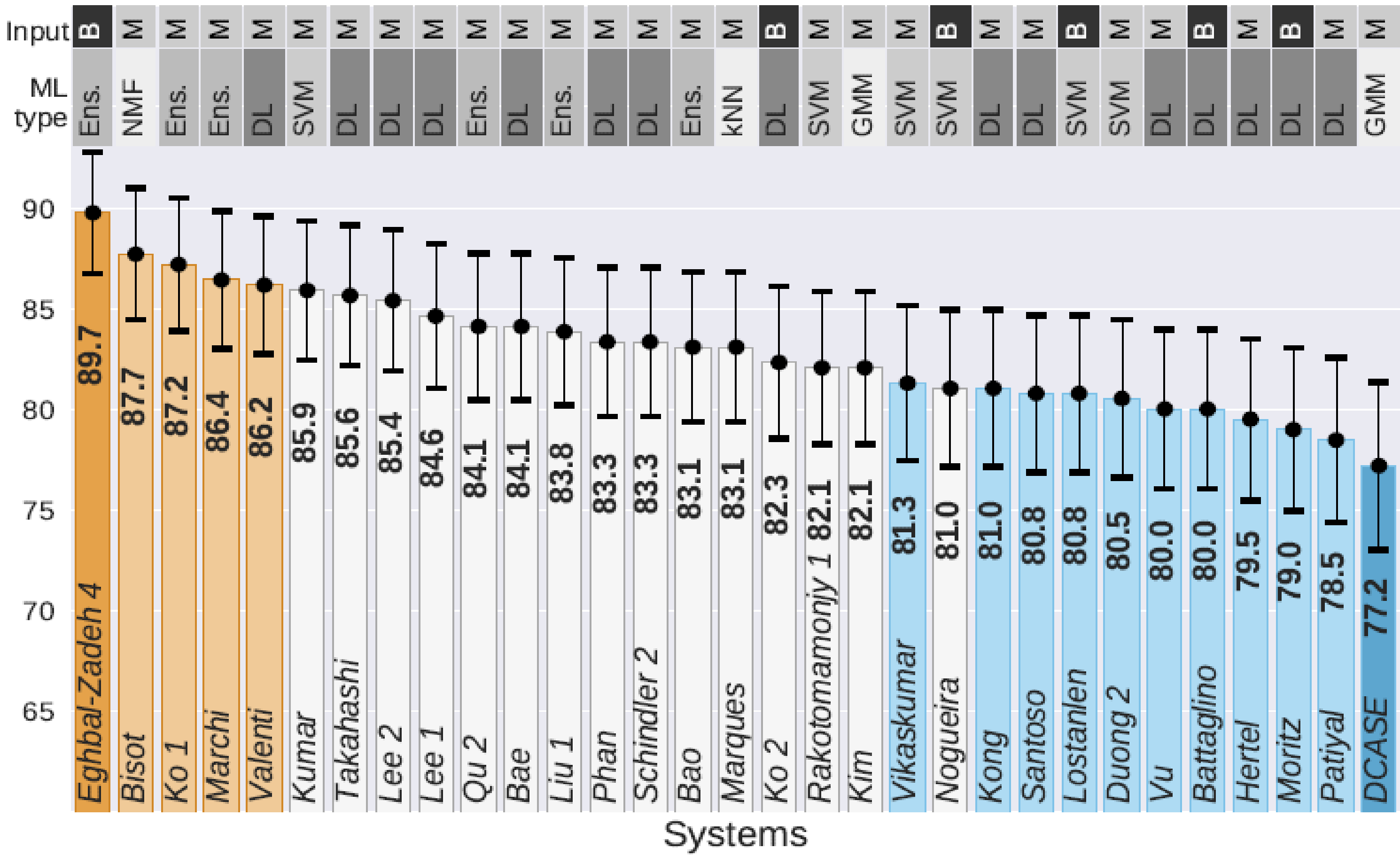
$$\underbrace{\boxed{\mathbf{V}} \approx \boxed{\mathbf{W}} \times \boxed{\mathbf{H}}}_{\min_{\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH})}$$

Example with DNN: acoustic scene recognition



V. Bisot & al., "Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, (2017),
V. Bisot & al., Leveraging deep neural networks with nonnegative representations for improved environmental s[…] classification IEEE International Workshop on Machine Learning for Signal Processing MLSP, Sep 2017, Toky[…]

Typical performances of Acoustic scene recognition (challenge DCASE 2016)



- *A Mesaros & al. Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 challenge IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (2), 379-393*

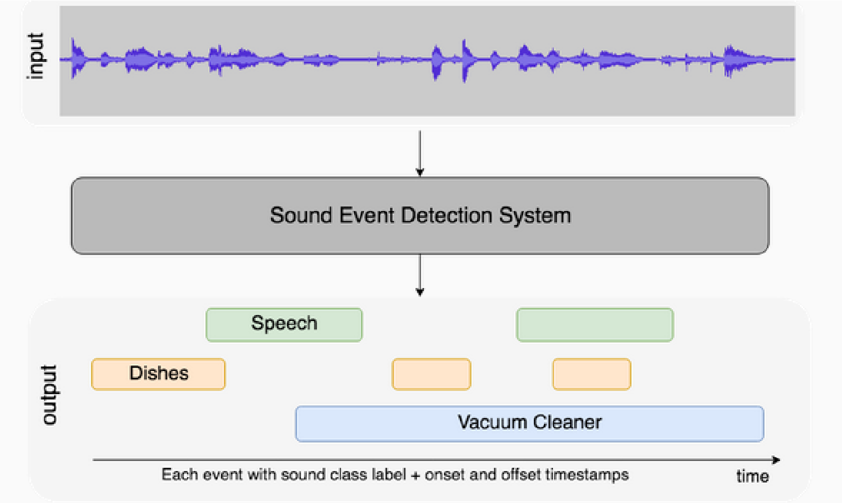
DCASE: Sound Event Detection and Separation in Domestic Environments

- **Goal**: the detection of sound events with their time localization using weakly labeled data (without timestamps).

- **Two subtasks in the challence DCASE 2021 (1/2)**

Domestic
Task 4

to provide the event class with event time localization given that multiple events can be present in an audio recording



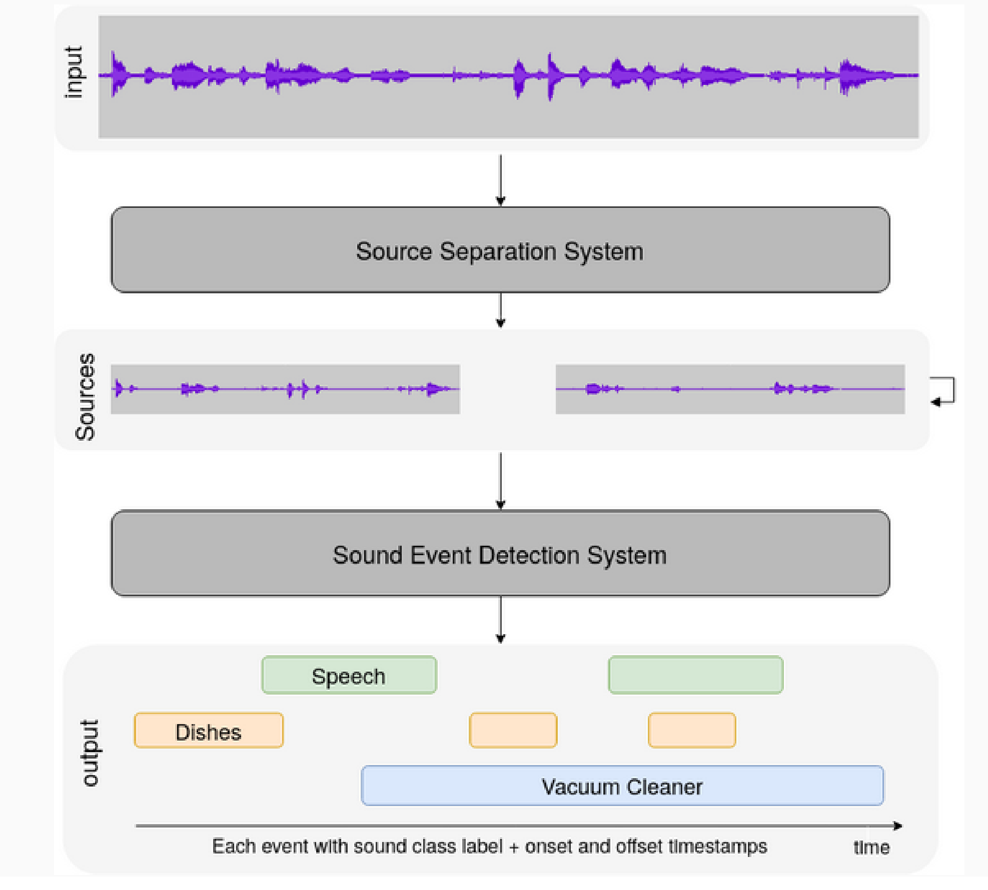
Dataset : many datasets *(see next slide)*

- DESED
- SINS
- TUT Acoustic scenes 2017
- FUSS
- FSD50K
- YFCC100M

DCASE: Sound Event Detection and Separation in Domestic Environments

- **Goal**: the detection of sound events with their time localization using weakly labeled data (without timestamps).

- **Possibility to use source separation**

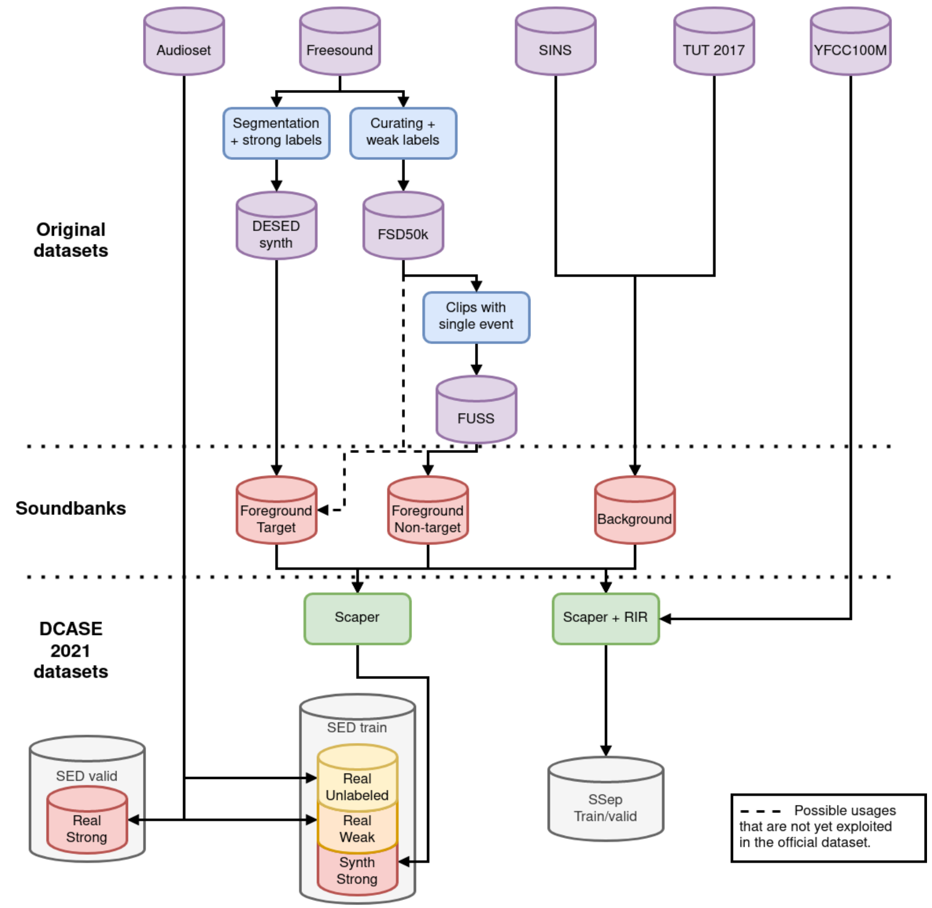


DCASE: task 4: datasets

Dataset	Subset	Type	Usage	Annotations	type	frequency
DESED	Real: weakly labeled	Recorded soundscapes	Training	Weak labels (no timestamps)	Target	44.1kHz
	Real: unlabeled	Recorded soundscapes	Training	No annotations	Target	44.1kHz
	Real: validation	Recorded soundscapes	Validation	Strong labels (with timestamps)	Target	44.1kHz
	Real: public evaluation	Recorded soundscapes	Evaluation **(do not use this subset to tune hyperparamters)**	Strong labels (with timestamps)	Target	44.1kHz
	Synthetic: training	Isolated events + synthetic soundscapes	Training/validation	Strong labels (with timestamps)	Target	16kHz
	Synthetic: evaluation	Isolated events + backgrounds	Evaluation **(do not use this subset to tune hyperparamters)**	Event level labels (no timestamps)	Target	16kHz
SINS		Background	Training/validation	No annotations	N/A	16kHz
TUT Acoustic scenes 2017, development dataset		Background	Training/validation	No annotations	N/A	44.1kHz
FUSS dataset		Isolated events + synthetic soundscapes	Training/validation	Weak annotations from FSD50K (no timestamps)	Target and non-target	16kHz
FSD50K dataset		Isolated events + recorded soundscapes	Training/validation	Weak annotations (no timestamps)	Target and non-target	44.1kHz
YFCC100M dataset		Recorded soundscapes	Training/validation	No annotations	Sound sources	44.1kHz

DCASE: sound event training set

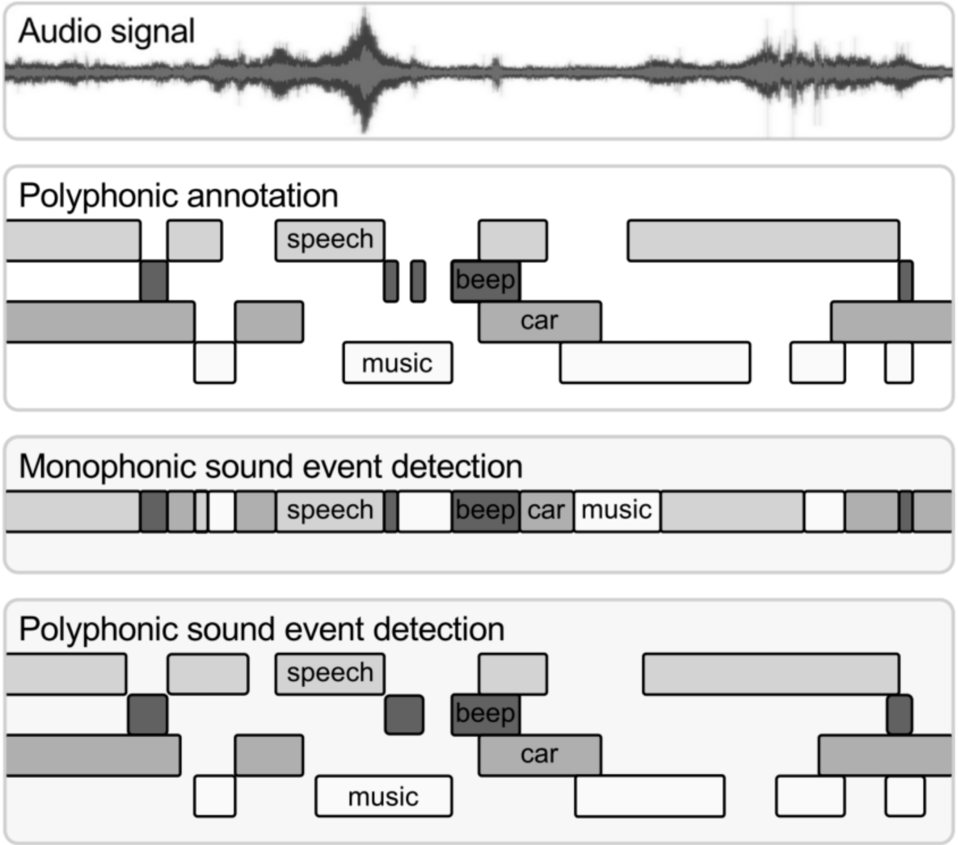
- Weakly labeled training set : 1578 clips (2244 class occurrences)
- 14,412 unlabeled clips
- 10000 strongly labeled synthetic clips generated with Scaper.
- Non-target events from FUSS.
- Validation set (manually verified) with similar class distribution than the weakly labeled training set.



https://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments
Salamon et. al. « Scaper: A Library for Soundscape Synthesis and Augmentation ». In *IEEE WASPAA 2017*
Wisdom et. al. « What's all the Fuss about Free Universal Sound Separation Data? » In IEEE *ICASSP 2021*

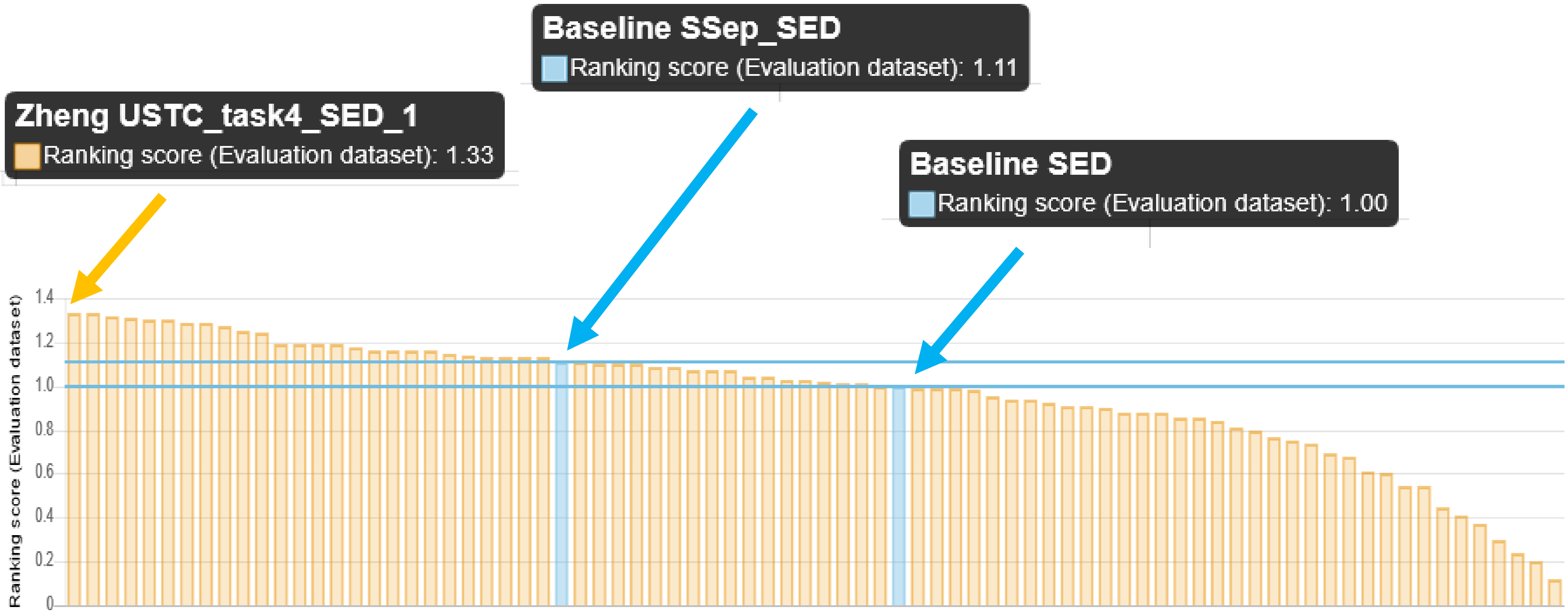
DCASE: Sound Event Detection and Separation in Domestic Environments

- **Evaluation: What is polyphonic event detection ?**



DCASE: Sound Event Detection and Separation in Domestic Environments

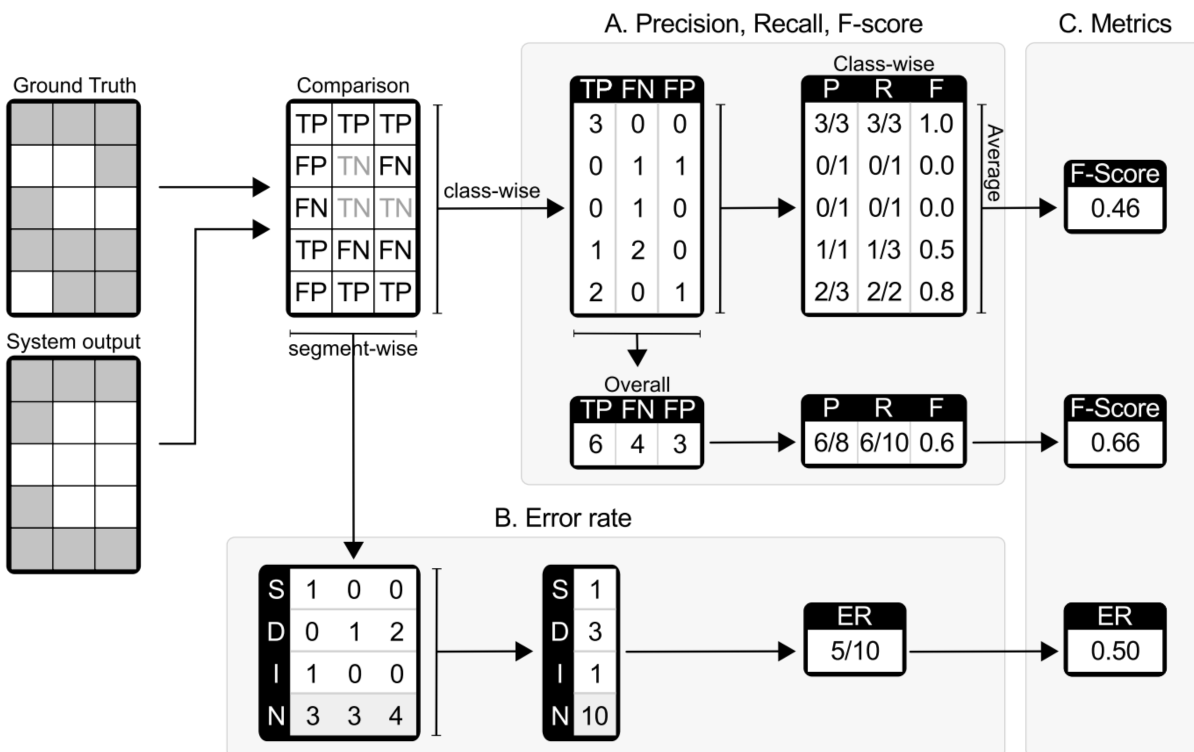
- **Performances**



Zheng, Xu and Chen, Han and Song, Zheng USTC Team's Submission For DCASE2021 Task4 – Semi-Supervised Sound Event Detection, DCASE2021 Challenge, Techn. Report

DCASE: Sound Event Detection and Separation in Domestic Environments

- How to evaluate Sound detection performances : **segment based metrics?**



TP/FP : True/False Positive
TN/FN: True/False Negative

P: Precision $= \frac{TP}{(TP+FP)}$

R: Recall $= \frac{TP}{(TP+FN)}$

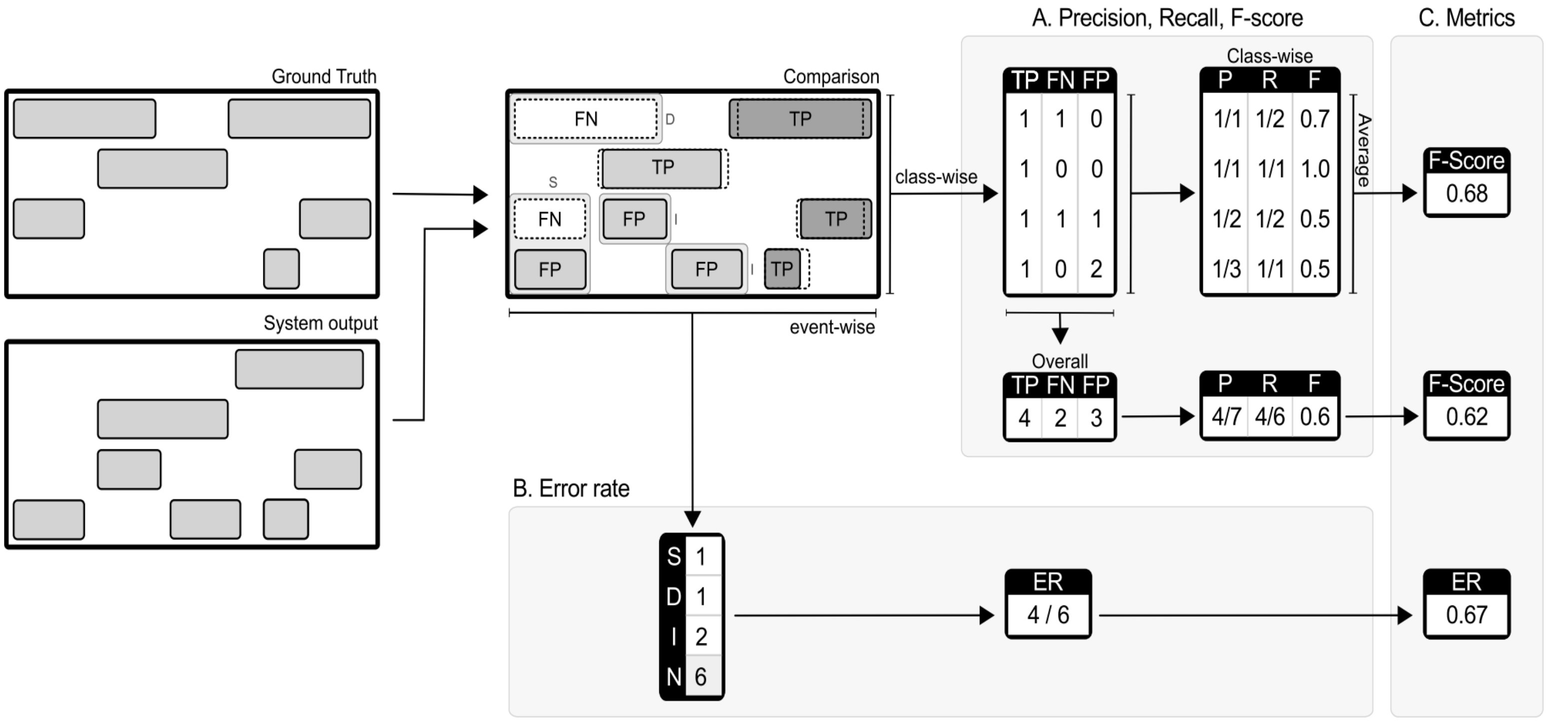
F: F-measure $= \frac{2.P*R}{(P+R)}$

Error types:
- S: Substitutions
- D: Deletions
- I: Insertions
- N: number of events
- active in a segment

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. Applied Sciences, 6(6):162, 2016. URL: http://www.mdpi.com/2076-3417/6/6/162, doi:10.3390/app6060162.

DCASE: Sound Event Detection and Separation in Domestic Environments

- How to evaluate Sound detection performances : **Event-based metrics?**



Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. Applied Sciences, 6(6):162, 2016. URL: http://www.mdpi.com/2076-3417/6/6/162, doi:10.3390/app6060162.

DCASE: Sound Event Detection and Separation in Domestic Environments

- **How to evaluate Sound detection performances ?**

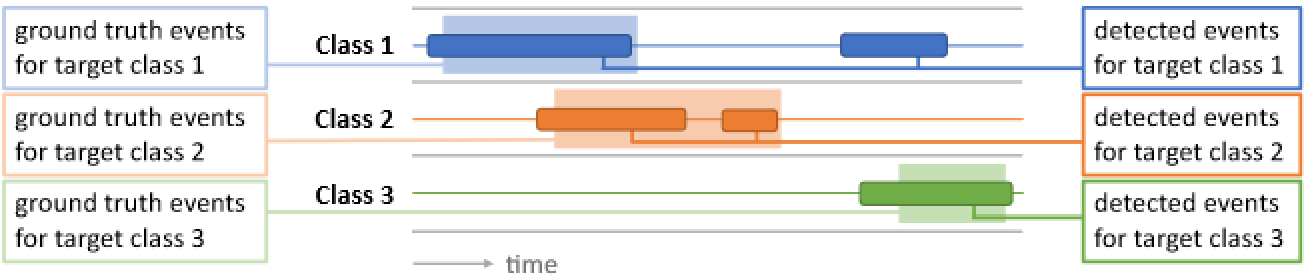
  - Polyphonic Sound event Detection Scores (PSDS)
    - computed over the real recordings in the evaluation set
    - PSDS values are computed using 50 operating points (linearly distributed from 0.01 to 0.99)
    - Event-based metrics

  - Many metrics « parameters »
    - Detection Tolerance criterion (DTC)
    - Ground Truth intersection criterion (GTC)
    - Cost of instability across class
    - Cross-Trigger Tolerance criterion
    - …

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. Applied Sciences, 6(6):162, 2016. URL: http://www.mdpi.com/2076-3417/6/6/162, doi:10.3390/app6060162.

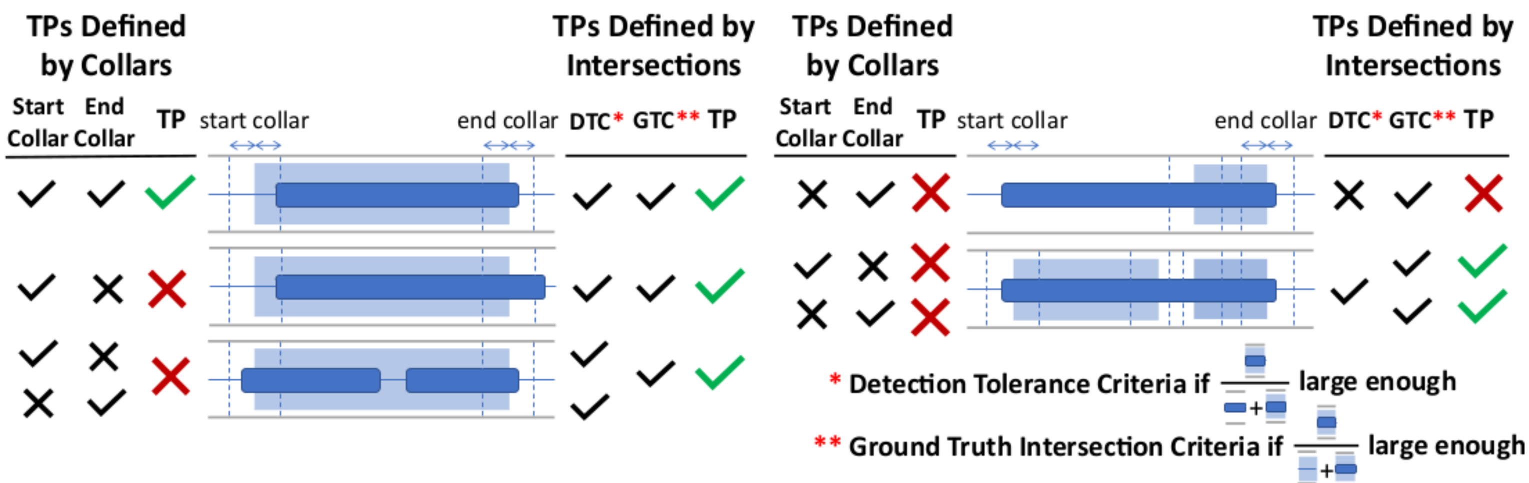
Evaluation of polyphonic sound event detection

- **Detected events vs Ground truth events**



Bilen et. al.. « A Framework for the Robust Evaluation of Sound Event Detection ». In *ICASSP 2020*

Metrics : Polyphonic sound event detection score (PSDS)



(a) TP decisions made by collars (left) vs. *DTC/GTC* (right).

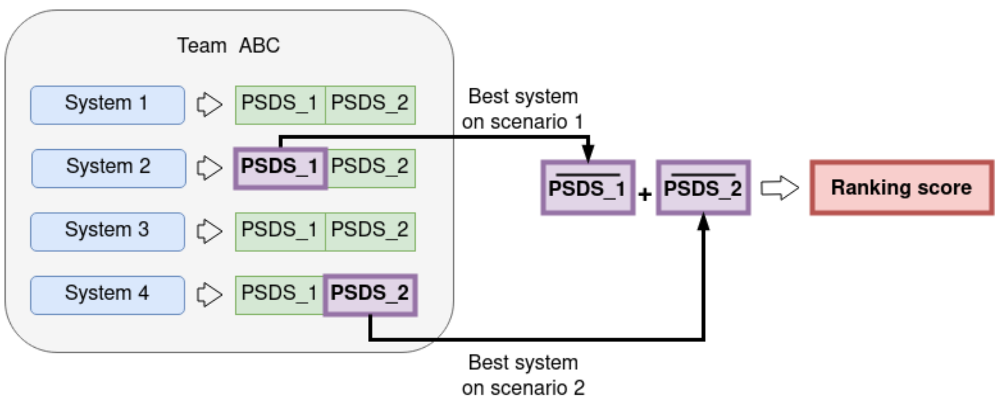
- **Detection Tolerance Criteria:** controls how precise a system detection must be with respect to all the ground truths of the same class that it intersects.
- **Groudtruth Intersection Criteria:** defines the amount of minimum overlap necessary to count a ground truth as correctly detected.

4
7

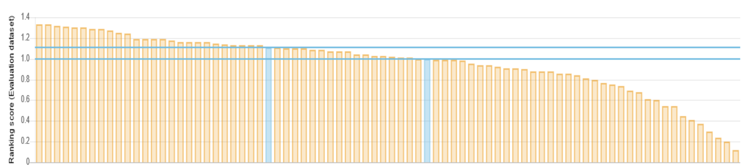
Bilen et. al.. « A Framework for the Robust Evaluation of Sound Event Detection ». In *ICASSP 2020*

Evaluation

- **Ranking teams with their two best systems on each scenario :**
  1. The system needs to react fast upon an event detection (e.g. to trigger an alarm, adapt home automation system...). The localization of the sound event is then really important.
  2. The system must avoid confusing between classes but the reaction time is less crucial than in the first scenario.

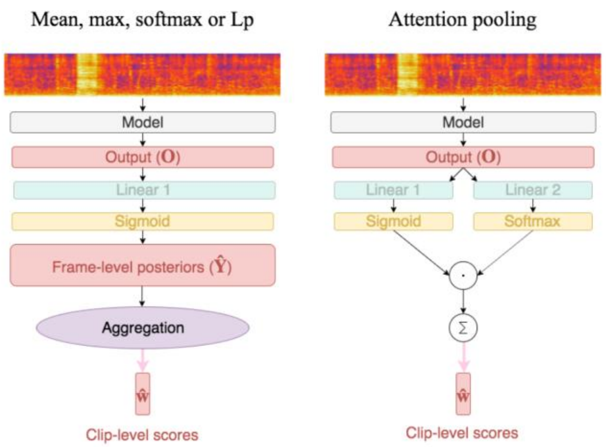


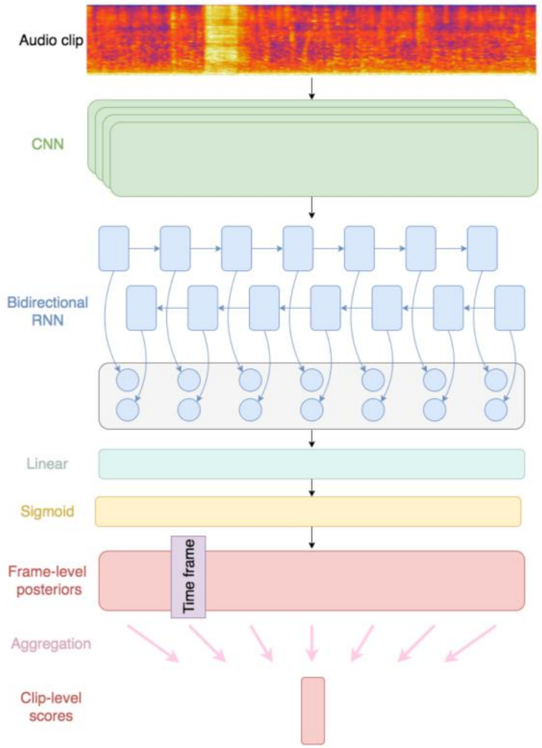
Ranking score



Baseline System : CRNN & Mean Teacher

- Encoding frames with a CRNN
- Frame-level classification using dense layers
- Aggregation of frame-level output to get clip-level prediction

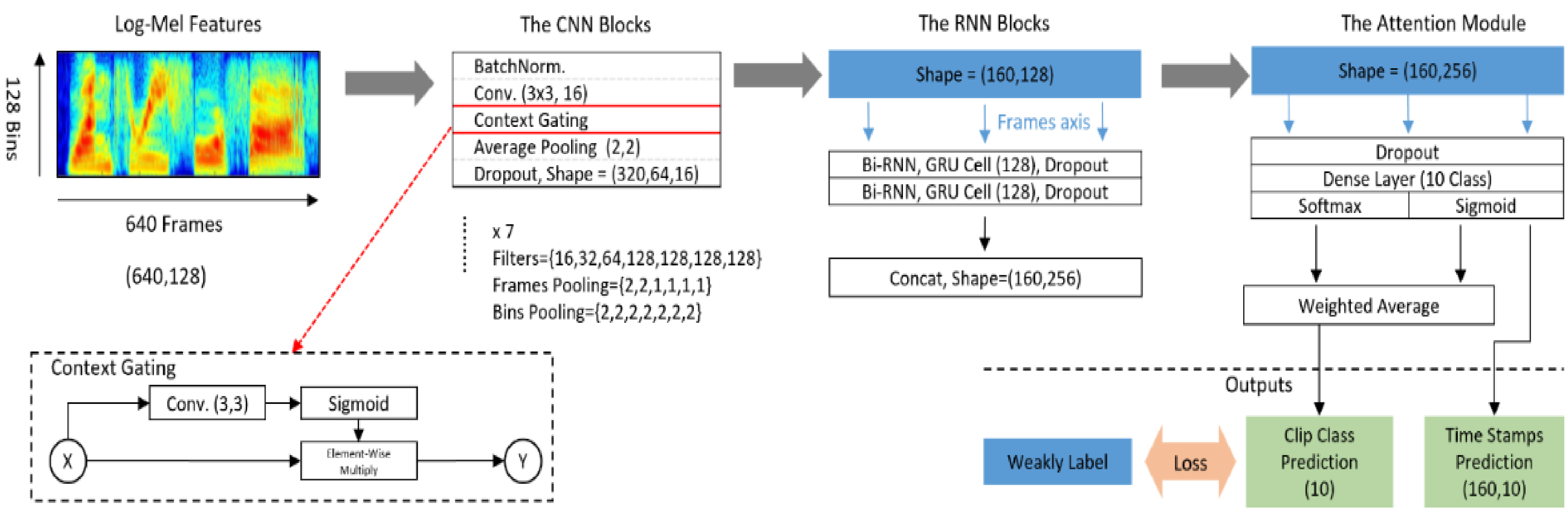




Turpault et. al. « Analysis of weak labels for sound event tagging». HAL-Inria 2021

DCASE: Sound Event Detection and Separation in Domestic Environments

- **Baseline system (another view..)**

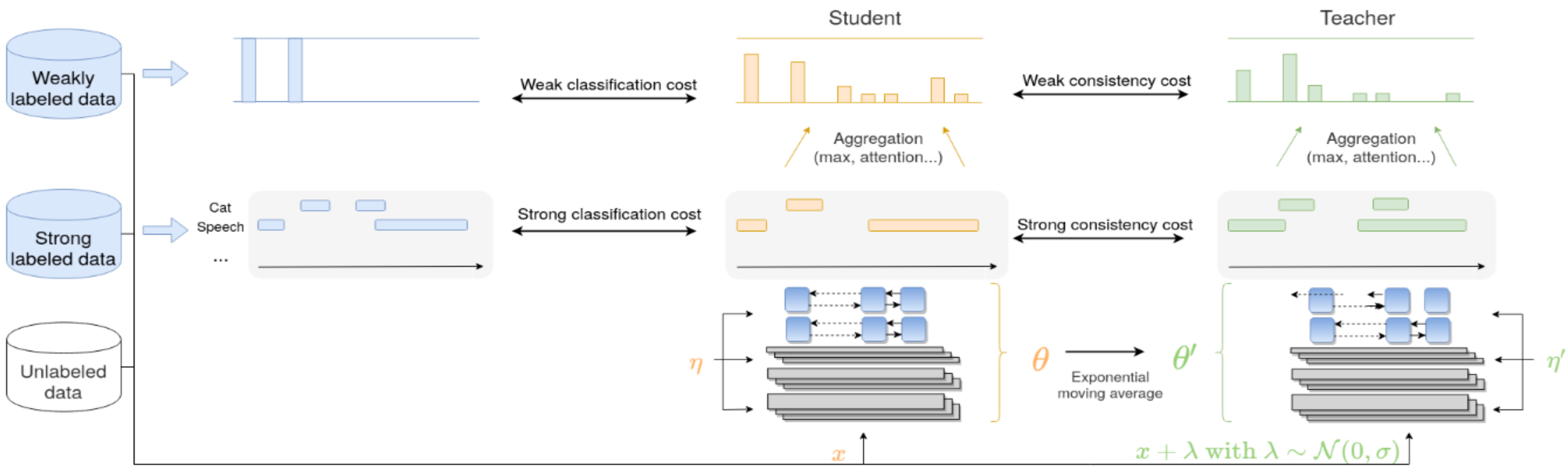


L. JiaKai, "Mean teacher convolution system for dcase 2018, task 4," DCASE2018 Challenge, Tech. Rep., September 2018

DCASE: Baseline System

- The student model parameters are updated based on a classification loss and a consistency loss between the student outputs and the teacher outputs.
- *The teacher model is not trained and is an average of consecutive student models*
- *The student model is used at inference time*

$$\begin{aligned} L(\theta) = & L_{class_w}(\theta) + \sigma(\lambda) L_{cons_w}(\theta) \\ & + L_{class_s}(\theta_s) + \sigma(\lambda) L_{cons_s}(\theta_s) \end{aligned}$$



Nicolas Turpault, Romain Serizel. Training Sound Event Detection On A Heterogeneous Dataset. DCASE Workshop, Nov 2020, Tokyo, Japan. hal-02891665v2
A. Tarvainen, H. Valpola. « Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results ». In Advances in Neural Information Processing Systems

Summary

- **Machine listening: a domain of growing interest**
- **… with many applications**

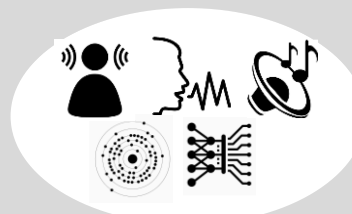
Audio surveillance, Audio scene analysis

Security, Health monitoring, bioacoustics



Transport & Communications

Autonomous cars, audio enhancement

Industry

Predictive maintenance

- **Some difficulties:**
  - Obtaining real-case annotated databases
  - Towards few-shot learning, unsupervised learning, …
  - … and distributed or sensor-based learning

A few additional references…

- ***Acoustic Scene and event recognition***

  - *V. Bisot & al., "Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, (2017),*
  - *V. Bisot & al., Leveraging deep neural networks with nonnegative representations for improved environmental sound classification IEEE International Workshop on Machine Learning for Signal Processing MLSP, Sep 2017, Tokyo,*
  - A Mesaros & al. Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 challenge IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (2), 379-393
  - D. Barchiesi, D. Giannoulis, D. Stowel, and M. D. Plumbley, “Acoustic scene classification: Classifying environments from the sounds theyproduce,”IEEE Signal Processing Magazine, vol. 32, no. 3, pp. 16–34, 2015
  - P. Lopez & al. “Ensemble of Convolutional Neural Networks”, in DCASE 2020 Acoustic Scene Classification Challenge
  - T. Virtanen, M. Plumbley, D. Ellis, Computational Analysis of Sound Scenes and Events, Springer, 2018
  - R. Serizel, V. Bisot, S. Essid, G.Richard, Acoustic Features for Environmental sound Analysis, in Computational Analysis of Sound Scenes and Events, T. Virtanen, D. Ellis, M. Plumbley Eds., Springer International Publishing AG, pp 71-101, 2018