ARIA Signal

Classification de chants d'oiseaux pour la conservation des espèces

NATHALIE HEINZELMEIER CHLOÉ SCHOLENT

20 NOVEMBRE 2025

Table des matières

Introduction

Contexte

L'essor du commerce international a favorisé la propagation d'espèces invasives, menaçant ainsi les écosystèmes locaux et les espèces endémiques. Chez les oiseaux, l'apparition d'une nouvelle espèce invasive peut entraîner une compétition accrue pour les ressources alimentaires et la dégradation des habitats naturels, mettant en péril la survie des espèces autochtones. Dans cette étude, nous nous intéressons à deux espèces d'oiseaux aux statuts écologiques opposés : *Psittacula krameri* (la perruche à collier), espèce invasive récemment introduite en Europe, et *Carduelis carduelis* (le chardonneret élégant), espèce endémique et protégée en France.

La perruche à collier, originaire d'Afrique et d'Inde, s'est établie en Europe suite à l'importation croissante d'animaux exotiques. Classée invasive par les autorités françaises, elle n'est pas protégée sur le territoire national, comme le rappelle l'arrêté du 29 octobre 2009 fixant la liste des oiseaux protégés en France [1]. Cette espèce se distingue notamment par un cri strident, facilement identifiable. À l'inverse, le chardonneret élégant est une espèce native d'Europe, menacée principalement par le braconnage. Reconnu comme une espèce protégée en France, sa conservation repose sur des mesures strictes interdisant la perturbation de son habitat naturel, le prélèvement d'individus, ainsi que la destruction de ses nids ou œufs. La protection efficace de cette espèce en danger est essentielle pour assurer la préservation de l'écosystème.

L'identification des espèces peut être facilitée par l'analyse acoustique des chants d'oiseaux, chaque espèce possédant un chant caractéristique. Ceci permettrait de reconnaître des espèces en danger même dans des environnements difficiles d'accès. L'efficacité de cette méthode est fortement dépendante de la qualité des signaux enregistrés. L'environnement naturel comporte de nombreux bruits parasites. Plusieurs étapes de traitement de ces signaux sont donc essentielles. La reconnaissance automatique des chants doit être suffisamment précise afin de bien différencier les espèces protégées des espèces invasives.

Chez l'homme, la vocalisation se caractérise par plusieurs métriques acoustiques. Le chant présente un timbre, correspondant à sa composition harmonique spécifique, et une hauteur, correspondant à la fréquence fondamentale. L'intensité de la voix, mesurée en décibels, constitue également une métrique importante. De plus, les silences entre les mots participent à la caractérisation de la parole. Ces différentes métriques, couramment utilisées en acoustique pour l'étude de la voix humaine, constituent des pistes intéressantes pour l'identification des sons naturels enregistrés. Dans cette étude, nous avons collecté 40 enregistrements naturels contenant chacun un cri d'oiseau (20 par espèce), anonymisés et mélangés en un même jeu de données. L'objectif est d'identifier, à partir de ces enregistrements, les signaux appartenant à la même espèce. Nous visons à classifier ces signaux en utilisant trois paramètres pertinents, sélectionnés au cours de l'analyse des données.

Peut-on, à partir du traitement et de l'analyse d'un ensemble de signaux naturels comprenant les cris du chardonneret élégant et de la perruche à collier, extraire trois paramètres pertinents permettant de classifier ces signaux avec efficacité ?

Pour répondre à cette problématique, nous étudierons tout d'abord les caractéristiques des signaux naturels enregistrés. Nous présenterons ensuite les méthodes mises en œuvre pour réduire les bruits parasites. Les deux stratégies employées pour sélectionner les trois paramètres de classification seront détaillées. Enfin, nous utiliserons ces paramètres pour regrouper les signaux en deux classes distinctes, puis analyserons les résultats obtenus.

1 Données utilisées

1.1 - Base de données

Les signaux constituant notre jeu de données proviennent de la base Xeno-canto, une plateforme collaborative dédiée à la collecte et au partage d'enregistrements de chants d'oiseaux. Chaque enregistrement y est annoté et validé par la communauté, garantissant la fiabilité des labels associés. La base fournit également des méta-données détaillées telles que la localisation géographique, la date d'enregistrement, l'altitude ainsi que le type de vocalisation (chant, cri d'alarme, appel, combat, etc...). Afin de limiter les variations acoustiques liées à la diversité géographique, nous avons retenu uniquement des enregistrements effectués en Europe.

Les signaux sélectionnés présentent une fréquence d'échantillonnage variable et une durée non homogène, nécessitant ainsi une phase de pré-traitement pour leur harmonisation. Les cris de la perruche à collier se caractérisent par un son aigu et puissant. Leur fréquence fondamentale est en moyenne centrée autour de 4000 ± 2500 Hz [2]. D'autre part, les vocalisations du chardonneret élégant présentent une fréquence fondamentale proche en moyenne de 4500 Hz [3].

Les enregistrements récupérés sont fortement bruités, principalement à cause des sons environnementaux (vent, sons réalisés par d'autres espèces et bruits urbains) [4]. On observe également la présence de silences entre les émissions sonores, qui peuvent constituer un élément distinctif du comportement vocal de chaque espèce [5].

1.2 - Observations et pré-traitements

1.2.1 - Sous-échantillonnage

Un des premiers éléments à prendre en compte pour le traitement de la base de données est le fait que les signaux ne possèdent pas tous la même fréquence d'échantillonnage. Certains signaux sont échantillonnés à 48000 Hz, d'autres à 44100 Hz et d'autres encore à 32000 Hz. Afin de faciliter le calcul de certains éléments, notamment au moment de la détection de motif, un sous-échantillonnage a été mis en oeuvre pour que tous les signaux soient échantillonnés à la plus petite fréquence d'échantillonnage au sein du jeu de données soit 32000Hz .

1.2.2 - Débruitage

Du fait de la nature des différents enregistrements, les signaux sélectionnés pour ce travail présentent une quantité plus ou moins importante de bruit. Afin de pouvoir appliquer les différents outils nécessaires à l'extraction de caractéristiques il est nécessaire de les débruiter. Plusieurs méthodes ont ainsi été mises en oeuvre. Un premier débruitage a été effectué à l'aide d'un filtre passe-bande proposant un découpage de 4000 à 5500 Hz. Ce filtre permet de conserver les fréquences autour des fréquences fondamentales décrites plus haut correspondant à la perruche à collier et au chardonneret élégant, mais également de corriger certaines erreurs potentiellement causées par les capteurs au moment de l'enregistrement.

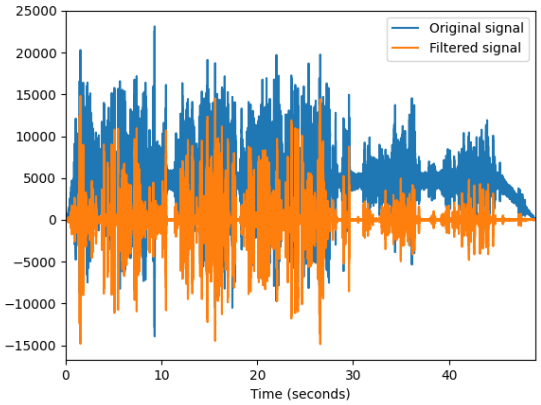


FIGURE 1.1 – Débruitage d'un signal avec un problème d'enregistrement

La figure 1.1 présente par exemple une moyenne globale autour de 5000 d'amplitude ainsi qu'une montée et une descente rapide de celle-ci au début et à la fin de l'enregistrement. Le filtre passe-bande permet alors de recentrer le signal autour de l'axe des abscisses et de supprimer une grande quantité de bruit. Cependant, la largeur de ce filtre passe-bande n'est pas adaptée à tous les signaux. Comme indiqué plus haut, la fréquence fondamentale du collier étant de 4000 Hz *en moyenne*, certains signaux possèdent de l'information à conserver à des fréquences en dessous de 4000 Hz. Un nouveau filtre passe-bande a donc été mis en place, cette fois de manière dynamique afin de s'adapter au mieux à chaque signal.

1.2.3 - Filtrage dynamique

Un filtrage dynamique a été mis en place pour s'adapter à chaque signal. Plusieurs paramètres ont été définis pour tout de même cadrer le filtrage. Celui-ci s'opère en deux temps :

— Pour le premier filtrage les fréquences conservées se situent nécessairement entre 3500 ± 50Hz et 8000 ± 50Hz. Afin de déterminer dynamiquement la bande de fréquence du filtre, l'amplitude maximale est d'abord identifiée dans le domaine de Fourier (en excluant celle à l'origine). Puis, la première limite de bande correspond à la première amplitude détectée atteignant au moins 20% de l'amplitude maximale. La dernière fréquence de la bande est alors la dernière fréquence détectée atteignant également 20% de l'amplitude maximale.
— Pour le second filtrage, les fréquences conservées se situent nécessairement entre 3500 ± 50Hz et 6000 ± 50Hz. La bande de fréquence du filtre est composée de la première fréquence détectée atteignant 50% de l'amplitude maximale, et la dernière fréquence de la bande est alors la dernière détectée à atteindre 50% de l'amplitude maximale.

Ce filtrage en deux temps permet plusieurs choses. Tout d'abord, le fait de passer par deux étapes de filtrage permet de minimiser le bruit résiduel au niveau des extrémités de bande (Figure 1.2). Ensuite, utiliser un filtre passe-bande centré sur les fréquences fondamentales moyennes observée permet d'ignorer des sources parasites parfois visibles après une Transformée de Fourier et qui seraient potentiellement captées par un filtre coupe-bande. De plus, un filtre passe-bas aurait laissé passer trop de bruit en basses fréquences, étant donné que les cris d'oiseaux emploient majoritairement les hautes fréquences. Enfin, un filtre passe-haut aurait pu être envisagé, mais certains signaux présentent du bruit dans de très hautes fréquences comme pour la figure 1.3. Le filtre passe-bande semble donc être le plus adapté au vu de la base de données.

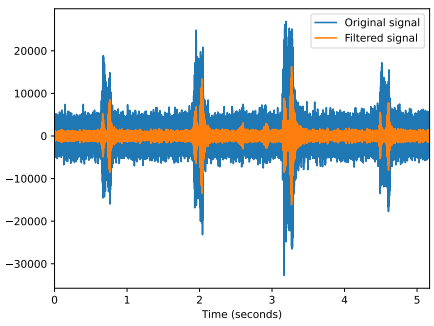


FIGURE 1.2 – Signal reconstruit à partir des fréquences filtrées

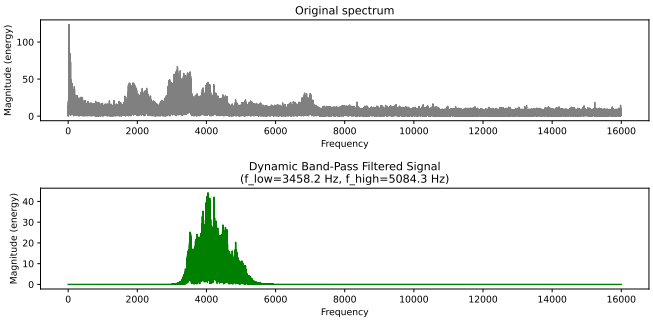


FIGURE 1.3 – Filtre passe-bande appliqué sur un signal bruité dans les hautes fréquences dans le domaine de Fourier

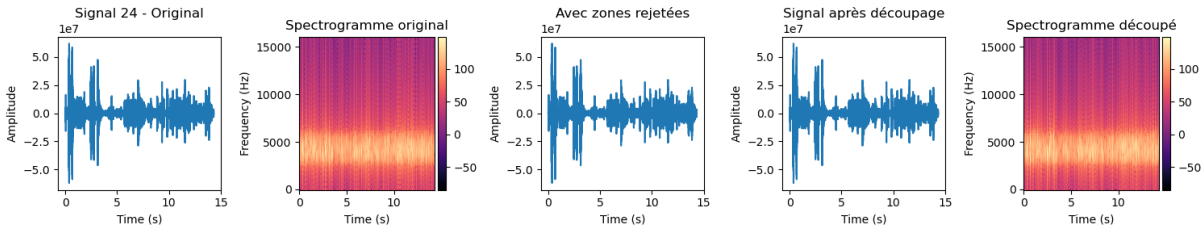
1.2.4 - Analyse de la stationnarité des signaux

Afin de réaliser une analyse rigoureuse des signaux, il est nécessaire de vérifier s'il suivent la stationnarité au sens large. Nous faisons l'hypothèse que le critère d'ergodicité s'applique à ces signaux, ceux-ci étant suffisamment longs. La stationnarité d'ordre 1 implique que la moyenne du signal reste constante au cours du temps, tandis que la stationnarité d'ordre 2 suppose une variance également constante (ou présentant de faibles variations temporelles). La stationnarité au sens large d'un signal peut être déterminée en observant son spectrogramme. Ainsi, si au cours du temps les fréquences caractérisant le signal et leur intensité restent identiques, alors on considère un signal comme stationnaire au sens large. Pour certains signaux, le spectrogramme présente une répartition en fréquence homogène dans le temps (Figure 1.4a), indiquant une stationnarité au sens large. Dans ce cas, aucune modification du signal n'est nécessaire. D'autres signaux présentent des discontinuités dans le spectrogramme, notamment liées à la présence de silences entre les émissions sonores. Afin d'homogénéiser ces spectrogrammes, nous avons cherché à supprimer les zones de silence. Une première approche visuelle globale des signaux bruts révèle une variance du signal non constante au cours du temps. La suppression des silences les plus longs ne parvient pas totalement à enlever ces variations.

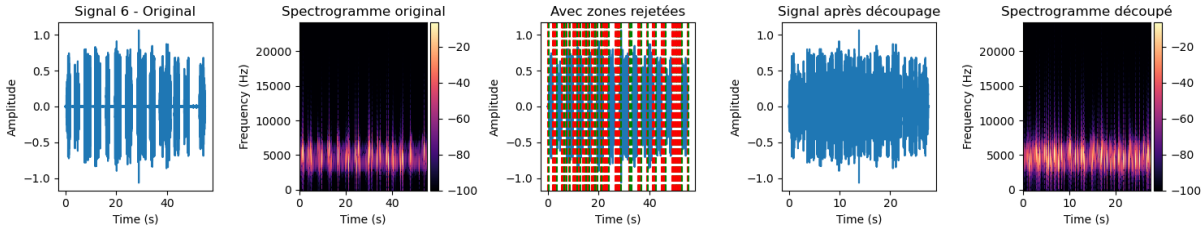
1.2.5 - Suppression des silences et découpage audio

De nombreux enregistrements présentent de longs silences entre certaines salves de chant d'oiseaux. Ces silences sont distincts de ceux présents au sein du chant, caractérisant la fin d'une syllabe ou d'un mot. Afin de supprimer ces silences, deux techniques ont été mises en place.

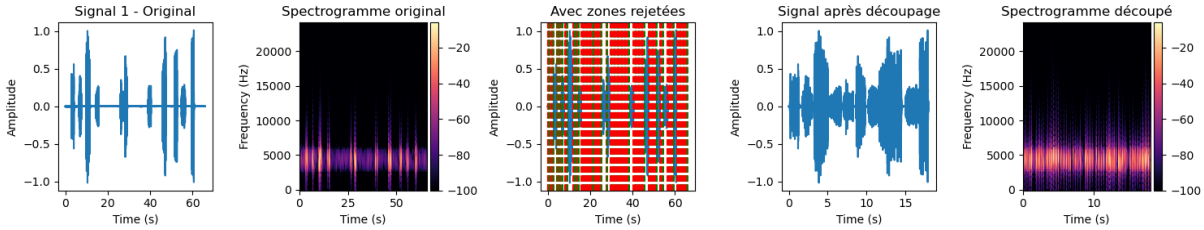
La première méthode est basée sur la détection de ruptures à partir de la mesure de la moyenne du signal (en considérant uniquement sa partie positive). En effet, les zones silencieuses se caractérisent par une moyenne d'intensité plus faible que les zones contenant un cri. Toutefois, les algorithmes de détection de rupture se sont révélés inefficaces sur les signaux, ceux-ci étant de grande taille et exigeant un temps de calcul trop important. Pour contourner cette limitation, une méthode de découpage adaptatif a été mise en oeuvre. Le signal positif est d'abord divisé en sous-séquences, puis la moyenne de chaque sous-séquence est calculée. Une moyenne globale est ensuite estimée à partir de ces valeurs. Les sous-séquences dont la moyenne est inférieure à un seuil relatif à la moyenne globale sont supprimées. Les paramètres de découpage (taille des sous-séquences et seuil de suppression) sont ajustés individuellement pour chaque signal. Cette approche permet d'obtenir des signaux présentant un spectrogramme plus homogène, satisfaisant ainsi la condition de stationnarité d'ordre 1.



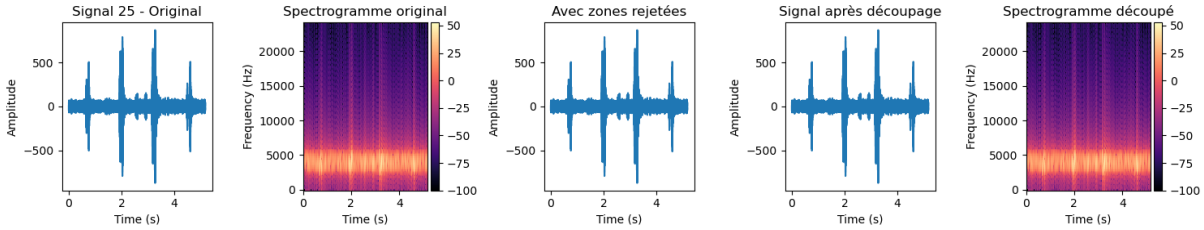
(a) Signal non modifié



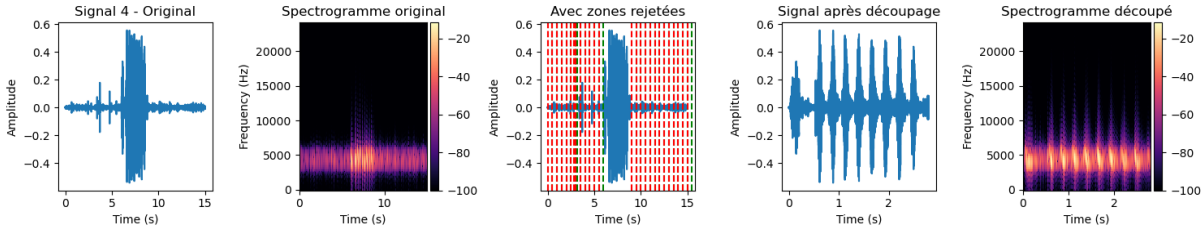
(b) Signal modifié sans silence



(c) Signal modifié avec reste de silence



(d) Signal bruité



(e) Signal avec variation d'intensité

FIGURE 1.4 – Audiogramme et spectrogramme des signaux originaux et des signaux après coupures (lignes rouges) réalisées en mesurant des différences de moyenne sur la partie positive du signal

Les signaux obtenus sont dépourvus de blancs et mieux adaptés à l'analyse. Cette méthode permet de bien supprimer les blancs (Figure 1.4b). Pour certains signaux, il reste toujours de petites coupures (Figure 1.4c). De plus, certains signaux sont encore trop bruités ce qui les rends insensibles à cette méthode(Figure 1.4d). Enfin, après traitement, on peut arriver à enlever les blancs mais on observe des variations d'intensité au sein du signal. Ces signaux sont alors stationnaires d'ordre 1 mais pas d'ordre 2 (Figure 1.4e).

Une autre méthode de découpage des signaux a également été mise en place de la manière suivante : les silences les plus longs sont détectés en prenant en compte l'amplitude du signal ainsi que la durée des faibles amplitudes. L'amplitude est déterminée ici en dB. Un silence "long" correspond à une portion de signal ayant une amplitude inférieure à 18 dB pour une durée supérieure à 0.5 s. Ces paramètres permettent de supprimer les silences entre les salves de chant sans interférer avec les silences au sein de ces salves. La figure 1.5 illustre cette suppression de longs silences. Les portions restantes du signal sont ensuite concaténées. Cette méthode apparaît comme plus efficace car elle prend en compte la durée des silences et évite d'interférer avec les silences au sein des salves.

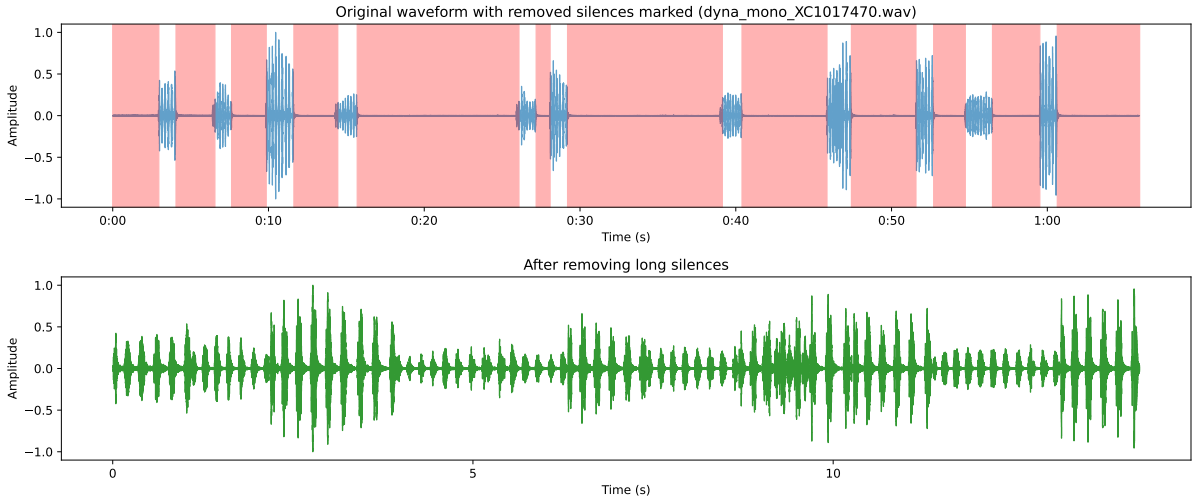


FIGURE 1.5 – Exemple de suppression de silences longs

Enfin, pour réaliser une analyse en composantes principales (ACP) qui sera utile par la suite, il est nécessaire que l'ensemble des signaux soient de la même taille. À cette fin, une procédure de normalisation de la taille des signaux a été adoptée. La méthode consiste à identifier la longueur du signal le plus court, puis à extraire, pour chaque enregistrement, un segment de même nombre d'échantillons. Afin d'optimiser la conservation de l'information acoustique pertinente lors de cette opération, le point de départ du découpage est déterminé de manière adaptative : nous recherchons dans chaque signal la première position où l'intensité dépasse 10 pourcent du maximum d'intensité du signal mesuré, puis extrayons le segment requis à partir de ce point. Cette approche permet de réduire la variabilité liée aux portions de silence restantes. En ce qui concerne la taille des signaux pour l'extraction de motifs, celle-ci a été fixée à 5 s pour chaque signal à l'exception de ceux pouvant avoir une durée plus courte qui reste alors inchangée.

2 Démarche utilisée pour sélectionner des métriques de classification de nos signaux

2.1 - Détection de silences au sein des salves de chant

Durant l'observation des signaux, il est apparut que certains d'entre eux semblaient présenter des motifs récurrents caractérisés par des salves sonores très brèves en grande quantité ou bien des salves plus larges avec des pauses plus longues entre chaque pic sonore. Dans le but de caractériser la dynamique temporelle des vocalisations, il a été décidé de s'intéresser à la durée des pauses entre les cris des oiseaux. À cette fin, l'algorithme de détection des longs silences précédemment utilisé a été adapté. Cet algorithme modifié a alors permis d'identifier les transitions entre les segments sonores et les périodes de silence.

La classification des signaux basée sur la durée des silences est effectuée de la manière suivant :

- Les périodes de silence sont d'abord détectée de la même manière que celle décrite plus haut, à la différence que cette fois-ci, le seuil d'amplitude à conserver se situe en-dessous de 25 dB.
- La longueur des silences est déterminée en nombre d'échantillons. Les silences conservés sont ceux contenant un nombre d'échantillons strictement inférieur à 6000. Ce nombre a été choisi après observation des salves de chant et semblait proposer un découpage cohérent par rapport aux signaux comme montré sur la figure 2.1. Seuls les signaux conservés sont comptabilisés.
- Le nombre de silences conservés par signal permet d'envisager une classification entre des signaux présentant un grand nombre de silences courts contre des signaux en contenant très peu voire pas du tout.

Il est important de noter que pour cette extraction des silences courts, la taille des signaux n'a pas été normalisée. Cette décision a été prise suite au fait que la suppression des silences longs réduit déjà particulièrement la taille des signaux. Le signal le plus long avant suppression des silences était d'environ 2 min, tandis qu'avec cette suppression, ce maximum passe à 40 s et ne représente pas l'entièreté des signaux. La médiane de temps se trouve à 12 s et la moyenne à 15 s. Il apparaît donc possible de considérer que la différence de durée entre les signaux n'induit pas de biais trop important.

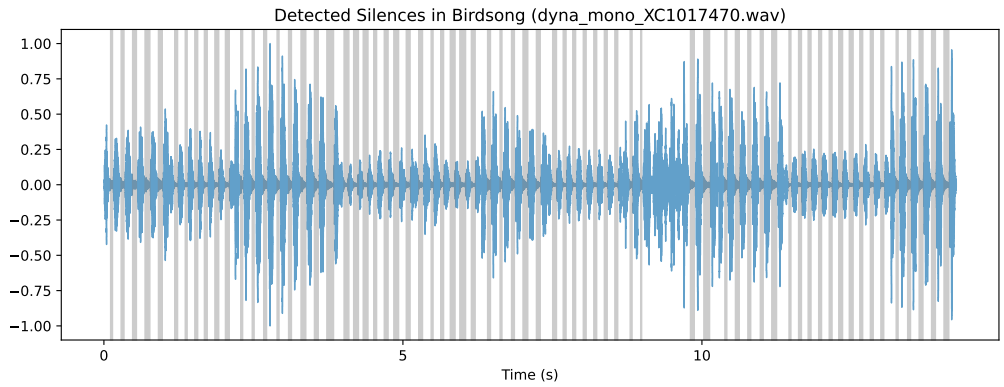


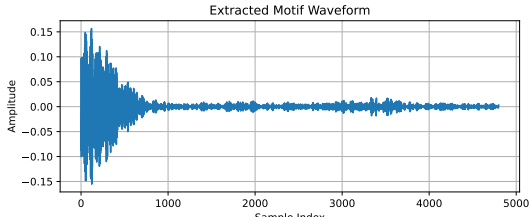
FIGURE 2.1 – Détection de silences courts (en gris) pris en compte pour la classification

2.2 - Détection et extraction de motifs

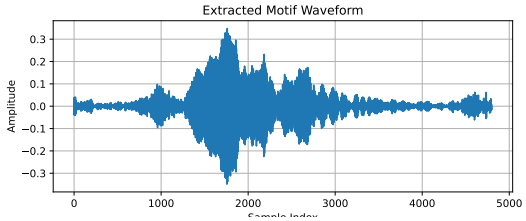
L'observation de certaines formes répétitives au sein des signaux a mené à l'idée d'extraire un motif pour chaque signal afin de procéder à une classification des signaux. Si certains motifs s'illustrent comme représentatifs d'une des deux classes, il semble alors possible d'opérer une classification des motifs en comparant la corrélation maximum de chaque signal avec chaque motif.

Afin de pouvoir extraire un motif par signal, il faut d'abord procéder à un raccourcissement de ces derniers pour éviter un temps de calcul trop important au moment de l'extraction. Comme indiqué précédemment, les signaux sont raccourcis à une taille de 5 secondes maximum afin de diminuer le temps de traitement. Ce découpage pourrait biaisé la recherche de motif si trop peu de chant d'oiseau est présent au sein des 5 secondes choisies, mais la suppression des silences effectuées précédemment résout une partie de ce problème en concentrant l'information sonore.

Une extraction de motif efficace repose sur la détermination de la durée d'un motif. Les salves de chant présentant des pics très brefs, il apparaît tout de suite clair que la taille du motif ne peut pas dépasser 0.5 secondes. D'après la bibliographie, la durée d'une syllabe du chardonneret élégant dure en moyenne 0.116 s [3]. Avec cette information en tête, la recherche de motif a été faite pour des motifs de 0.15s afin d'avoir une petite marge d'erreur. La taille du motif recherché est traduite en nombre d'échantillons. Ainsi, 40 motifs d'une taille de 4800 échantillons ont été extraits. Les motifs extraits ne semblent pas tous bien caractériser un signal. Ainsi, certains motifs comme celui de la figure 2.2b semblent plus représentatifs d'un signal que d'autres motifs comme celui de la figure 2.2a qui contiennent essentiellement du bruit. La suppression des silences comme précédemment expliqué ne semble pas suffire à enlever les silences dans les motifs.



(a) Exemple de motif présentant du bruit



(b) Exemple de motif isolé correctement

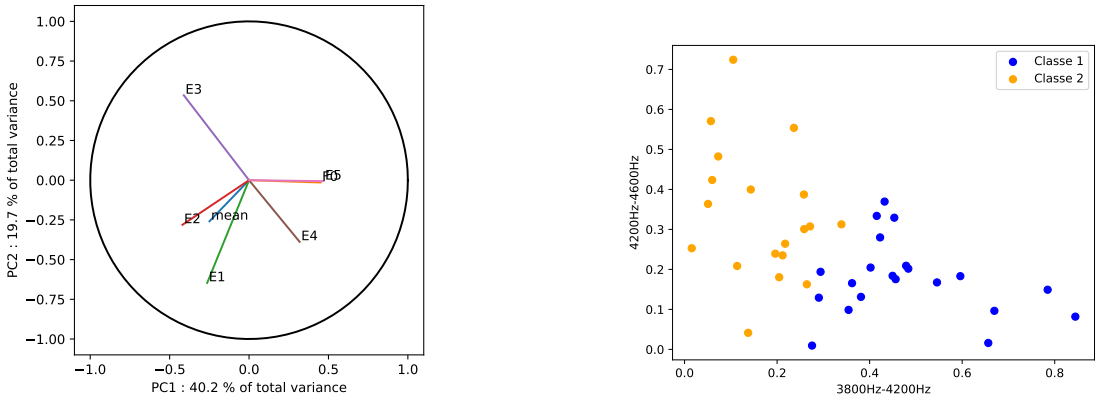
FIGURE 2.2 – Sélection de motifs présentant des niveaux de bruit plus au moins grands

La corrélation de chaque signal pour chaque motif a ensuite été calculée. Une classification de nos signaux basée uniquement sur ce critère est peu fiable car la présence de silence dans certains motifs augmente fortement leur corrélation avec l'ensemble des signaux. Malgré plusieurs tentatives et changement de paramètre, la démarche décrite ici semble être la meilleure, et également la plus simple. Les motifs extraient pour chaque signal ont tout de même été conservés en tant que métrique finale étant donné que certains d'entre eux semblent tout de même caractéristiques d'une classe comme sur la figure 2.2b.

2.3 - Analyse en Composantes Principales

Pour sélectionner d'autres critères de classification, une analyse en composantes principales a été appliquée avec des métriques calculées sur les signaux. Parmi celles-ci, la moyenne de la partie positive du signal ainsi que l'énergie relative répartie sur cinq bandes de fréquences ont été étudiées.

La perruche à collier est connue pour produire des signaux plus intenses que ceux du chardonneret élégant. Il est donc pertinent d'analyser la moyenne d'amplitude du signal positif, qui reflète l'intensité sonore globale et peut contribuer à la discrimination entre espèces. Par ailleurs, les deux espèces étudiées se distinguent principalement par leur hauteur et leur timbre. La hauteur correspond à la fréquence fondamentale du signal, qui, selon la littérature, est d'environ 4000 Hz pour la perruche à collier et 4500 Hz pour le chardonneret élégant. L'extraction de cette fréquence fondamentale pour chaque signal constitue donc un paramètre intéressant pour la classification. Le timbre, quant à lui, est lié à la composition harmonique du signal. L'énergie relative, calculée sur des bandes de fréquences définies, permet d'estimer cette composition harmonique et de quantifier la contribution de chaque bande au signal. Nous avons choisi de centrer ces bandes entre 3000 Hz et 5000 Hz, afin de couvrir les plages de fréquences les plus pertinentes pour la différenciation des espèces.



(a) ACP des signaux selon 7 métriques

(b) Classification des signaux selon E3 et E4

FIGURE 2.3 – Sélection d'une métrique pour la classification des signaux à partir d'une ACP.

L'interprétation des résultats, se concentre sur la première composante principale (PC1). Cette dernière, explique la proportion la plus importante de la variance totale et permet ainsi de classer les métriques selon leur capacité à classer les signaux. On observe sur la figure 2.3a, que la moyenne du signal positif ne permet pas de séparer les signaux. En effet, cette métrique présente une explicabilité de la variance très faible par rapport aux autres. Cette limitation est vraisemblablement liée à la variabilité des distances entre les enregistreurs et les oiseaux. Les différences d'intensité observées entre les signaux pourraient ainsi refléter principalement des effets de distance plutôt que des caractéristiques intrinsèques des espèces. Pour que cette métrique soit pertinente, il faudrait connaître exactement la distance entre l'enregistreur et l'oiseau, ce qui complique l'identification des oiseaux notamment dans des milieux naturels peu accessibles.

La PC1 semble plus fortement corrélée aux bandes d'énergie E2 et E5. Ces résultats suggèrent que les bandes d'énergie renferment une information acoustique plus discriminante pour la différenciation des signaux que la fréquence fondamentale (F0). Cependant, il est intéressant de noter que les bandes E3 et E4 semblent négativement corrélées, ce qui pourrait indiquer l'appartenance d'une classe des signaux à l'une des bandes ou l'autre. Il est également bon de noté que les bandes E3 et E4 pourraient correspondre respectivement au collier et au chardonneret, de part les fréquences qu'elles englobent (3800–4200 Hz pour E3 et 4200–4600 Hz pour E4). Les bandes E3 et E4 sont donc retenues pour poursuivre les recherches. Lorsque les signaux sont représentés selon les deux métriques retenues pour la classification (Figure 2.3b), leur distribution révèle la formation de deux groupes semblant distincts et homogènes. Plus précisément, 19 signaux présentent une intensité énergétique plus élevée dans la bande E4 que dans la bande E3, tandis que 21 signaux montrent la tendance inverse, avec une énergie dominante dans E3. Étant donné que le jeu de données comprend 20 signaux par espèce d'oiseau, cette répartition apparaît cohérente avec la structure attendue des données. Ces observations suggèrent que l'énergie relative entre les bandes E3 et E4 constitue un indicateur pertinent pour la différenciation des signaux et peut être utilisée comme métrique efficace dans la classification acoustique.

2.4 - Tentatives infructueuses

Une tentative d'ACP portant sur les motifs extraits a été tentée. Le but était de classer les motifs indépendamment des signaux pour ensuite extrapoler cette classification sur les signaux dont ils étaient tirés. L'intérêt de cette méthode venait de la taille réduite des signaux qui permettait un traitement plus rapide. Cependant, ce traitement plus rapide s'opérait au dépend d'une quantité d'information suffisante. En effet, à cause de la taille réduite des signaux, mais surtout de la quantité de bruit malgré tout présente en leur sein, cette classification s'est révélée inefficace, et n'a donc pas été retenue.

3 | Solution choisie

À la suite de plusieurs essais exploratoires, trois métriques ont été retenues pour la classification des signaux acoustiques. Ces métriques ont été sélectionnées en raison de leur capacité à capturer des variations pertinentes dans la structure des signaux. Étant donné que la catégorisation ne repose sur aucune information à priori concernant les classes, la classification finale se fait de manière non supervisée. La cohérence du résultat par rapport au jeu de données d'origine est ainsi uniquement basée sur le nombre de signaux présent par classe, soit 20 par classe.

Dans ce cadre, l'algorithme K-Means, dont l'objectif est de regrouper les observations en k-groupes en minimisant la variance intra-classe a été appliqué. Le nombre de groupes a été fixé à deux, conformément au nombre d'espèces d'oiseaux présentes dans le jeu de données. L'algorithme procède de manière itérative en assignant chaque signal au centroïde le plus proche, selon les trois métriques sélectionnées, puis en mettant à jour la position de ces centroïdes afin d'optimiser la distribution des groupes formés. Les centroïdes finaux obtenus représentent ainsi les positions moyennes les plus représentatives de la distribution des signaux dans l'espace défini par les trois métriques.

3.1 - Métriques choisies

Selon les observations précédemment réalisées, les métriques suivantes ont été sélectionnées :

— Les bandes d'énergie E3 (3800–4200 Hz) et E4 (4200–4600 Hz)
— Le nombre de silences courts entre les cris des oiseaux
— Les motifs extraits des signaux

Pour effectuer la classification selon ces trois métriques, les résultats ont été compilés pour chaque signal dans une matrice, sur laquelle l'algorithme K-Means a ensuite été appliqué.

3.2 - Démarche complète

Le jeu de données est en premier lieu anonymisé. Ensuite, les fichiers sont tous convertis au format wav. Les fichiers présentant des enregistrements stéréo sont transformés en mono en conservant une seule de leur deux pistes. Puis, ces fichiers sont remaniés en numpy array et exportés dans un fichier au format npz. Tout le reste de la démarche se base sur ce fichier.

Après avoir importé les bibliothèques python nécessaires pour le traitement du jeu de données, les signaux sont récupérés depuis le fichier npz. Les valeurs d'échantillons sont stockées dans une liste 'signals' et les fréquences d'échantillonnage associées dans une liste 'sample_rates'. Les signaux sont aussi tous affichés en même temps pour vérifier que le jeu de données est complet.

Afin de ne pas rencontrer de problème dans la suite du traitement, les signaux subissent un sous-échantillonnage. Avec l'analyse préalable du jeu de données, il a été établi que la fréquence d'échantillonnage la plus faible est de 32000 Hz. Cette fréquence d'échantillonnage est alors choisie comme fréquence cible pour tous les autres signaux. La fréquence d'échantillonnage de chaque signal est prise en compte en observant la liste 'sample_rates' et les signaux sont ré-échantillonnés grâce à la fonction resample() de la bibliothèque SciPy puis ajoutés à une nouvelle liste 'resampled_signals'. La fréquence d'échantillonnage prise en compte pour tous les signaux est alors de 32000 Hz pour le reste du traitement.

Á ce stade, les signaux contiennent toujours leur bruit d'origine. Un débruitage dynamique est donc appliqué sur les signaux de la liste 'resampled_signals'. Les fonctions de filtrage sont d'abord définies :

- Une fonction de filtre passe-bande prenant en argument un signal, une bande de fréquences et la fréquence d'échantillonnage du signal.
- Une fonction de détection dynamique de la bande de fréquences à conserver pour le filtre passe-bande. Cette fonction prend en argument le signal, sa fréquence d'échantillonnage, un seuil de détection des fréquences, une marge à appliquer à chaque extrémité de la bande de fréquence, un seuil de fréquences minimales à exclure, et un seuil de fréquences maximales à exclure. La bande de fréquences est choisie en fonction des pics de fréquence observés après une Transformée de Fourier du signal. La fonction applique d'abord une Transformée de Fourier sur le signal. La zone d'analyse est ensuite réduite en fonction des seuils de fréquences définis. Les fréquences où l'énergie est la plus forte sont détectées, puis une bande est centrée sur ces fréquences, et élargie d'une marge. La fonction applique ensuite le filtre passe-bande en prenant en compte la bande de fréquence établie.

Ce filtrage dynamique est ainsi appliqué à l'ensemble des signaux. Chaque signal filtré est ajouté à une nouvelle liste : 'denoised_signals'. Tous les signaux filtrés sont ensuite affichés pour les comparer avec les signaux non-filtrés.

Les signaux subissent ensuite une suppression des longs silences en leur sein. Deux paramètres sont pris en compte : un seuil de décibel fixé à 18 dB et un seuil de durée fixé à 0.5 s. Pour chaque signal, les intervalles de silence sont détectées selon ces deux paramètres puis extraites. Le signal est ensuite reconstruit en concaténant les intervalles restantes et ajouté à la liste 'trimmed_signals'. Cette étape affiche également le découpage effectué par signal. Les zones supprimées apparaissent grisées sur le signal original, et le signal reconstruit est ensuite affiché pour visualiser l'effet de la concaténation après découpage.

Les prochaines étapes visent à préparer les signaux pour une Analyse en Composantes Principales (ACP). Pour ce faire, les signaux doivent tous avoir la même taille, et doivent donc être redimensionnés. Pour chaque signal, on recherche d'abord le premier échantillon où l'amplitude dépasse 10% d'un maximum détecté au préalable et cette position est enregistrée. Ensuite, le signal avec la plus petite longueur est détecté et cette longueur est choisi comme longueur cible. Une nouvelle liste de signaux 'court' est créée en ajoutant des versions tronquées de chaque signal, toutes de la même taille et alignées à partir du moment où le signal dépasse le seuil d'amplitude défini.

Plusieurs fonctions sont définies pour extraire les métriques à utiliser pour l'ACP :

- Une fonction pour extraire tous les échantillons positifs d'un signal.
- Une fonction calculant la répartition de l'énergie au sein de plusieurs bandes de fréquences. Cette fonction prend en argument un signal, un nombre de bandes de fréquences, la fréquence d'échantillonnage du signal, les bornes d'intérêt. Elle effectue d'abord une Transformée de Fourier pour obtenir le spectre du signal, puis n'en conserve que les fréquences positives dans la bande d'intérêt. Cette bande est découpée selon le nombre de sous bandes souhaitées, et l'énergie contenue dans chacune est calculée. L'ensemble est ensuite normalisé pour obtenir des énergies relatives (leur somme valant alors 1). La fonction retourne l'énergie relative dans chaque bande, les bornes fréquentielles correspondantes, ainsi que la fréquence fondamentale du signal.
- Une fonction pour calculer les métriques suivantes pour chaque signal : la moyenne des échantillons positifs, l'énergie du signal sur 5 bandes de fréquences et la fréquence fondamentale.
- Une fonction pour réaliser l'ACP.
- Une fonction pour afficher le résultat de cette ACP avec un cercle de corrélation.

L'ACP est réalisée à l'aide des fonctions précédentes et le cercle de corrélation est affiché. L'importance de chaque métrique est également calculée et affichée. Comme indiqué plus haut, les métriques choisies pour la suite de l'analyse sont E3 (3800-4200 Hz) et E4 (4200-4600 Hz) qui semblent présenter les caractéristiques les plus intéressantes compte tenu du jeu de données étudié.

Un algorithme de K-means est alors appliqué sur ces données pour vérifier la cohérence de la sélection de E3 et E4 avec une classification automatique. Cette classification cherche à établir deux classes, et afin de garantir que cette classification puisse être reproduite, une valeur de random state a été précisée. La classification affiche alors 21 points dans la classe 1 et 19 dans la classe 2, ce qui semble cohérent avec le jeu de données de départ. Les énergies de chaque signal au sein des bandes E3 et E4 sont ainsi combinées en une matrice représentant la première métrique extraite : 'relative_energy_feature'.

L'étape suivante reprend les signaux de la liste 'trimmed_signals' et compte le nombre de silences courts détectés pour chaque signal. La liste 'trimmed_signals' est réutilisée parcequ'elle contient plus d'informations d'intérêt que les signaux redimensionnés. Comme pour l'extraction de silences longs, cette détection se base sur deux paramètres : un seuil de décibel fixé à 25 dB ainsi qu'une durée à ne pas dépasser qui est équivalente à 6000 échantillons. Á l'aide de la bibliothèque Librosa, les intervalles non silencieuses de plus de 25 dB sont détectées. On en déduit les portions silencieuses en conservant les segments entre les parties sonores. La durée de chaque silence est ensuite calculée, et seuls les silences de moins de 6000 échantillons sont pris en compte. Pour chaque signal, on calcul ensuite le nombre de silences conservés et ce nombre enregistré pour chaque signal constitue la métrique 'silence_feature'.

La prochaine étape permet l'extraction, pour chaque signal, d'un motif. La liste 'trimmed_signals' est de nouveau utilisée. Cependant, la taille des signaux est réduite par l'algorithme de détection de motif pour diminuer le temps de calcul du *matrix profile*. Cette section de code cherche également à déterminer si un GPU est à disposition pour accélérer le calcul qui s'avère très long sur CPU. Si un GPU est détecté, le code s'exécute dessus. La durée maximale de chaque signal est d'abord limitée à 5 s pour réduire le temps de calcul, puis une fenêtre d'analyse de longueur L est définie. Cette longueur L correspond à la taille du motif recherché qui est de 0.15 s. Par la suite la taille du motif est reconvertie en nombre d'échantillons. Le profil de similarité du signal est calculé soit sur CPU avec stumpy.stump() soit sur GPU avec stumpy.gpu_stump(), la bibliothèque Stumpy étant dédiée au calcul de *matrix profile*. Le minimum atteint par ce profil correspond à la portion la plus représentative (le "meilleur motif") du signal. Cette portion est extraite et enregistrée dans la liste 'best_motifs. Cette liste de motif correspond à la troisième métrique extraite pour la classification des signaux. Avant de continuer, les motifs sont tous affichés en même temps afin de vérifier que l'extraction s'est bien déroulée.

Une fois les 3 métriques extraites, celles-ci sont combinées en une matrice de caractéristiques. Pour chaque signal, le nombre de silences (feature_1), les énergies relatives dans E3 et E4 (feature_2) et le motif extrait (feature_3) sont combinés. Ces informations sont ensuite converties en tableaux numpy, puis concaténées pour former un seul vecteur de caractéristiques par signal. Tous ces vecteurs sont assemblés dans une matrice X. Cette matrice est alors exportée en fichier npz. Enfin, la matrice contenue dans ce fichier npz est utilisée pour classifier tous les signaux à l'aide d'un algorithme K-Means.

3.3 - Résultats

Après application de l'algorithme K-Means, les signaux sont répartis en deux classes. La classe 1 compte 18 signaux, et la classe 2 en compte 22. Ce résultat est proche des attentes de 20 signaux par classes, ce qui indique une possible réussite de la tâche de classification. Cela montre également que les caractéristiques extraites pour chacun des signaux étaient bel et bien représentatives d'une classe. L'implémentation complète de ce travail est disponible sur Github.

Conclusion

En conclusion, cette étude a permis de classifier des signaux enregistrés dans des environnements naturels selon deux espèces d'oiseaux. Les signaux étant enregistrés en conditions réelles, un débruitage a d'abord été appliqué via un filtre passe-bande, conservant uniquement les fréquences d'intérêt. Ces signaux se caractérisent par la répétition de motifs, et l'extraction de ceux-ci a permis d'identifier, pour chaque signal, un motif représentatif correspondant à une première métrique. Comme chez les humains, la communication des oiseaux repose également sur la présence de silences entre les vocalisations. La longueur de ces silences étant spécifique à chaque espèce, la deuxième métrique choisie correspond au nombre de silences courts par signal. Enfin, les signaux se distinguent par leur timbre, défini par leur composition en harmoniques. La dernière métrique retenue est donc l'énergie relative de deux bandes de fréquences, déterminées par une ACP comme représentant au mieux la variance du jeu de donnée. L'ensemble de ces caractéristiques a permis de regrouper les signaux en deux classes distinctes, reflétant les différences acoustiques entre les deux espèces étudiées.

Dans cette étude, la sélection du jeu de données repose sur un critère de proximité géographique des enregistrements. Ceci permet de centrer notre méthode de classification sur les caractéristiques vocales des espèces présentes en Europe. Cependant, cette sélection ne prend pas en compte les caractéristiques individuelles des oiseaux émettant ces signaux. Ainsi, il est possible que les échantillons étudiés ne reflètent pas pleinement la diversité des vocalisations naturelles de chaque espèce. En effet, les propriétés acoustiques des signaux peuvent varier en fonction de l'âge, du sexe ou du comportement des oiseaux, entraînant une variabilité importante qui complique leur identification et leur classification. Pour renforcer la robustesse et la généralisation de notre approche, il serait pertinent de tester les métriques définies sur des jeux de données plus représentatifs de la diversité des vocalisations observées dans la nature.

Enfin, en ce qui concerne l'extraction de motifs, la méthode utilisée tend à sélectionner des motifs contenant de longues portions de silence, malgré la suppression de nombre d'entre eux. Or, ces silences n'apportent pas d'information discriminante entre les espèces. Il serait donc pertinent d'introduire des contraintes supplémentaires dans l'algorithme d'extraction afin de pénaliser les motifs présentant des séquences de silence trop longues, et ainsi favoriser l'identification de motifs réellement informatifs pour la classification.

Bibliographie

[1] *Arrêté du 29 octobre 2009 fixant la liste des oiseaux protégés sur l'ensemble du territoire et les modalités de leur protection*. Journal Officiel de la République Française. 2009.

[2] M. M. MOGHAL et al. « Bird Calls Frequency Distribution Analysis to Correlate with Complexity of Syrinx ». In : *Journal of Advanced Scientific Research* (2015).

[3] G. C. CARDOSO, Y. HU et P. G. MOTA. « Birdsong, sexual selection, and the flawed taxonomy of canaries, goldfinches and allies ». In : *Animal Behaviour* (2012).

[4] Christine ERBE et Jeanette A. THOMAS. « Exploring Animal Behavior Through Sound ». In : *Proceedings of the Acoustical Society of America Meeting*. 2022.

[5] T. JOSE et J. A. MAYAN. « Real-Time Sound Detection of Rose-Ringed Parakeet Using LSTM Network with MFCC and Mel Spectrogram ». In : *Annual International Conference on Emerging Research Areas : International Conference on Intelligent Systems*. 2023.