# PLDAC

## Chants d'oiseaux

25 mai 2023

Valentin Bencheci, Aymeric Delefosse

Master DAC - Sorbonne Université

# DCASE

# Contexte

## Detection and Classification of Acoustic Scenes and Events

- Workshop & Challenge annuels : rassemblement de la communauté de chercheurs du domaine du traitement du signal audio .
- Se concentre principalement sur la détection et la classification des scènes et événements acoustiques.

## Exemples

- 2022 & 2023 : *Few-shot Bioacoustic Event Detection*
- 2021 : *Automated Audio Captioning*
- 2018 : *Bird audio detection* ← notre problématique

# Quel lien avec DAC ?

Qui dit détection ou classification... implique souvent des techniques issues de l'apprentissage automatique.

## Mais...

La communauté du traitement du signal audio a connu un certain décalage par rapport aux développements dans les domaines de l'informatique et de la data science.

## Pourquoi ?

- Complexité et de la spécificité des données audio.
- Communauté qui a évolué de manière relativement isolée ⇒ manque de transfert de connaissances.

# Quel lien avec DAC ?

Mais depuis ces dernières années, il existe une convergence plus étroite entre les deux communautés.

## Les données au cœur de tout

- Avancées dans le *deep learning* dans ce domaine, grâce aux réseaux de neurones convolutionnels et récurrents.
- Soutenues par la disponibilité de grandes quantités de données audio.
- Progrès de l'informatique distribuée.

# Challenge et données

# *Bird Audio Detection*

SCIENCES SORBONNE UNIVERSITÉ

## Challenge

- Développer un système capable de détecter la présence ou l'absence de sons d'oiseaux dans des enregistrements audio.
- Décision binaire ou probabiliste $\in [0,1]$.
- $\Rightarrow$ Capacité à généraliser (défi important).

# Données

## Freefield

- Enregistrements de terrain à travers le monde.
- Diversité des emplacements et des environnements.
- ⚠Classes déséquilibrées : 25 % sons d'oiseaux

## Warblr

- Enregistrements audio provenant d'utilisateurs de l'application Warblr, couverture variée des emplacements et des environnements britanniques.
- Bruits de fond tels que le trafic, les voix humaines et les imitations d'oiseaux par les humains.
- ⚠Classes déséquilibrées : 75 % sons d'oiseaux

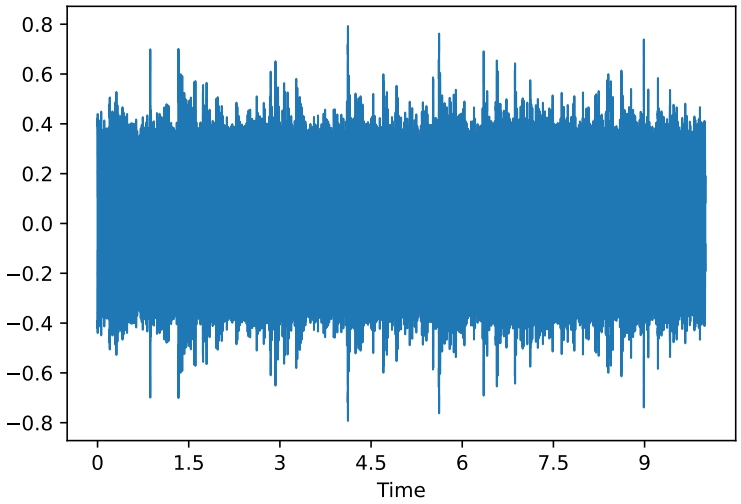
# Données

## BirdVox

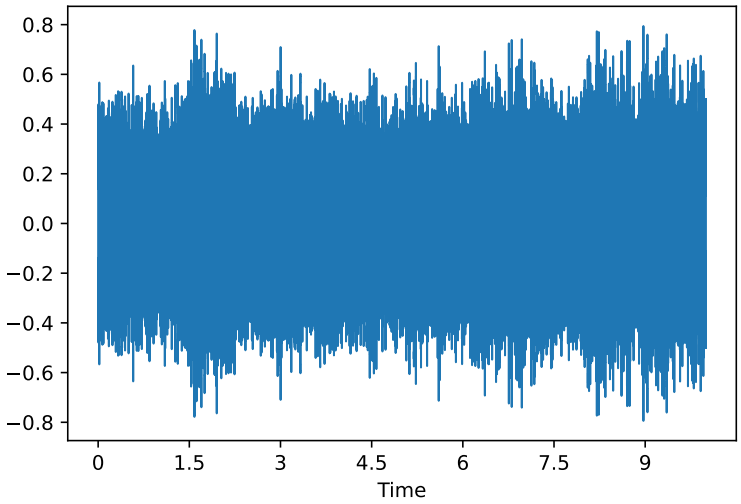
- Enregistrements effectués par des unités de surveillance à distance.
- Échantillonnage uniforme sur différents moments de la journée et conditions météorologiques.
- Classes équilibrées

Au total : 35 690 enregistrements avec 17 997 sons contenant des chants d'oiseaux et 17 693 sons n'en contenant pas.

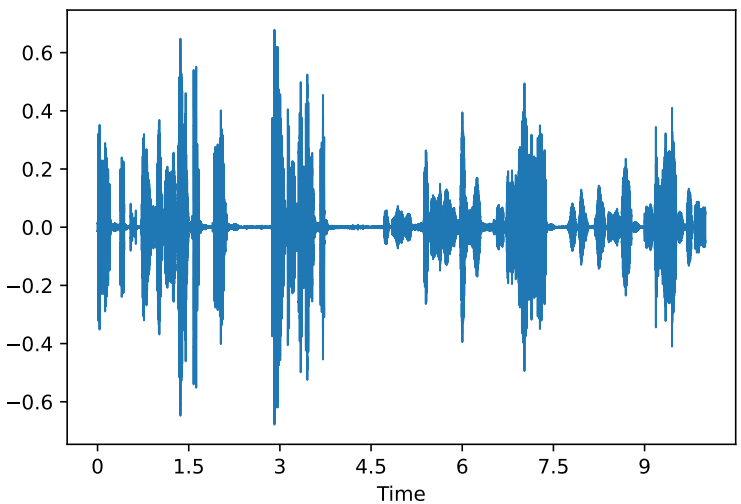
# Principal défi

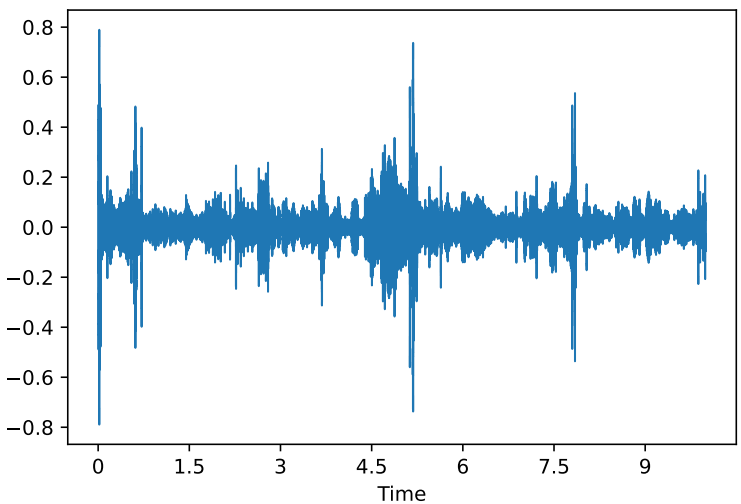



(a) Enregistrement contenant des oiseaux




(b) Enregistrement ne contenant pas d'oiseaux



# Représentation temporelle (forme d'onde)

## Avantages

- Visualiser intuitive du signal audio.
- Identifier les moments de silence ou de faible amplitude, ainsi que les pics et les moments d'intensité sonore élevée.
- Repérer des motifs ou des caractéristiques spécifiques dans le signal audio.

## Inconvénients

- Pas d'informations détaillées sur la nature exacte des composantes fréquentielles du son.
- Ne permet pas d'identifier précisément les différentes sources sonores présentes dans le signal.
- Limité par la résolution temporelle de l'affichage.

# Représentation spectrale

Nécessité de se tourner vers des représentations **spectrales**.

## Avantages

- Visualiser des composantes fréquentielles du signal audio.
- Identifier les changements de fréquence, les harmoniques et les caractéristiques spectrales spécifiques.
- Utile pour la **détection d'événements sonores**, la classification d'instruments, l'analyse musicale...

## Inconvénients

- Peut nécessiter une résolution temporelle plus fine pour représenter les variations rapides dans le signal.
- Processus parfois irréversible...

# Représentation spectrale

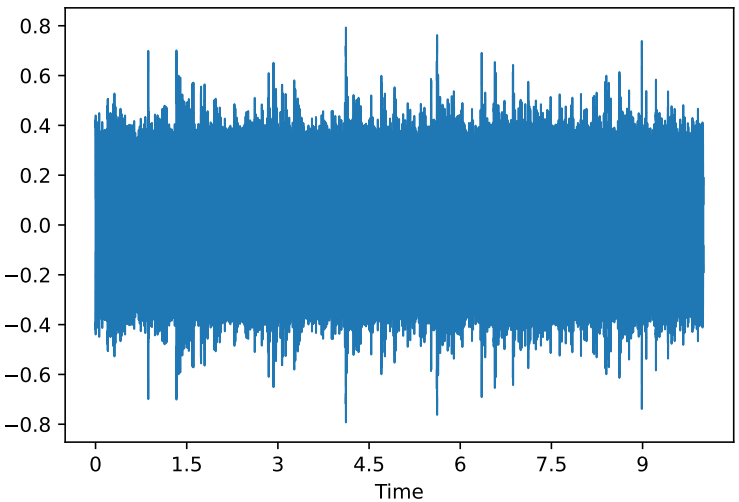
Mais il existe plusieurs représentations... laquelle choisir ?

## Les plus communes...

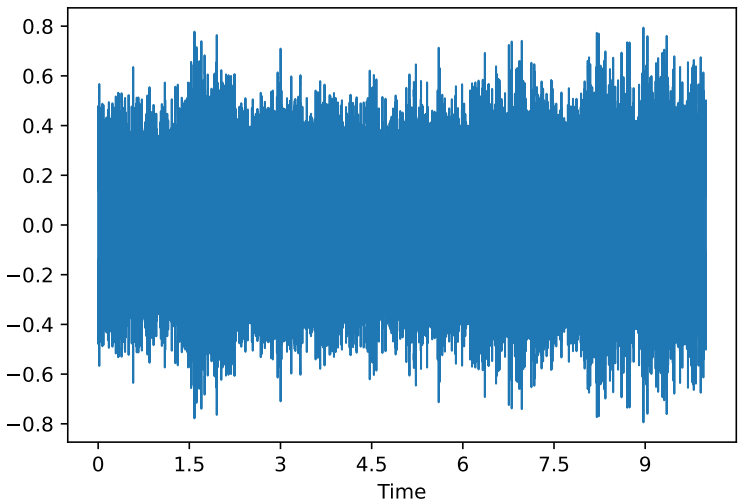
- Spectrogramme ?
- Chromagramme ?
- Cepstre ?

⇒ Dans tous les cas : Fourier !

# Représentation spectrale : Chromagramme




(a) Enregistrement contenant des oiseaux




(b) Enregistrement ne contenant pas d'oiseaux

# Représentation spectrale : Chromagramme

Mais il existe plusieurs représentations... laquelle choisir ?

## Chromagramme

- Représentation temps/notes de musique.
- Met l'accent sur les informations tonales et harmoniques d'un signal audio.
- Largement utilisé dans des applications telles que la transcription musicale automatique, la reconnaissance des accords et l'analyse comparative de morceaux de musique.

⇒ Représentation **non pertinente** pour notre problème.

# Représentation spectrale : Spectrogramme

Mais il existe plusieurs représentations... laquelle choisir ?

## Spectrogramme

- Représentation temps/fréquence.
- Met l'accent sur la répartition spectrale de l'énergie ou de la puissance du signal audio.
- Utile pour observer les changements de fréquences, les harmoniques, les transitions tonales et les événements sonores dans le temps.
- Possibilité de passer en 3D : temps/fréquence/amplitude.

⇒ Représentation **pertinente** pour notre problème.

# Représentation spectrale : Cepstre

Mais il existe plusieurs représentations... laquelle choisir ?

## Cepstre

- Met l'accent sur l'étude structures périodiques temporelles des caractéristiques spectrales d'un signal audio.
- Fournit des informations sur les enveloppes spectrales et les variations dans le domaine fréquentiel du signal.
- Utile pour l'analyse des formants vocaux, la détection des harmoniques, la reconnaissance de la parole et d'autres applications liées aux caractéristiques temporelles du signal audio.
- Réversible !

⇒ Représentation **pertinente** pour notre problème.

# Expérimentations et résultats

# Performance

| Modèle | Pooling | Freefield1010 | | Warblrb10K | | BirdVox | |
|---|---|---|---|---|---|---|---|
| | | AUC | F1 | AUC | F1 | AUC | F1 |
| CNN | Max probability | .847 | .730 | .849 | .926 | .873 | .860 |
| CNN | Temporal | .842 | .706 | .853 | **.937** | .888 | **.885** |
| $\text{CRNN}_{RNN}$ | Temporal | .841 | .708 | .789 | .921 | .877 | .868 |
| $\text{CRNN}_{LSTM}$ | Temporal | .857 | .783 | .868 | .906 | .882 | .873 |
| $\text{CRNN}_{LSTM}$ | Soft Attention | .849 | .783 | .858 | .908 | .870 | .857 |
| $\text{CRNN}_{Transformers}$ | Temporal | .820 | .755 | .844 | 884 | .836 | .811 |
| $\text{CRNN}_{GRU}$ | Max probability | **.861** | .778 | .860 | .908 | .882 | .872 |
| $\text{CRNN}_{GRU}$ | Temporal | .840 | .719 | .818 | .923 | .890 | .883 |
| $\text{CRNN}_{GRU}$ | Soft Attention | .859 | **.784** | **.870** | .916 | **.891** | .883 |

Table 1 – Performance de différents modèles sur Freefield, Warblr et BirdVox

# Some papers for noisy images datasets

SCIENCES SORBONNE UNIVERSITÉ

## « Convolutional Recurrent Neural Networks for Bird Audio Detection » Cakir et al. 2017

- Define a regularized loss for training the CNN
- Can be seen as looking for the label of similar images for regularization
- Results slightly better than Sukhbaatar model

# References

# References I

Cakir, Emre et al. (août 2017). « Convolutional Recurrent Neural Networks for Bird Audio Detection ». In : *2017 25th European Signal Processing Conference (EUSIPCO)*. Kos, Greece : IEEE, p. 1744-1748. isbn : 978-0-9928626-7-1. doi : 10.23919/EUSIPCO.2017.8081508. (Visité le 21/03/2023).