

# 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*
  - 2016/2017/2018 : *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

## 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

Warbler

- 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

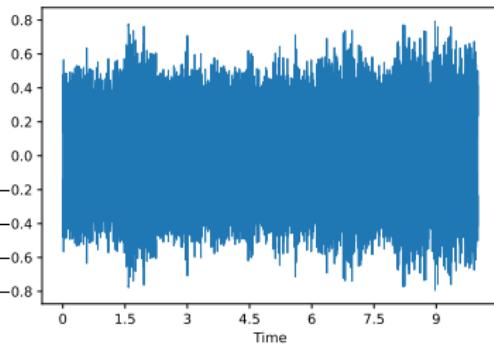
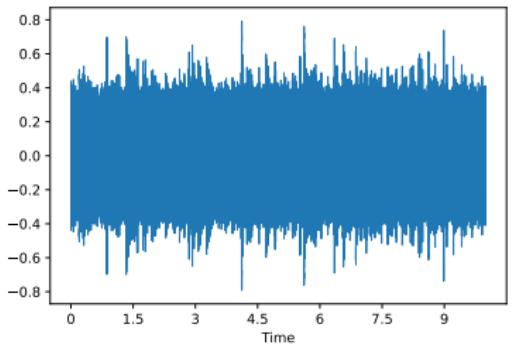
Données

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 de 10 secondes avec 17 997 sons contenant des chants d'oiseaux et 17 693 sons n'en contenant pas.

## Principal défi



## 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



Mais il existe plusieurs représentations... laquelle choisir ?

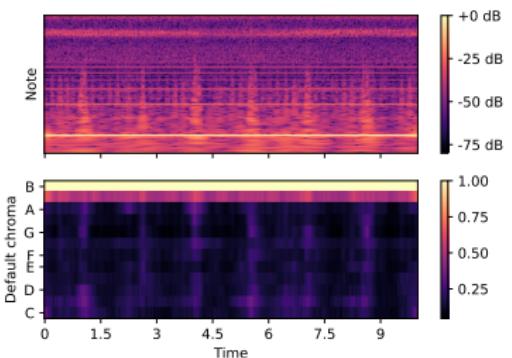
## Chromagramme

- 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.

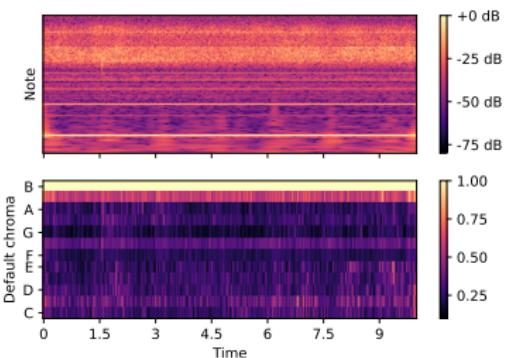
⇒ Représentation non pertinente pour notre problème.

## Représentation spectrale : Chromagramme

## Sur des données issues de BirdVox...



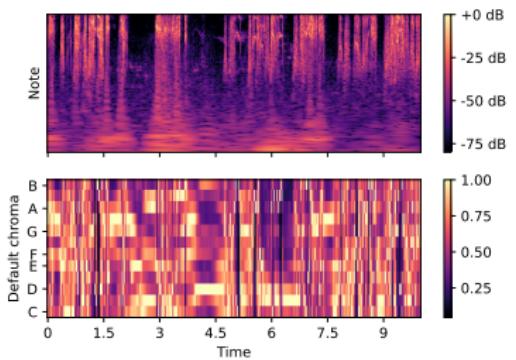
(a) Enregistrement contenant des oiseaux



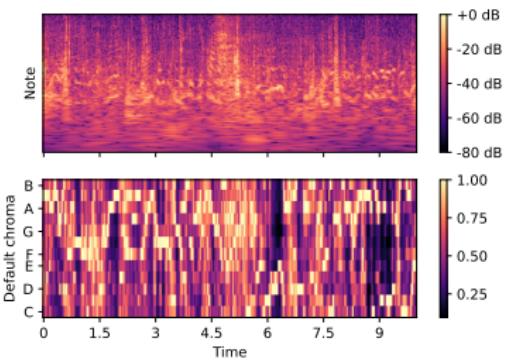
(b) Enregistrement ne contenant pas d'oiseaux

## Représentation spectrale : Chromagramme

## Sur des données issues de Freefield...



(a) Enregistrement contenant des oiseaux



(b) Enregistrement ne contenant pas d'oiseaux

## Représentation spectrale : Spectrogramme



Mais il existe plusieurs représentations... laquelle choisir ?

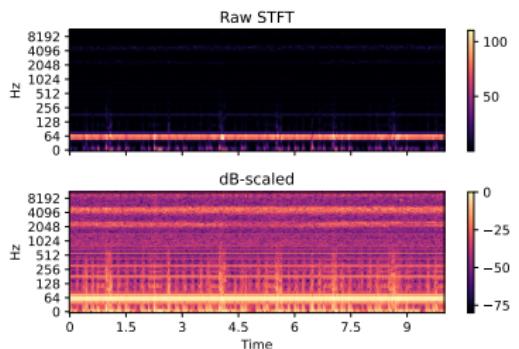
## Spectrogramme

- 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.

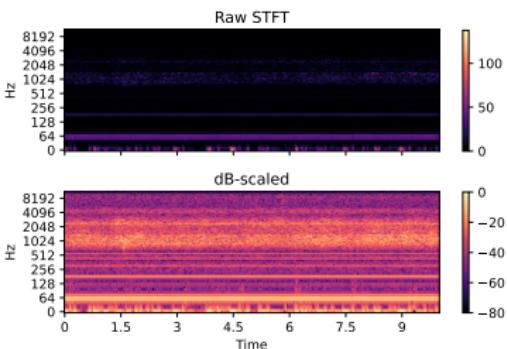
⇒ Représentation **intéressante** pour notre problème. Est-ce la plus pertinente ?

## Représentation spectrale : Spectrogramme

Sur des données issues de BirdVox...



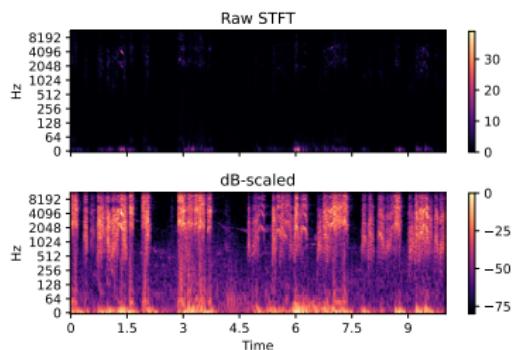
(a) Enregistrement contenant des oiseaux



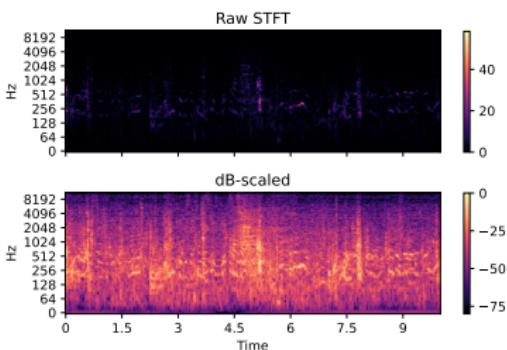
(b) Enregistrement ne contenant pas d'oiseaux

# Représentation spectrale : Spectrogramme

Sur des données issues de Freefield...



(a) Enregistrement contenant des oiseaux



(b) Enregistrement ne contenant pas d'oiseaux

# Représentation spectrale : Cepstre

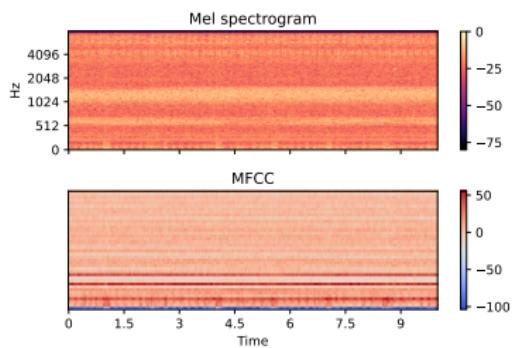
Mais il existe plusieurs représentations... laquelle choisir ?

## Cepstre

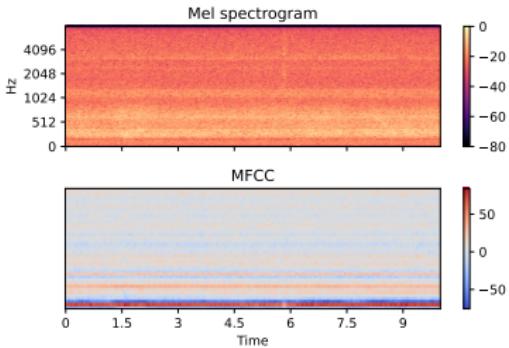
- Met l'accent sur l'étude structures périodiques temporelles des caractéristiques spectrales d'un signal audio.
- Fournit des informations sur les enveloppes spectrales et les variations dans le domaine fréquentiel du signal.
- Utile pour l'analyse des formants vocaux, la détection des harmoniques,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

# Représentation spectrale : Cepstre

Sur des données issues de BirdVox...



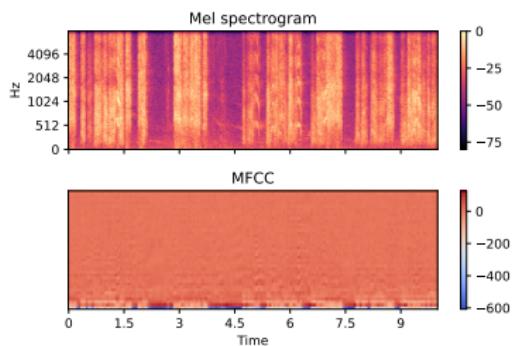
(a) Enregistrement contenant des oiseaux



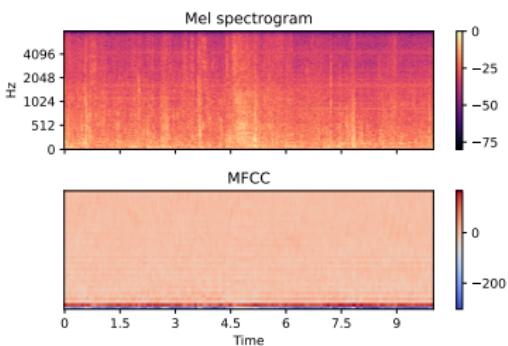
(b) Enregistrement ne contenant pas d'oiseaux

# Représentation spectrale : Cepstre

Sur des données issues de Freefield...



(a) Enregistrement contenant des oiseaux



(b) Enregistrement ne contenant pas d'oiseaux

## Et maintenant ?

En plus de transformer la représentation, il y a d'autres façons de pré-traiter les données en traitement du signal. Entre autres :

### Pré-traitements

- Normalisation
- Réduction de bruit
- Élimination des artefacts et des imperfections
- Augmentation de données
- Modification temporelle et fréquentielle
- Détection et suppression de "mots-clés"

Beaucoup de pistes : tout explorer est **chronophage** mais serait intéressant, d'un point de vue théorique et fondamental.

## Pré-traitements



## Normalisation

- Normalisation de l'amplitude
  - Compression dynamique
  - Ajustement des niveaux de volume

## Noise reduction

- Filtrage adaptatif
  - Soustraction spectrale
  - Réduction de bruit par seuillage
  - Réduction de bruit par modélisation statistique

## Pré-traitements

### Élimination des artefacts et des imperfections

- Élimination des silences et des clics
- Détection et suppression des artefacts
- Correction des distorsions
- Suppression de la réverbération

### Data augmentation

- Variations de vitesse et de tonalité
- Mixage de sources audio
- Perturbations et déformations synthétiques
- **Chunking**

## Pré-traitements



## Modification temporelle et fréquentielle

- Normalisation temporelle (time-stretching)
  - Modification de la vitesse (time-scaling)
  - Modification de la tonalité (pitch-shifting)
  - Time-domain resampling

## Détection et suppression de "mots-clés"

- Séparation de parole et de musique
  - Détection et suppression de mots-clés spécifiques

Liste non-exhaustive... mais le pré-traitement n'est qu'une première étape, son efficacité dépendra également du modèle d'apprentissage derrière !

## En résumé...

Les sons d'oiseaux peuvent être distingués en chants ou cris, en se basant sur la complexité, la longueur et le contexte. Ici, on ne les distinguerà pas.

### Les sons d'oiseaux

- Plus ou moins long (chant vs cri).
- Motifs répétitifs structurés qui peuvent varier dans le temps (rythme, tempo, puissance, trilles, glissandos, vibrato...).
- Couvre une large gamme de fréquence mais le chant d'une espèce oiseau peut occuper une plage de fréquence spécifique/limitée (en fonction de l'espèce).  
→ Notion de *syllabe*.

# Modélisation

État de l'art



Ce qui marche le mieux : les réseaux de neurones.

## Grand gagnant du challenge 2016/2017 :

- Architecture CNN classique.
  - Pré-traitements : silence/noise trimming + data augmentation.
  - PCA + Agglomerative Clustering à partir des features du spectrogramme (moyenne, écart-type, 1-percentile, 99-percentile).
  - Adaptation de domaine : pseudo-labeling.
  - Ensembling : model averaging.

## État de l'art



Ce qui marche le mieux : les réseaux de neurones.

## Prix du jury du challenge 2016/2017 :

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

- Modèle « hybride » : CNN + RNN = CRNN.
  - Résultats très proches du modèle le plus performant (88.5 % vs 88.7 %).
  - Bien moins computationally intensive que le modèle le plus performant.
  - Pas de data augmentation, d'adaptation de domaine ou d'ensembling.

## État de l'art



Je crois que c'est celui là pour les chunks

« Acoustic Bird Detection with Deep Convolutional Neural Networks » Lasseck 2018

## État de l'art



Et aussi parler des SVM / GMM / HMM très vite fait car c'est ce qui se faisait bcp avant DEEP et se fait encore parfois

Cakir et al. 2017

- Modèle développé dans le cadre de la détection de sons polyphoniques, c'est-à-dire la détection de plusieurs événements sonores se chevauchant.
  - Postulat que CNN et RNN sont deux méthodes complémentaires.
    - ⇒ CNN identifie des features spatiales.
    - ⇒ RNN gère des séquences et des dépendances temporelles.

Mais...

...modèle toujours d'actualité ?

CRNN - Ce n'est pas une révolution...



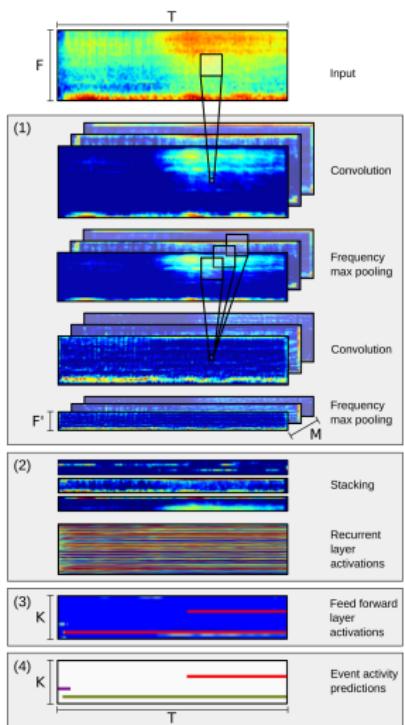
Pas nouveau... mais différent des architectures CRNN précédentes :

Cakir et al. 2017

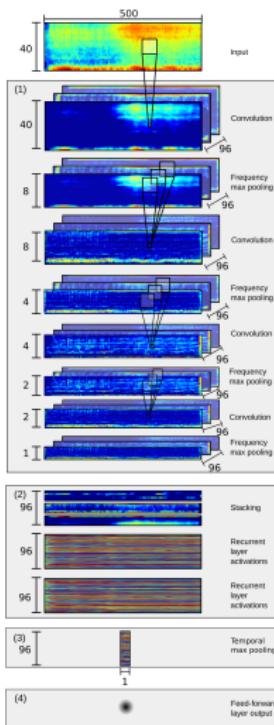
- Aucune utilisation de couche linéaire après le CNN ou RNN.
  - Kernels de convolution plus petits.
  - Jusqu'à 4 couche de convolution, 3 couche de récurrence.
  - Utilisation de GRU au lieu de LSTM.
  - Séquences plus longues.
  - Utilisation optionnelle d'une couche supplémentaire de *temporal pooling* à la sortie du RNN.

⇒ Surpasse les GMM, FNN, CNN et RNN classiques.

## CRNN - Architecture



### (a) Architecture générique du CRNN



(b) Architecture proposée pour la détection d'oiseaux

## CRNN - Que peut-on en dire ?



On travaille sur des *images* (spectrographes), dans le temps ⇒ le modèle est entièrement fondé, mais pouvons-nous apporter des améliorations ?

## La couche récurrente : GRU ?

GRU se "limite" à capturer des dépendances à court terme.

- En a-t-on *réellement* besoin ?
  - Un RNN suffirait-il ?
  - Un LSTM serait-il plus apte à capturer des informations ?
  - Méthodes plus "modernes"... ?

CRNN - Que peut-on en dire ?

On travaille sur des *images* (spectrographes), dans le temps ⇒ le modèle est entièrement fondé, mais pouvons-nous apporter des améliorations ?

### La couche de *temporal pooling*

But : réduire la dimension temporelle tout en préservant les features les plus importantes.

- En a-t-on *réellement* besoin ?
  - Pourquoi ne pas regarder directement sur la sortie temporelle ?
  - Méthodes plus "modernes" ... ?

CRNN - Méthodes plus modernes ?



## La révolution de l'attention.

« Attention Is All You Need » Vaswani et al. 2017

Se concentrer sur des parties spécifiques d'une séquence, afin de capturer des relations à longue portée.

- Des Transformers à la place d'une couche récurrente ?
  - De l'attention *pure* (soft ? hard ?) à la place d'une couche de pooling temporelle ?

# Expérimentations et Résultats

## Méthode d'évaluation



Seulement nos trois jeux de données. Pas les mêmes données d'évaluation que le challenge.

## AUC, F1, (Accuracy + Confusion Matrix)

## Performance

Modèle	Pooling	Freefield1010		Warblr10K		BirdVox	
		AUC	F1	AUC	F1	AUC	F1
CNN	Max probability	.847	.730	.849	.926	.873	.860
CNN	Temporal	.842	.706	.853	<b>.937</b>	.888	<b>.885</b>
CRNN <sub>RNN</sub>	Temporal	.841	.708	.789	.921	.877	.868
CRNN <sub>LSTM</sub>	Temporal	.857	.783	.868	.906	.882	.873
CRNN <sub>LSTM</sub>	Soft Attention	.849	.783	.858	.908	.870	.857
CRNN <sub>Transformers</sub>	Temporal	.820	.755	.844	884	.836	.811
CRNN <sub>GRU</sub>	Max probability	<b>.861</b>	.778	.860	.908	.882	.872
CRNN <sub>GRU</sub>	Temporal	.840	.719	.818	.923	.890	.883
CRNN <sub>GRU</sub>	Soft Attention	.859	<b>.784</b>	<b>.870</b>	.916	<b>.891</b>	.883

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

# Bibliographie

## Bibliographie 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).
  -  Grill, Thomas et Jan Schluter (août 2017). « Two Convolutional Neural Networks for Bird Detection in Audio Signals ». In : *2017 25th European Signal Processing Conference (EUSIPCO)*. Kos, Greece : IEEE, p. 1764-1768. isbn : 978-0-9928626-7-1. doi : 10.23919/EUSIPCO.2017.8081512. (Visité le 21/03/2023).
  -  Lasseck, Mario (2018). « Acoustic Bird Detection with Deep Convolutional Neural Networks ». In : *Workshop on Detection and Classification of Acoustic Scenes and Events*. (Visité le 23/05/2023).
  -  Vaswani, Ashish et al. (2017). « Attention Is All You Need ». In : *CoRR* abs/1706.03762. arXiv : 1706.03762. url : <http://arxiv.org/abs/1706.03762>.