



HAL
open science

Classification non-supervisée des productions vocales chez le bébé humain entre 0 et 12 mois

Guillem Bonafos

► **To cite this version:**

Guillem Bonafos. Classification non-supervisée des productions vocales chez le bébé humain entre 0 et 12 mois. Statistiques [math.ST]. Aix-marseille University, 2023. Français. NNT : 2023AIXM0487 . tel-04607513

HAL Id: tel-04607513

<https://hal.science/tel-04607513v1>

Submitted on 10 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0
International License

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 11 Décembre 2023 par

Guillem BONAFOS

Classification non-supervisée des productions vocales chez le
bébé humain entre 0 et 12 mois

Discipline

Mathématiques

École doctorale

ED 184 Mathématiques et Informatique

Laboratoire/Partenaires de recherche

Institut de Mathématiques de Marseille (UMR 7373)

Laboratoire de Psychologie Cognitive (UMR 7290)
Résurgences R&D

Composition du jury


Bertrand MICHEL	Rapporteur
Professeur, École Centrale de Nantes	
Florence LEVRERO	Rapporteuse
Maîtresse de Conférence, Université Jean Monnet	
Marianne CLAUSEL	Examinatrice
Professeure, Université de Lorraine	
Vincent VANDEWALLE	Président du jury
Professeur, Université Côte d'Azur	
Pierre PUDLO	Directeur de thèse
Professeur, Université d'Aix-Marseille	
Jean-Marc FREYERMUTH	Codirecteur de thèse
Maître de Conférence, Université d'Aix-Marseille	
Samuel TRONÇON	Encadrant de thèse
Docteur, Résurgences R&D	
Arnaud REY	Membre invité
Chargé de recherche, CNRS - Université d'Aix-Marseille	

Affidavit

Je soussigné, Guillem Bonafos, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Pierre Pudlo, Jean-Marc Freyermuth, Samuel Tronçon et Arnaud Rey, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le 9 Octobre 2023



Cette œuvre est mise à disposition selon les termes de la [Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Liste de publications et participation aux conférences

Liste des publications et/ou brevets réalisées dans le cadre du projet de thèse :

1. L. Tosatto, **G. Bonafos**, J-B. Melmi, A. Rey. 2022. Detecting Non-Adjacent Dependencies Is the Exception Rather than the Rule. PLOS ONE, 17(7) :e0270580. doi :10.1371/journal.pone.0270580.
2. A. Rey, J. Fagot, F. Mathy, L. Lazartigues, L. Tosatto, **G. Bonafos**, J-M. Freyermuth, F. Lavigne. 2022. Learning Higher-Order Transitional Probabilities in Nonhuman Primates. Cognitive Science, 46(4). doi :10.1111/cogs.13121.
3. A. Rey, L. Bogaerts, L. Tosatto, **G. Bonafos**, A. Franco, B. Favre. 2020. Detection of Regularities in a Random Environment. Quarterly Journal of Experimental Psychology. doi :10.1177/1747021820941356.
4. **G. Bonafos**, P. Pudlo, J-M. Freyermuth, T. Legou, J. Fagot, S. Tronçon, A. Rey. 2023. Detection and classification of vocal productions in large scale audio recordings. doi :[10.48550/arXiv.2302.07640](https://doi.org/10.48550/arXiv.2302.07640)
5. **G. Bonafos**, J-M. Freyermuth, P. Pudlo, S. Tronçon, A. Rey. 2023. Topological data analysis of human vowels : Persistent homologies across representation spaces. doi :[10.48550/arXiv.2310.06508](https://doi.org/10.48550/arXiv.2310.06508)

Participation aux conférences et écoles d'été au cours de la période de thèse :

1. Octobre 2023. International Bio Acoustics Congress 2023, Sapporo, Japon. Poster.
2. Juillet 2023. 54^{ème} journées de Statistique de la SFdS, Bruxelles, Belgique. Présentation.
3. Avril 2023. Statlearn Conference, Montpellier, France. Poster.

Résumé

Aux alentours de son premier anniversaire, l'enfant humain prononce son premier mot. Cette première production n'est pourtant pas le début de son apprentissage de la langue. Celui-ci commence dès sa naissance. En effet, tout au long de sa première année, l'enfant développe des capacités motrices lui permettant de produire une gamme de vocalisations de plus en plus large, en les calibrant au langage qui l'entoure. Les moyens d'enregistrements et de stockage récents permettent de construire de nouvelles bases de données de vocalisations produites tout au long de l'année. Nous construisons dans cette thèse une telle base et présentons trois contributions pour aider à étudier la question des vocalisations infantiles pré-langagières.

Nous proposons d'abord une méthodologie pour détecter et classifier automatiquement les vocalisations dans les enregistrements audios massifs. Elle permet l'apprentissage d'un réseau de neurones à partir d'un peu plus d'une heure de données étiquetées, qui fait ensuite le travail d'extraction de vocalisations d'enregistrements naturels massifs. Elle a été appliquée sur deux ensembles d'enregistrements, prouvant son adaptabilité : les enregistrements de bébé récoltés pour ce travail de thèse ainsi que des enregistrements d'un mois d'un enclos de singe, permettant de produire deux nouveaux ensembles de données, un de vocalisation de bébé et un de vocalisation de singe. Nous avons rendu ce dernier librement accessible, tout comme le code permettant de reproduire la méthodologie. Cette contribution a donné lieu à un papier, actuellement soumis et accessible en *preprint*, qui a été présenté sous forme de poster à la conférence *Statlearn 2023* à Montpellier, France, et à l'*IBAC* de 2023 à Sapporo, Japon.

Nous emmenons ensuite des preuves empiriques de l'intérêt d'incorporer une information topologique dans la représentation d'un signal vocal humain pour une tâche de classification. Nous quantifions la plus-value d'une approche topologiquement augmentée et les différences selon l'objet représentant une vocalisation identique. On montre que l'information topologique est complémentaire à une information fréquentielle et que les homologies persistantes calculées sur chaque objet sont complémentaires entre elles. Pour répondre à cette question, nous avons construit une nouvelle base d'enregistrements de 11 200 voyelles, que nous avons rendu librement accessible. Nous avons comparé les résultats sur trois tâches de classification selon que la représentation du signal est topologiquement augmentée ou non, ainsi que la meilleure façon de vectoriser l'information contenue dans un diagramme de persistance. Ce travail a été présenté oralement à la conférence de la *SFds* de 2023 à Bruxelles, Belgique.

Enfin, nous avons classifié de manière non-supervisée, par une modélisation bayésienne non-paramétrique, les vocalisations produites par un enfant durant sa pre-

mière année de vie, à partir d'une représentation topologiquement augmentée du signal. On découvre huit classes de vocalisations, dont la proportion de production varie selon le développement, et avec des caractéristiques fréquentielles différentes.

Mots clés : classification, apprentissage statistique, développement du langage, statistique bayésienne, analyse topologique des données

Abstract

Around her first birthday, the human child utters her first word. This first utterance is not, however, the beginning of language learning. This begins at birth. Throughout the first year of life, children develop motor skills that enable them to produce an increasingly wide range of vocalizations, calibrated to the surrounding language. Recent recording and storage systems have made it possible to build new databases of vocalizations produced throughout the year. In this thesis, we build such a database and present three contributions to help study the question of pre-language infant vocalizations.

First, we propose a methodology for automatically detecting and classifying vocalizations in massive audio recordings. It enables a neural network to be trained from just over an hour's worth of labeled data, which then does the job of extracting vocalizations from massive natural recordings. It has been applied to two sets of recordings, proving its adaptability: the baby recordings collected for this thesis and one month's recordings from a monkey enclosure, producing two new data sets, one of baby vocalizations and one of monkey vocalizations. We made it freely accessible, as is the code used to reproduce the methodology. This contribution has resulted in a paper, currently submitted and available as a preprint, which was presented as a poster at the *Statlearn 2023* conference in Montpellier, France, and at the *IBAC 2023* in Sapporo, Japan.

We then provide empirical evidence of the value of incorporating topological information into the representation of a human speech signal for a classification task. We quantify the added value of a topologically augmented approach and the differences depending on the object representing an identical vocalization. We show that topological information is complementary to frequency information, and that the persistent homology computed on each object is complementary to each other. To answer this question, we built a new, freely accessible database of 11,200 vowel recordings. We compared the results on three classification tasks, depending on whether the signal representation is topologically augmented or not, as well as the best way to vectorize the information contained in a persistence diagram. This work was presented orally at the *SFds 2023* conference in Brussels, Belgium.

Finally, we performed clustering, using non-parametric Bayesian modeling, of the vocalizations produced by a child during its first year of life, based on a topologically augmented representation of the signal. Eight classes of vocalizations were discovered, with different proportions of production depending on development, and with different frequency characteristics.

Keywords: clustering, machine learning, language development, Bayesian statistics,

Topological Data Analysis

Remerciements

Je tiens tout d'abord à remercier l'ensemble des membres de mon jury, Florence Levrero et Bertrand Michel, mes rapporteurs, pour le temps qu'ils ont accepté de consacrer à la relecture de cette thèse, ainsi que Marianne Clausel et Vincent Vandewalle, d'avoir accepté de faire partie de mon jury. C'est un honneur de tous les retrouver autour de ce travail.

Je remercie mes directeurs de thèse, Pierre Pudlo, Jean-Marc Freyermuth, Samuel Tronçon et Arnaud Rey. Merci pour leurs conseils, leur apprentissage, leur encadrement, leur bienveillance. Merci d'avoir imaginé ce beau projet de recherche. Merci de m'avoir appris à mener ma barque.

Je remercie l'ensemble des personnes qui ont une implication, de près ou de loin, avec ce travail, et qui ont tous permis à cette thèse d'être écrite aujourd'hui. À tous les membres des laboratoires, administratif, ingénieur ou chercheur, qui ont pu me donner des conseils et répondre à mes questions. Au service de pédiatrie et de médecine néonatale de l'hôpital de Saint-Joseph, notamment aux docteurs Edwin Quarello et Jean-Michel Bartoli, pour m'avoir permis de leur présenter ce projet et d'entrer en contact avec des familles. Je remercie également l'ensemble des membres de ces services, pour avoir participé à faire circuler l'information.

Je remercie bien entendu les premiers acteurs et actrices de cette thèse, les enfants que j'ai eus la chance d'enregistrer et leurs parents, qui ont eu la gentillesse de participer à cette étude. C'est un vrai travail qu'ils ont fourni, et je leur en suis infiniment reconnaissant.

Merci à toutes les personnes qui font de Saint-Charles un lieu si sympathique, à tous les travailleurs du site, au personnel du CROUS avec qui il est toujours agréable de discuter, à Marie-Hélène grâce à qui nos bureaux restent propres.

Je remercie tous les doctorants et post-doctorants de la Fed3C, pour m'avoir accueilli, et pour tous les bons moments que nous avons passés.

Merci Rebecca, pour toi et pour tout.

Enfin, merci à toute ma famille, et à mes parents, grâce à qui j'en suis là aujourd'hui.

Table des matières

Affidavit	2
Liste de publications et participation aux conférences	3
Résumé	4
Abstract	6
Remerciements	8
Table des matières	9
Table des figures	13
Liste des tableaux	16
1. Introduction	17
1.1. Le problème : les productions vocales du bébé	18
1.1.1. De la naissance au premier mot	18
1.1.2. Un détail plus fin, au-delà du babillage	20
1.1.3. Intérêt des enregistrements massifs	22
1.2. Réponse au premier problème : la détection des vocalisations	23
1.2.1. Revue de littérature	24
1.2.2. Retour sur l'apprentissage statistique	25
1.2.3. Points à résoudre pour son utilisation dans notre cas	26
1.3. Réponse au deuxième problème : la représentation du signal	27
1.3.1. La physique du son	27
1.3.2. Hypothèse géométrique de la variété	28
1.3.3. Lien avec la TDA	29
1.3.4. Revue de littérature	30
1.4. Réponse au troisième problème : la classification des vocalisations	31
1.4.1. Présentation du paradigme bayésien	31
1.4.2. Le non-paramétrique	33
1.4.3. Revue de littérature	34
2. Réseau de neurones	39
2.1. Remise en contexte et résumé de l'étude	39
2.2. Construction d'un réseau de neurones	41
2.2.1. Architecture d'un réseau de neurones	41

2.2.2.	Les fonctions d'activation	43
2.2.3.	Apprentissage de représentation	45
2.2.4.	La fonction objectif	50
2.3.	Apprentissage	50
2.3.1.	La descente de gradient	51
2.3.2.	La rétro-propagation	53
2.3.3.	L'algorithme de descente de gradient	54
2.4.	La régularisation, le sur-apprentissage	56
2.4.1.	Data augmentation	56
2.4.2.	Batch normalization	57
2.4.3.	Dropout	57
2.4.4.	Contrainte sur la norme des paramètres	58
2.4.5.	Early-stopping	58
2.5.	Apprentissage des hyper-paramètres	59
2.5.1.	Le choix des hyper-paramètres	59
2.5.2.	L'optimisation bayésienne	60
3.	Detection and classification of vocal productions in large scale audio recordings	63
3.1.	Introduction	64
3.2.	Methodology	66
3.2.1.	Data	66
3.2.2.	Network architecture	67
3.2.3.	Fit on the data	69
3.2.4.	Vocalization delineation and classification	71
3.3.	Experimental Validation	71
3.3.1.	From audio recordings to data banks for our method	72
3.3.2.	Performance of our deep learning architecture	73
3.3.3.	New large-scale databases of vocalizations	74
3.4.	Conclusion and Discussion	75
3.5.	Supplementary materials	77
4.	Analyse Topologique du Signal	82
4.1.	Remise en contexte et résumé de l'étude	82
4.2.	La forme des données	84
4.2.1.	Complexe simplicial géométrique	84
4.2.2.	Complexe simplicial abstrait	85
4.2.3.	Espace sous-jacent	86
4.3.	Notions d'équivalence	86
4.3.1.	Homéomorphisme	86
4.3.2.	Équivalence homotopique	87
4.3.3.	Théorème du Nerf	88
4.4.	Homologie	90
4.4.1.	Chaînes, cycles et bords	90

4.4.2. Groupe d'homologie	92
4.5. Persistence	93
4.5.1. Topologie et échelle	93
4.5.2. Filtration	94
4.5.3. Homologies persistantes	95
4.6. Diagramme de persistance	97
4.6.1. Définition d'un diagramme de persistance	97
4.6.2. Stabilité d'un diagramme de persistance	98
4.6.3. Consistance d'un diagramme de persistance	99
5. Topological data analysis of human vowels : Persistent homologies across representation spaces	102
5.1. Introduction	104
5.2. The problem	104
5.2.1. Motivations for TDA	104
5.2.2. Representation space	105
5.2.3. Same signal, different representations, different topological characteristics	107
5.3. Topological Data Analysis : An overview	108
5.3.1. Persistent homology	109
5.3.2. Exploitation of the information from the diagram space	110
5.4. Experiments	111
5.4.1. Presentation of the data	111
5.4.2. Comparison on three supervised classification tasks	112
5.4.3. Computation of the homological information	112
5.4.4. Extraction of the information from the persistence diagram and comparison of the persistent variables	112
5.4.5. Step-wise selection of the variables	116
5.5. Results	117
5.5.1. Supervised task	117
5.5.2. Best topological variables	121
5.5.3. Unsupervised analysis, learning the manifold	121
5.6. Discussion	123
5.6.1. On the improvements on the prediction of labels	123
5.6.2. Different objects, different topologies	123
5.6.3. On the more present persistent variables	124
5.7. Conclusion	125
6. Modélisation bayésienne	127
6.1. Remise en contexte et résumé de l'étude	127
6.2. Définition du processus de Dirichlet	128
6.2.1. Définition formelle	128
6.2.2. Stick-breaking	130
6.2.3. Processus du Restaurant Chinois	131

6.3. Modèle de mélange avec processus de Dirichlet	134
6.3.1. Le modèle de mélange classique	135
6.3.2. Cas non-paramétrique	136
6.4. Estimation du modèle	138
6.4.1. Les propositions algorithmiques	138
6.4.2. Retour sur le modèle	140
6.4.3. Détermination de la partition	142
7. Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development	145
7.1. Introduction	146
7.2. Data	147
7.3. Modelling	148
7.3.1. Topologically augmented signal representation	148
7.3.2. Nonparametric Bayesian modelling for clustering	152
7.3.3. Depth of the clusters for representative vocalizations	154
7.4. Data analysis	154
7.4.1. Partition	154
7.4.2. Comparison of clusters	156
7.5. Conclusion and discussion	160
7.6. Supplementary materials	161
Conclusion	162
Bibliographie	167
ANNEXES	200
A. Documents à destination des parents	200

Table des figures

1.1. Développement de la perception du langage et du développement de la production vocale du bébé sur la première année de vie. Image tirée de Patricia K. KUHL 2004.	19
2.1. Schéma de l'architecture d'un réseau de neurones. Une couche d'entrée prend les données. Une succession de couches cachées applique des transformations non-linéaires aux données, les projetant dans un espace latent. À partir de cette représentation, la couche finale produit \hat{y} . La profondeur, la largeur, les connections entre les couches déterminent l'architecture du réseau. Ici, l'architecture est composée de deux modules à partir de la représentation apprise, permettant de produire deux sorties, la probabilité que l'entrée, de dimension 16000, contienne une vocalisation, et la probabilité de la classe de la vocalisation.	42
2.2. L'architecture du réseau de notre première contribution (BONAFOS, PUDLO, FREYERMUTH et al. 2023). Nous transférons les couches avant du modèle de YamNet, qui calcule pour une fenêtre d'une seconde d'audio son spectrogramme de 96×64 , suivi d'une concaténation de couches de convolution <i>depth-wise</i> (en bleu) et <i>step-wise</i> (en violet), opérations de convolution introduites par HOWARD, M. ZHU, B. CHEN et al. 2017. Elles se terminent par une couche de <i>pooling</i> pour créer une représentation de dimension 1024. A partir de cette représentation, nous créons une double architecture : un module pour détecter s'il y a une vocalisation dans l'image ou non, un module pour classifier la vocalisation. Les couches avant sont totalement gelées pendant l'apprentissage, tandis que nous apprenons l'arrière de l'architecture sur les données. Une fois entraîné, le modèle peut traiter des données audio non étiquetées pour extraire des segments de vocalisation et prédire leur classe simultanément.	49
S3.1. Confusion matrices for the <i>baboon</i> data.	77
S3.2. Confusion matrices for the <i>human baby</i> data.	78
4.1. Complexe de chaînes d'un complexe, consistant en une séquence linéaire de groupes de chaînes, de cycles et de bords, connectés par des homomorphismes, l'opérateur de bord. Les groupes des p -bords sont des sous-groupes des groupes des p -cycles, qui sont eux-mêmes des sous-groupes des p -chaînes. Schéma tiré de H. EDELSBRUNNER et HARER 2009.	92

5.1.	Different representation of the same signal. 5.1a is the initial wave of the sound; 5.1b) is the surface of its spectrogram; 5.1c is the zeros of its spectrogram in the time-frequency plane; 5.1d is its Taken's embedding.	106
5.2.	Three persistent diagrams computed for the three different representation of the same signal of Figure 5.1. 5.2a is the persistence diagram of the spectrogram's surface using sublevel sets; 5.2b is the persistence diagram of the zeros of the spectrogram using an Alpha complex; 5.2c is the persistence diagram of the Taken's embedding using an Alpha complex. The points in the diagram represent the computed persistent homologies. The different colors represent the different dimensions : black for $p = 0$, red for $p = 1$ and blue for $p = 2$.	108
5.3.	Projection of the records on the 2d space learned with UMAP, depending on the input of the algorithm : MFCC, topological variables from the spectrogram's surface, the spectrogram's zeros or the Taken's embedding.	122
6.1.	Illustration de la construction <i>stick-breaking</i> de l'ensemble infini de poids $\pi \sim GEM(\alpha)$ correspondant à la mesure $G \sim DP(\alpha, G_0)$. À gauche, un bâton de longueur un est cassé à un point aléatoire β_1 tiré dans $Beta(1, \alpha)$. La longueur de la partie cassée produit π_1 . On casse récursivement le bout de bâton restant pour produire π_2, π_3, \dots À droite, les $K = 20$ premiers poids générés par quatre constructions <i>stick-breaking</i> aléatoires, deux pour $\alpha = 1$, deux pour $\alpha = 5$. Image tirée de E. B. (B. SUDDERTH 2006.	132
6.2.	Illustration du processus du restaurant chinois. Les diamants représentent les clients, associés à l'observation correspondante, les cercles sombres. Les grands cercles représentent les tables, qui sont donc les classes, ainsi que les paramètres θ associés. Ces derniers ne font pas partie du processus du restaurant chinois en tant que tel mais plutôt de modèle du mélange. Image tirée de GERSHMAN et D. M. BLEI 2012.	133
6.3.	Représentation graphique du modèle utilisé pour la classification non-supervisée des voyelles dans le chapitre 7. Modèle de mélange gaussien avec un processus de Dirichlet comme loi a priori sur la mesure aléatoire. La mesure de base est une loi normale Wishart-Inverse. La loi normale prend elle-même deux paramètres, un paramètre de moyenne et un paramètre de variance, sur lesquels on pose une distribution a priori, respectivement une loi normale et une loi Gamma. La loi de Wishart-Inverse prend, elle aussi, deux paramètres. On pose une loi a priori sur la matrice d'échelle, une loi de Wishart. Reste en hyper-paramètres tous les paramètres des hyper-priors, le degré de liberté de la loi de Wishart-Inverse et le paramètre de concentration α du processus de Dirichlet.	140

7.1. Representation of the signal, by the surface of its spectrogram or by its Taken's embeddings. On each of these representations, we compute the persistent homology that we resume in the persistence diagram. We plot here the spectrogram surface and the Taken's embeddings for the deepest vocalization of each of the cluster we found ($N = 8$) with our Dirichlet process mixtures model.	148
7.1. Representation of the signal, by the surface of its spectrogram or by its Taken's embeddings. On each of these representations, we compute the persistent homology that we resume in the persistence diagram. We plot here the spectrogram surface and the Taken's embeddings for the deepest vocalization of each of the cluster we found ($N = 8$) with our Dirichlet process mixtures model (continuation).	149
7.1. Representation of the signal, by the surface of its spectrogram or by its Taken's embeddings. On each of these representations, we compute the persistent homology that we resume in the persistence diagram. We plot here the spectrogram surface and the Taken's embeddings for the deepest vocalization of each of the cluster we found ($N = 8$) with our Dirichlet process mixtures model (continuation).	150
7.2. Manifold of the vocalizations of the baby produced during one year, learned on the topologically augmented representation of the vocalizations, on which we project the clustering. We highlight the deepest vocalization of each cluster. Each point is a vocalization, each color a cluster. The diamonds correspond to the deepest vocalization of each class.	155
7.3. Proportion of monthly production of vocalization per cluster. Parents did not record during three months, yet the gap.	156
7.4. Spectrograms of the deepest vocalization of each cluster.	158
7.6. Persistence diagrams of the Taken's embeddings of the deepest vocalization of each cluster. The black, red and blue points correspond respectively to homological features of dimension 0, 1 and 2.	159
7.5. Persistence diagrams of the surface of the spectrogram of the deepest vocalization of each cluster. The black, red and blue points correspond respectively to homological features of dimension 0, 1 and 2.	159
S7.1. Initial manifold of the vocalizations of the baby produced during one year, on which we project the clustering of the Dirichlet process mixtures model, before to remove the "garbage cluster". This class is separated from the other classes, composed by baby vocalizations.	161

Liste des tableaux

3.1. Network Architecture	68
3.2. Hyper-parameters of the model	70
3.3. Performance of the deep learning in the baboon and human studies . .	74
S3.1.Total and per partition distribution of the baboon labeled data set. . .	78
S3.1.Total and per partition distribution of the baboon labeled data set. . .	79
S3.2.Seconds of vocalization for baboons, for each class, over the month. . .	79
S3.3.Total and per partition distribution of the human baby labeled data set	80
S3.3.Total and per partition distribution of the human baby labeled data set	81
S3.4.Seconds of vocalization for human babies, for each class, over the year	81
5.1. Comparison of Out Of Bag error (OOB) for different signal representa- tions. Bold numbers emphasize the best signal representation for each task	118
5.2. Most frequently kept persistent variables in each best model. The va- riables present in more of 50% of the best models are emphasized in bold	120
7.1. Number of vocalizations per month in the massive audio recordings, as well as the mean and the standard deviation of the duration of the vocalizations produced per month.	148
7.2. Number of vocalizations per cluster, as well as the mean and the stan- dard deviation of the duration of the vocalizations of the cluster.	154
7.3. Proportion of production for each cluster during the year.	157

1. Introduction

Au cours de sa première année de vie, l'enfant apprend à maîtriser son appareil vocal. D'importants changements anatomiques ont lieu et l'enfant développe des capacités motrices qui lui permettront, à l'issue des douze premiers mois, de produire son premier mot.

On trouve dans la littérature une description des étapes du chemin que suit l'enfant, ainsi que des preuves de l'importance des productions vocales durant cette première année pour les capacités langagières futures. La possibilité de prédire de possibles troubles ou retards développementaux à partir des productions vocales de la première année de vie motive une étude approfondie de celles-ci. Les nouveaux moyens et capacités d'enregistrement et de stockage ouvrent des opportunités pour décrire plus finement les productions vocales des enfants entre 0 et 12 mois.

Ce travail de thèse s'attache à proposer une description quantitative des premières phases de l'émergence du langage chez l'humain, ses variations inter-individuelles, ses liens avec des phases postérieures. Plus particulièrement, l'objectif est d'affiner la description des classes de vocalisations produites au cours de la première année de vie.

Pour cela, nous avons enregistré pendant un an, de leur naissance à leur premier anniversaire, 16 enfants. Nous avons donné à chaque famille un micro mobile en leur demandant d'enregistrer l'enfant le plus possible, dans toutes les conditions possibles et en posant le micro au plus proche de l'enfant, au minimum trois jours par mois. Nous avons ainsi constitué une base de données massives, constituée de centaines d'heures d'enregistrement par enfant, réalisés tout au long d'une année. Nous avons également récolté des informations sur le contexte socio-économique de la famille et sur les conditions d'accouchement. C'est une base de données riche, massive, regroupant des enregistrements faits en condition naturelle.

L'analyse de cette base de données, permettant *in fine* de proposer une classification non-supervisée des vocalisations d'un enfant entre sa naissance et son premier anniversaire, soulève plusieurs questions et défis. Nous les résumons en trois principaux problèmes, que cette thèse propose de résoudre : l'extraction des vocalisations dans les enregistrements continus, la représentation du signal étant donné notre problème et la classification de celui-ci. Après avoir présenté la question de la production vocale des enfants entre 0 et 12 mois, nous présentons nos choix pour y répondre.

Organisation générale de la thèse

La thèse s'organise autour de chacun de ces obstacles et de nos propositions pour les résoudre, qui sont les trois contributions de cette thèse. Pour chacune de ces contributions, le chapitre précédent l'introduit et présente les outils théoriques sur

lesquels elle se base. Le Chapitre 3 est notre contribution afin d’extraire les vocalisations des enregistrements continus. Elle se base sur un réseau de neurones, que l’on introduit dans le Chapitre 2. Le Chapitre 5 est notre contribution sur la représentation d’un signal sonore. Elle se base sur l’Analyse Topologique des Données, que l’on introduit dans le Chapitre 4. Enfin, le Chapitre 7 est notre contribution pour la classification non-supervisée des productions vocales. Elle se base sur un modèle bayésien non-paramétrique, que l’on introduit dans le Chapitre 6.

1.1. Le problème : les productions vocales du bébé

1.1.1. De la naissance au premier mot

Au cours de sa première année de vie, l’enfant développe des capacités décisives pour la maîtrise du langage. Il apprend à discriminer des sons, les phones, et des catégories, les phonèmes. Les phones sont les réalisations concrètes, les sons du langage que l’enfant entend, alors que les phonèmes sont des représentations abstraites. Elles forment les éléments constitutifs du langage et vont aider l’enfant à segmenter du flux de parole. Un même phonème pourra être prononcé de différentes manières, par différents phones, mais il aura à chaque fois le même sens. Les enfants apprennent, au cours de leurs premiers mois, à distinguer les phonèmes de leur langue maternelle à partir des phones qu’ils entendent.

Les mécanismes neuronaux qui encodent le son dans le cortex auditif, notamment la parole, évoluent énormément au cours des premiers mois de vie (WILD, LINKE, ZUBIAURRE-ELORZA et al. 2017; ZUBIAURRE-ELORZA, LINKE, HERZMANN et al. 2018; GORINA-CARETA, RIBAS-PRATS, ARENILLAS-ALCÓN et al. 2022). In-utero, les enfants perçoivent les sons, notamment leur fréquence fondamentale F_0 , tant et si bien, qu’à la naissance, les nouveaux-nés sont capables de suivre la fréquence fondamentale aussi bien que les adultes. En revanche, le ventre de la mère agissant comme un filtre passe bas, il leur manque l’information relative aux autres harmoniques F_1, F_2 . Après l’accouchement, l’environnement acoustique est bien plus riche et des fréquences autrefois imperceptibles dans le ventre de la mère deviennent identifiables pour l’enfant. S’ils sont capables d’encoder l’information relative à la fréquence fondamentale des sons qu’ils entendent grâce à leur expérience in-utero, les enfants doivent apprendre à encoder l’information relative aux formants (RIBAS-PRATS, ALMEIDA, COSTA-FAIDELLA et al. 2019; ARENILLAS-ALCÓN, COSTA-FAIDELLA, RIBAS-PRATS et al. 2021; ARENILLAS-ALCÓN, RIBAS-PRATS, ESCERA et al. 2021).

Cet apprentissage est très rapide et les enfants parviennent rapidement à faire des distinctions fines dès les premiers mois (MEHLER, P. JUSCZYK, LAMBERTZ et al. 1988; CHEOUR-LUHTANEN, ALHO, KUJALA et al. 1995). Les enfants sont capables d’apprentissage statistique, *i.e.*, d’apprendre les probabilités de cooccurrences entre événements, et peuvent alors apprendre les probabilités transitionnelles entre séquences tonales (J. R. SAFFRAN, R. N. ASLIN et E. L. NEWPORT 1996; Richard N. ASLIN, Jenny R. SAFFRAN et Elissa L. NEWPORT 1998; Jenny R SAFFRAN, E. K. JOHNSON, Richard N ASLIN et al.

1. Introduction – 1.1. Le problème : les productions vocales du bébé

1999; Jenny R. SAFFRAN 2003). Ils apprennent le langage en découvrant les caractéristiques de la parole (consonne, voyelle, combinaison des deux) et commencent à comprendre assez rapidement, dès 6 mois, certains mots (E. BERGELSON et SWINGLEY 2012). Ils sont capables à cet âge de lier des mots non seulement à des objets, mais aussi à des catégories des objets, *e.g.*, ils associent le mot "maman" à leur mère, mais également le mot "main" à une main (TINCOFF et P. W. JUSCZYK 2012).

Les enfants sont donc déjà capables à la naissance de percevoir et distinguer des différences dans les sons auxquels ils sont confrontés, et leur perception et capacité de discrimination s'affinent très rapidement. Si on s'attend à un décalage temporel entre la production et la perception, cette première évolue aussi rapidement.

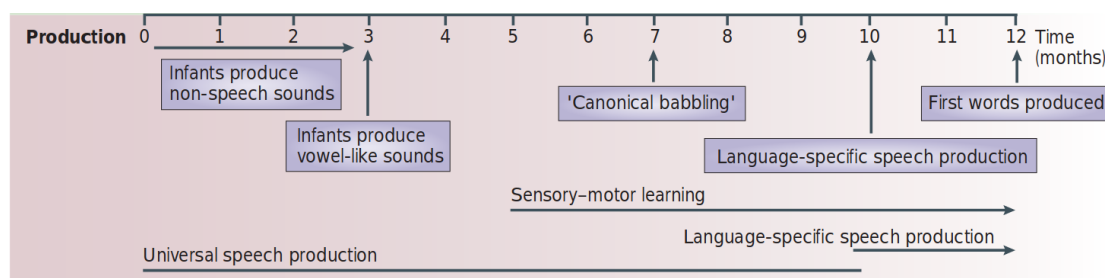


FIGURE 1.1. – Développement de la perception du langage et du développement de la production vocale du bébé sur la première année de vie. Image tirée de Patricia K. KUHL 2004.

La Figure 1.1 est tirée de Patricia K. KUHL 2004 et reprend le déroulement temporel classique de la production vocale des enfants au cours de leur première année de vie. Dès 2 mois, les enfants commencent à gazouiller, ce qui implique des articulations de la langue à l'arrière de la bouche. Suivent des quasi-voyelles, qui sont les premières vocalisations volontaires (D. Kimbrough OLLER 2000). Ces dernières émergent du gazouillage, qui est graduellement augmenté de voyelles de plus en plus claires, et de sons ressemblant de plus en plus à des syllabes, avec des combinaisons voyelle-consonne. Ce chemin emmène au babillage au sixième mois, que l'on distingue en deux types. Il est d'abord marginal, une combinaison unique de consonne et de voyelle, *e.g.*, "ba". Puis cette combinaison est répétée dans une vocalisation, c'est le babillage canonique, ou babillage dupliqué, *e.g.*, "baba". Dans une seconde étape, le babillage est non dupliqué. Différentes combinaisons de consonnes et voyelles se retrouvent dans une même vocalisation, avec une plus grande variété pour combiner différentes syllabes, *e.g.*, "bamapa". Cette deuxième étape emmènera *in fine* à la production du premier mot au premier anniversaire.

On ajoute en général à cette description les autres productions vocales, comme les pleurs présents dès la naissance, que l'on considère comme non volontaires. Ces vocalisations sont des indicateurs d'états affectifs. On fait une distinction entre vocalisations verbales et les vocalisations non verbales. Dans cette perspective, le babillage est considéré comme un tournant. Les productions vocales des enfants deviennent plus précises, de meilleures approximations des sons ambiants. Une information

1. Introduction – 1.1. Le problème : les productions vocales du bébé

spécifique aux phonèmes de la langue maternelle est graduellement consolidée. Le volume et la complexité de la production augmente, avec de grosse variation individuelle (MORGAN et WREN 2018). La proportion de vocalisation canonique dépasse 0.15 au-delà de 7 mois (Margaret CYCHOSZ, CRISTIA, Erika BERGELSON et al. 2021) : les enfants développent les outils nécessaires pour faire la transition de la production pré-langagière à la première production langagière. Un babillage tardif est alors même prédictif de troubles développementaux. Les enfants montrant un retard dans le babillage canonique ont un vocabulaire moins important à 18, 24 et 30 mois (D. K. OLLER, R. E. EILERS, A. R. NEAL et al. 1998; D. Kimbrough OLLER, Rebecca E EILERS, A. Rebecca NEAL et al. 1999). Le seuil de 0.15 du ratio de babillage canonique permet de discriminer entre les groupes d'enfants sans troubles et ceux atteints d'un syndrome de Rett (BARTL-POKORNY, POKORNY, GARRIDO et al. 2022).

Le babillage est un apprentissage des productions vocales, une exploration de l'espace acoustique et un entraînement pour reproduire des sons cibles entendus dans l'environnement. On peut ainsi voir le babillage comme une "calibration", que l'on retrouve par ailleurs chez d'autres espèces que l'humain (TER HAAR, FERNANDEZ, GRATIER et al. 2021).

1.1.2. Un détail plus fin, au-delà du babillage

Le babillage se spécialise selon les interactions de l'enfant (BOYSSON-BARDIES, SAGART et C. DURAND 1984; S. A. S. LEE, B. DAVIS et MACNEILAGE 2010). L'émergence des syllabes canoniques avant 9 mois est commun à tous les langages et se retrouve dans différents contextes culturels, mais celles-ci approximent les phonèmes de la langue ambiante au fil du temps. Dès 10 mois, les enfants français produisent plus de vocalisations nasales que les enfants anglais (Margaret CYCHOSZ, CRISTIA, Erika BERGELSON et al. 2021). Néanmoins, d'autres moments clés se jouent avant. La flexibilité fonctionnelle, condition *sine qua non* du langage humain, apparaît avant le babillage. D. Kimbrough OLLER, BUDER, RAMSDELL et al. 2013 distinguent trois types de vocalisations infantiles à 3-4 mois qui montrent une flexibilité fonctionnelle : les vocalisations de type voyelle, les couinements et les grognements. La flexibilité fonctionnelle est la capacité d'attribuer des valences affectives différentes à des mots ou phrases selon le contexte. Alors que des vocalisations comme les pleurs ou les rires, vocalisations partagées avec d'autres primates, ont une stabilité fonctionnelle (les pleurs sont négatifs, les rires positifs), d'autres vocalisations que le babillage ont une flexibilité fonctionnelle. Cette condition très importante du langage humain apparaît très tôt dans le pré-langage humain, dès le troisième mois (JHANG 2017).

De plus, la spécialisation selon la langue maternelle se retrouve avant le babillage, pour des vocalisations non verbales. Les pleurs des enfants s'adaptent à la prosodie de la langue maternelle (MAMPE, FRIEDERICI, CHRISTOPHE et al. 2009). Ces vocalisations pourraient ainsi avoir aussi un rôle important à jouer dans le processus d'apprentissage qui emmènera à la production du premier mot. Tout comme les quasi-voyelles pavent le chemin pour arriver au babillage, qui débouche *in fine* sur le premier mot, les autres vocalisations de l'enfant permettent à celui-ci de développer et d'apprendre

1. Introduction – 1.1. Le problème : les productions vocales du bébé

à maîtriser son appareil buco-phonatoire, et de mieux calibrer ses vocalisations étant donné la langue ambiante qu'il entend. Certes, l'apparition de certaines vocalisations sont des moments clés dans le processus d'apprentissage de la langue, et leur absence ou retard permet même de détecter de possibles troubles développementaux, mais les vocalisations antérieures, que l'on ne considère pas comme étant directement liées au langage, aident à l'apprentissage du contrôle moteur et donc à la capacité de produire des sons adaptés à la langue cible.

L'enfant doit apprendre à coordonner plus 70 muscles et 10 parties corporelles pour articuler et produire les sons désirés, et donc développer une capacité motrice pour maîtriser son appareil vocal (THELEN, ULRICH et WOLFF 1991). En même temps, l'appareil vocal connaît un développement anatomique, qui impacte la façon dont les enfants produisent les gestes articulatoires. La croissance du conduit vocal conditionne l'espace acoustique de la production vocale (BARBIER, BOË, CAPTIER et al. 2015). Dès quatre mois, les enfants établissent une correspondance entre des voyelles qu'ils entendent et leur articulation, et cherchent à imiter les voyelles en question (P. K. KUHL et MELTZOFF 1982; P. K. KUHL et MELTZOFF 1996). Dès la première semaine même, les nouveaux nés imitent des sons en ouvrant ou fermant leur bouche, même les yeux fermés (X. CHEN, STRIANO et RAKOCZY 2004). Ils font une correspondance entre l'auditif et l'oral, et l'imitation est une capacité clé.

Tout comme le babillage permet à l'enfant d'explorer son espace acoustique, d'autres vocalisations de l'enfant ont ce rôle antérieurement. La complexité mélodique des pleurs, que l'on regarde par la variation de la fréquence fondamentale, augmente au cours des six premiers mois. De plus, les pleurs portent une signature individuelle, qui évolue elle aussi au cours du développement (LOCKHART-BOURON, ANIKIN, PISANSKI et al. 2023). C'est également le cas d'autres vocalisations qui ne sont pas des pleurs. Cette complexification des productions vocales est une pierre angulaire du développement du langage et la reconnaître comme une des composantes importantes est utile non seulement pour mieux comprendre l'apprentissage du langage, mais aussi pour avoir un marqueur de troubles développementaux (WERMKE, ROBB et SCHLUTER 2021). L'analyse des pleurs des nouveaux nés prématurés est en effet utile pour évaluer leur développement (CABON, MET-MONTOT, POREE et al. 2021). Leurs pleurs n'ont pas les mêmes caractéristiques mélodiques que les enfants nés à terme et on note une relation entre la variation de la fréquence fondamentale des pleurs et de meilleurs résultats pour des tests cognitifs et langagiers à 18 mois. Les expériences vocales postnatales ont une importance (SHINYA, KAWAI, NIWA et al. 2017).

Enfin, si un retard du babillage canonique est prédictif d'un retard de la parole et d'un vocabulaire moins important (D. K. OLLER, R. E. EILERS, A. R. NEAL et al. 1998; D. Kimbrough OLLER, Rebecca E EILERS, A. Rebecca NEAL et al. 1999), et que la proportion de babillage canonique est un bon marqueur pour discriminer entre des enfants atteints d'un trouble développemental (syndrome de l'X fragile, de Rett ou autisme) ou ceux sans troubles (LANG, BARTL-POKORNY, POKORNY et al. 2019; BARTL-POKORNY, POKORNY, GARRIDO et al. 2022), il convient de s'intéresser plus en détail aux propriétés de celui-ci, ainsi qu'à l'influence des autres vocalisations et aux liens qui existent ou pas entre les différentes classes de vocalisations. PAUL, FUERST,

1. Introduction – 1.1. Le problème : les productions vocales du bébé

RAMSAY et al. 2011 notent l'intérêt de regarder d'autres vocalisations précoces de l'enfant en plus des vocalisations canoniques quand on suspecte des troubles du spectre autistique. On note des retards dans l'acquisition de différentes vocalisations (consonnes particulières, formes des syllabes, contours prosodiques). Ce niveau de détail est utile pour voir comment le développement du langage se fait chez les enfants à risque. Les babillages canoniques comptent, sont importants, mais d'autres éléments plus fins sont à analyser aussi.

1.1.3. Intérêt des enregistrements massifs

Afin de pouvoir étudier plus en détail les vocalisations produites par l'enfant au cours de sa première année de vie, analyser leurs évolutions et le décours développemental, nous avons commencé à enregistrer massivement, en continu, des enfants de la naissance à leur premier anniversaire. Des enregistrements continus, dans leur environnement naturel, fournissent une base de données de production plus variée, permettant donc un niveau d'analyse plus fin. De plus, le caractère longitudinal permet de suivre le développement des productions vocales et voir en quoi certaines catégories conditionnent l'apparition d'autres. Le fait que les enregistrements soient continus et massifs, couplé au fait qu'ils soient suivis sur un an, permet d'avoir une idée plus précise de l'exploration par l'enfant de son espace acoustique et la trajectoire développementale vers le premier mot.

En effet, suivre les productions vocales d'un enfant permet d'apprendre beaucoup mieux comment le développement se passe et de réévaluer l'importance de certains facteurs, par exemple autre que la parole (B. C. ROY, FRANK, DECAMP et al. 2015). Comme on le retrouve pour l'étude des productions vocales de différentes espèces animales et pour la question de l'évolution du langage (DE BOER 2019), des enregistrements plus écologiques apportent une information possiblement absente des données actuellement à disposition. GILKERSON, RICHARDS, WARREN et al. 2017 notent l'importance de l'étude des productions vocales de l'enfant humain dans son environnement naturel avec ses parents. En quantifiant ses vocalisations au fil du temps, on peut estimer la relation entre certaines covariables et le développement de l'enfant. De plus, ces enregistrements ouvrent la possibilité de poser de nouvelles questions, *e.g.*, les propriétés des pleurs, qui contiennent des ultrasons modulant la réponse hémodynamique de la mère (DOI, SULPIZIO, ESPOSITO et al. 2019).

L'utilisation des technologies portables pour récolter des données massives et plus écologiques est une des pistes sur lesquelles voudraient s'appuyer les chercheurs en développement précoce de l'enfance afin de dépasser les problèmes de reproductibilité que la communauté rencontre (LICHAND, LEAL NETO, PHUKA et al. 2022). Des enregistrements continus devraient permettre non seulement de voir l'évolution et le développement des paramètres acoustiques de l'enfant jouant un rôle clé dans la maîtrise du langage, mais aussi de distinguer entre les vocalisations des enfants sans troubles et ceux ayant un retard ou un trouble développemental. Les analyses pourraient alors se faire à des échelles sans précédent (D. K. OLLER, P. NIYOGI, GRAY et al. 2010). Les avancées récentes des réseaux de neurones dans le domaine du son,

1. Introduction – 1.2. Réponse au premier problème : la détection des vocalisations

qui sont très prometteurs pour aider à la détection de problèmes de santé et troubles développementaux, voient leur application limitée du fait de données accessibles en trop petit nombre (MILLING, POKORNY, BARTL-POKORNY et al. 2022).

Néanmoins, si des enregistrements continus nous permettront de constituer une nouvelle base de données de vocalisations de bébés plus riche, à partir de laquelle nous pourrions caractériser plus finement les classes de vocalisation et leur évolution, il nous faut d’abord extraire les vocalisations des enregistrements. Encore récemment, des bases de données comme Meg CYCHOSZ, SEIDL, Elika BERGELSON et al. 2019 ont été construites en s’appuyant sur du *crowdsourcing* pour faire le travail d’annotation. Ce travail laborieux et chronophage oblige de passer par de la main d’œuvre extérieure quand les enregistrements deviennent massifs. Mais cette solution n’est pas optimale pour annoter des vocalisations à un grain plus fin, et pose des soucis sur la qualité des annotations, les annotateurs faisant des erreurs et n’étant pas toujours d’accord entre eux (Margaret CYCHOSZ, CRISTIA, Elika BERGELSON et al. 2021).

On pourrait remplacer ce travail manuel par des méthodes automatiques plus simples et efficaces, pour lesquelles le risque d’erreur peut être minimisé. De plus, l’erreur en question ne changerait plus d’un individu à un autre. On pourrait estimer le biais de la méthode et travailler en connaissance de cause. La segmentation des vocalisations des enregistrements continus des bébés est le premier obstacle qui se pose dans cette thèse. La construction d’une méthodologie permettant de faire ce travail automatiquement en est la première contribution.

1.2. Réponse au premier problème : la détection des vocalisations

La première tâche est d’extraire les vocalisations des enregistrements continus. Ce sont des enregistrements continus, massifs et bruités, récoltés dans l’environnement naturel du bébé, la maison, où l’on retrouve énormément de bruits parasites (travail ménager, discussion des parents, radio, télévision, sons de la rue) et de moments sans vocalisations. Les segments qui nous intéressent, quand l’enfant vocalise, apparaissent *in fine* en faible quantité. Ce sont ces segments, les vocalisations, qui constitueront la base de données sur laquelle on pourra travailler, que l’on veut trouver et extraire.

Comme les enregistrements sont massifs, il nous faut une méthode qui soit capable de traiter efficacement et rapidement une quantité importante de données de grande dimension comme du signal sonore. De plus, cette méthode doit être robuste au bruit.

Le *deep learning* est une méthode qui a fait ses preuves dans plusieurs domaines pour lesquels elle s’est imposée pour résoudre ce type de problème. Cela en fait le parfait candidat pour notre problème de détection des vocalisations dans des enregistrements continus. Des problèmes techniques devaient être levés pour pouvoir l’utiliser étant donné notre contexte : notre base d’apprentissage est restreinte et nos capacités de calcul sont limitées. On résout ces problèmes dans notre première contribution.

1.2.1. Revue de littérature

Les méthodes classiques de détection et de classification d'un signal peuvent se décomposer en trois étapes. (1) Filtration du signal, (2) extraction de caractéristiques dans un vecteur qui représente le signal, et (3) utilisation de ce vecteur en entrée d'un modèle de classification. À partir de cette structure commune, de multiples stratégies sont possibles, adaptées à chaque problème (DIETRICH, PALM, RIEDE et al. 2004; X. XIA, TOGNERI, SOHEL et al. 2018; T. T. NGUYEN, T. T. T. NGUYEN, PHAM et al. 2016; T. T. T. NGUYEN, T. T. NGUYEN, LIEW et al. 2018; STRISCIUGLIO, VENTO et PETKOV 2019). À chaque fois, la question de la représentation est primordiale. Le *deep learning*, comme apprentissage de représentation, permet de dépasser ces problèmes de représentation en traitant des données naturelles brutes (Y. BENGIO, COURVILLE et VINCENT 2013). L'apprentissage profond, et plus précisément les réseaux neurones convolutifs (CNN), découvrent automatiquement la représentation utile pour la classification, à partir des données brutes. L'apprentissage de la représentation effectué par le biais des couches cachées trouve des structures dans les données de grande dimension et découvre des structures hiérarchiques dans les signaux naturels (Yann LECUN, Y. BENGIO et G. HINTON 2015). Cette structure hiérarchique représente bien les données sonores, notamment la parole : des phonèmes aux phrases, en passant par les syllabes et les mots (FARABET, COUPRIE, NAJMAN et al. 2013). Des structures hiérarchiques similaires peuvent être supposées chez d'autres espèces pour différents systèmes de communication vocale, ce qui rendrait un CNN utile pour extraire des caractéristiques hiérarchiques apprises des données brutes et ainsi réduire le risque de créer des représentations anthropocentriques (PRAT 2019).

Les réseaux de neurones ont déjà prouvé leur intérêt pour résoudre des tâches proches de la nôtre (GU, Zhenhua WANG, KUEN et al. 2017). Alors que nous devons détecter des événements, on sait que les CNN permettent d'excellents résultats en détection d'objet (GIRSHICK, DONAHUE, DARRELL et al. 2014; REN, HE, GIRSHICK et al. 2016). Ils sont par ailleurs capables de modéliser la dimension temporelle, que ce soit pour des vidéos (JI, W. XU, M. YANG et al. 2013; TRAN, BOURDEV, FERGUS et al. 2015) ou des séries temporelles plus larges (ISMAIL FAWAZ, FORESTIER, WEBER et al. 2019). On trouve les bons résultats des CNN pour traiter différentes données audio (Jongpil LEE, T. KIM, PARK et al. 2017; PALANISAMY, SINGHANIA et YAO 2020), que ce soit de la classification musicale (W. ZHANG, LEI, Xiangmin XU et al. 2016; CHOI, Gyorgy FAZEKAS, Mark SANDLER et al. 2017; Jongpil LEE, PARK, K. L. KIM et al. 2017; DONG 2018; M.-T. CHEN, B.-J. LI et CHI 2019), la reconnaissance de parole (ABDEL-HAMID, MOHAMED, JIANG et al. 2014; Z. ZHU, J. H. ENGEL et HANNUN 2016; Y. ZHANG, PEZESHKI, BRAKEL et al. 2017; SCHINDLER, LIDY et RAUBER 2018) ou même de la génération de parole (MEHRI, K. KUMAR, GULRAJANI et al. 2017), notamment WaveNet (OORD, DIELEMAN, ZEN et al. 2016) qui peut par ailleurs être utilisé pour d'autres tâches de synthèse (J. ENGEL, RESNICK, A. ROBERTS et al. 2017). On retrouve ces résultats pour des problèmes de bioacoustique (PANDEYA, D. KIM et Joonwhoan LEE 2018; STOWELL, STYLIANOU, WOOD et al. 2018; BERGLER, SCHRÖTER, CHENG et al. 2019; OIKARINEN, SRINIVASAN, MEISNER et al. 2019) et sur des tâches de classification

1. Introduction – 1.2. Réponse au premier problème : la détection des vocalisations

de sons de l'environnement (PICZAK 2015; TOKOZUME et HARADA 2017; KHAMPARIA, GUPTA, N. G. NGUYEN et al. 2019; Xinyu LI, CHEBIYYAM et KIRCHHOFF 2019; GUZHOV, RAUE, HEES et al. 2020) démontrant la capacité d'un réseau de neurones à traiter efficacement des données massives et fortement bruitées. Ces modèles sont adaptés pour traiter des signaux bruyants et à grande échelle (S. CHEN, ZHENG, L. YANG et al. 2019).

1.2.2. Retour sur l'apprentissage statistique

Tous ces résultats indiquent que les CNN sont de bons candidats pour la modélisation de notre tâche. Ils sont rapides pour l'inférence, ils s'adaptent aux données massives, ils sont robustes au bruit et ont de bons résultats de prédiction. En outre, cette approche évite le problème de la construction d'une représentation du signal à la main, ce qui rend le *pipeline* de traitement du signal plus accessible et généralisable. Elle évite le recours à un traitement manuel souvent utilisé dans les approches traditionnelles. Il s'agit d'une stratégie *end-to-end*, on part des données brutes et on produit la sortie.

Concrètement, on décide d'apprendre un modèle $f(\mathbf{x}; \boldsymbol{\theta})$ qui prédise la présence d'une vocalisation y dans un segment sonore \mathbf{x} , *i.e.*, la distribution $p(y|\mathbf{x})$. Étant donné les arguments que l'on vient d'exposer, on décide que f sera un réseau de neurones. Il nous reste à déterminer les paramètres du modèle $\boldsymbol{\theta}$.

Pour cela, nous avons à disposition des exemples de vocalisation étiquetés. Ces vocalisations sont des exemples tirés de la distribution d'intérêt $p(y|\mathbf{x})$ que l'on ne connaît pas. L'ensemble forme la distribution empirique $\hat{p}(y|\mathbf{x})$. Nous faisons de l'apprentissage statistique : à partir de \hat{p} , on apprend un modèle f qui estime la distribution p . Cette estimation est donnée par $f(\mathbf{x}; \boldsymbol{\theta}) = p_{\text{modèle}}(y|\mathbf{x})$. Le modèle produit \hat{y} , sa prédiction de y pour des données \mathbf{x} . On a ainsi, pour un segment \mathbf{x} , la probabilité selon le modèle qu'il y ait une vocalisation ou pas.

Une tâche d'apprentissage n'est pas une tâche de pure optimisation. Notre modèle apprend en minimisant sa fonction objectif $J(\boldsymbol{\theta})$, qui mesure l'adéquation aux données d'entraînement des prédictions du modèle de paramètre $\boldsymbol{\theta}$. Notre but n'est pas tant de minimiser J que de réduire l'écart entre \hat{y} et y . Plus précisément, on ne veut pas réduire l'écart entre \hat{y} et y seulement pour les vocalisations que nous connaissons, celles qui constituent notre ensemble d'entraînement de \hat{p} , on veut réduire cet écart pour celles que nous ne connaissons pas, que le modèle n'a jamais vu, tirée dans la distribution des données p . Pour avoir une idée de cet écart-là, nous faisons une partition de notre ensemble de données étiquetées en trois sous-ensembles : un ensemble d'entraînement, un de validation et un de test. Nous utilisons les deux premiers lors de l'apprentissage. Le troisième est laissé de côté jusqu'au moment où on estime la capacité de généralisation de notre modèle, l'écart entre \hat{y} et y sur l'ensemble de test. C'est ce qui différencie un travail d'apprentissage d'un travail de pure optimisation : on optimise indirectement le score qui nous intéresse, notre précision de prédiction sur l'ensemble de test, en optimisant une autre valeur, la fonction de perte du modèle calculée sur la distribution empirique.

1. Introduction – 1.2. Réponse au premier problème : la détection des vocalisations

La minimisation d'un score estimé sur la distribution empirique risque d'aboutir à une situation de sur-apprentissage. Des modèles avec une capacité importante, comme c'est le cas des réseaux de neurones, sont capables d'apprendre par cœur un ensemble d'entraînement, sans avoir pour autant de capacité de généralisation. Ce modèle serait inutile, car incapable d'être utilisé sur des données non vues. Toute la tâche consiste donc à réussir à effectivement minimiser notre erreur de généralisation à partir d'un ensemble de données étiquetées potentiellement assez réduit.

1.2.3. Points à résoudre pour son utilisation dans notre cas

Deux des principaux facteurs de la réussite du *deep learning* sont l'explosion des capacités de calcul et la construction d'ensembles de données étiquetées massifs. Ces deux points sont justement les éléments qui nous manquent. On trouve une bonne description de ces verrous techniques chez STOWELL 2022 pour l'utilisation des réseaux de neurones dans la bioacoustique, où la communauté doit travailler avec des ensembles de données SUNG (*Small, Unbalanced, Noisy, but Genuine*) (ARNAUD, PELLEGRINO, KEENAN et al. 2023).

Des données massives sont la clé de voûte d'un apprentissage de représentation réussi, mais nous n'avons pas à disposition des ensembles de vocalisations de bébé assez importants permettant l'apprentissage d'un réseau de neurones de zéro. Nous avons besoin d'une méthodologie permettant à un réseau de neurones d'apprendre à partir d'un ensemble d'entraînement réduit (≈ 60 minutes) et des moyens de calcul raisonnable. Il devra malgré tout être capable de traiter des fichiers audios enregistrés dans différentes conditions (orientation du microphone, position du sujet par rapport au microphone) et dans des conditions non contrôlées (bruits de fond). On cherche à ce que notre méthodologie soit flexible pour être potentiellement utilisée sur des tâches similaires, mais pour des espèces différentes.

Le Chapitre 3 est la première contribution de la thèse et propose une méthodologie pour dépasser ces limites. Pour supporter cela, nous avons testé notre *pipeline* sur deux problèmes différents, afin de montrer sa capacité d'adaptation à des contextes différents. Dans les deux cas, ce ne sont pas des données synthétiques, mais des problèmes réels.

Dans le premier cas, nous traitons un problème de bioacoustique : un groupe de babouin a été enregistré en continu pendant un mois dans son habitat, un centre de primatologie. Deux micros ont été placés à proximité de l'enclos et ont enregistré le groupe en permanence. Les singes sont en semi-liberté : ils peuvent bouger librement au sein de leur enclos. Il en résulte que les sources (les singes) sont en constant mouvement par rapport aux micros. À cela s'ajoute tous les bruits d'enregistrements sonores faits en extérieur.

La deuxième étude s'inscrit plus directement dans le sujet de cette thèse : des enfants ont été enregistrés à intervalles réguliers, trois jours par mois, pendant un an. Les parents ont placé un micro à proximité de l'enfant afin d'enregistrer d'éventuelles vocalisations. Les micros changeaient ainsi d'orientation et de positionnement par rapport à la source potentiellement à chaque nouvel enregistrement. Là encore, les

1. Introduction – 1.3. Réponse au deuxième problème : la représentation du signal

enregistrements sont massifs, avec beaucoup de sons parasites et le signal d'intérêt en faible quantité par rapport à la durée des enregistrements continus.

Le Chapitre 2 introduit plus en détail l'étude, les éléments constitutifs de la méthodologie permettant l'apprentissage d'un réseau de neurones dans ces conditions.

1.3. Réponse au deuxième problème : la représentation du signal

Une fois que l'on a été capable de construire les bases de données de vocalisations sur lesquelles on va travailler, on se pose la question de la représentation du signal. C'est en effet la question classique en traitement du signal, question qui est dépendante du problème à résoudre. Quelle est la bonne façon de représenter le signal étant donné la question que l'on se pose? Il convient de choisir la bonne représentation, permettant de mettre en lumière l'information que l'on recherche, tout en restant dans une dimension qui nous permette de travailler.

Nous avons choisi ici une approche topologique. L'Analyse Topologique des Données (TDA) est un champ de recherche en fort développement. Elle s'appuie sur des arguments théoriques solides et intéressants, et elle a été utilisée dans différents contextes. De plus, la TDA, par sa procédure et son formalisme, nous permet de faire un lien avec une hypothèse géométrique (de type distribution autour d'une variété dans l'espace des sons), et par là avec la physique du signal.

1.3.1. La physique du son

Un signal sonore $x(t)$ est une perturbation périodique de la pression de l'air (on se restreint pour nous au cas où le son est véhiculé dans l'air), produit par un objet vibratoire (SUEUR 2018). Une source, par exemple les poumons d'un individu dans le cas de la parole, émet un flux d'air. Celui-ci passe les cordes vocales, l'excitateur, suit le conduit qui va du pharynx aux lèvres, le résonateur. L'onde sonore se propage dans le médium, l'air, mettant les particules en mouvement. L'onde en question a une amplitude qui fait varier les particules d'air le long de sa propagation, jusqu'au récepteur, et a une durée, plus ou moins longue. Cette onde est un son.

Les vocalisations ne vont pas occuper l'intégralité de l'espace sonore, car elles n'utilisent pas l'intégralité des gammes de sons existants. On ne vocalise par exemple pas en haute fréquence. Des contraintes physiques et motrices nous empêchent d'utiliser l'intégralité des gammes de son qui existent. On ne produit que dans un sous-espace de l'espace des sons. L'anatomie et la physiologie du conduit vocal et des cordes vocales structurent la gamme des possibles (BOË, BERTHOMMIER, LEGOU et al. 2017). En plus de l'anatomie, qui conditionne déjà les sons produits possibles, les caractéristiques articulatoires de l'appareil buco-phonatoire façonnent l'étendue de la gamme des sons que l'humain peut produire (FAGOT, BOË, BERTHOMMIER et al. 2019). L'utilisation de la langue et des lèvres permet de modifier la cavité buccale et

1. Introduction – 1.3. Réponse au deuxième problème : la représentation du signal

contraindre le conduit vocal. Ce contrôle moteur se fait à partir de quelques paramètres : des modèles articulatoires géométriques de l'enfant (qui n'ont par ailleurs pas de données expérimentales permettant de les valider) sont par exemple contrôlés à partir de sept paramètres articulatoires (BOË, BADIN, MÉNARD et al. 2013). Si les capacités motrices évoluent au cours de la première année de vie, il devrait en être de même de la possibilité de modifier la forme du conduit vocal, et donc de la répartition des productions vocales dans l'espace des vocalisations. Celles-ci devraient évoluer à mesure que l'enfant apprend à maîtriser son appareil.

1.3.2. Hypothèse géométrique de la variété

Le degré de liberté limité de la production vocale humaine, du fait de ses contraintes physiologiques, permet de faire le lien avec une hypothèse communément admise en apprentissage, l'hypothèse de la variété (GOODFELLOW, Y. BENGIO et COURVILLE 2016) : les ensembles de données de haute dimension que l'on rencontre se situeraient proche d'une variété sous-jacente de plus basse dimension dans cet espace de haute dimension.

Alors que l'on récupère des données de grande dimension quand on digitalise les signaux sonores, celles-ci auraient une dimension intrinsèque inférieure (BERENFELD 2022). La dimension intrinsèque des données est liée à un petit ensemble de paramètres reliés aux caractéristiques physiques, au nombre de paramètres implicites qui gouvernent l'ensemble des données. Comme on s'intéresse à la production vocale des enfants au fil de la première année de vie et que les productions vocales sont déterminées par des contraintes physiques, on espère tirer une information en s'intéressant à la structure de l'objet géométrique sous-jacent.

Pour une vocalisation, on aurait ainsi $X = \{\mathbf{x}_i\}_{i=1}^n$ inclus dans un espace ambiant $\mathcal{X} \in \mathbb{R}^D$, que l'on considère générée d'une distribution \mathbb{P} . On fait l'hypothèse que le support de \mathbb{P} est le fermé d'une variété de dimension d plongée dans \mathcal{X} pour $d < D$ inconnu (formellement, le support est toujours un ensemble fermé, on parle généralement de sous-variété ou on considère le support comme le fermé de la variété) (TINARRAGE 2020). La dimension intrinsèque d renvoie à la dimension de la variété autour de laquelle les données se situent, une dimension inférieure à la dimension de l'espace ambiant D dans laquelle la variété est plongée. Nous intéresser à la dimension intrinsèque au lieu de la dimension de l'espace ambiant permet d'éviter l'écueil de la malédiction de la dimension. Cela est d'autant plus important si on garde à l'esprit le troisième problème, la modélisation. Une structure de basse dimension émerge à cause de contraintes provenant de lois physiques et la masse de probabilité se concentre dans des régions qui ont une dimension inférieure à celle de l'espace ambiant.

On retrouve l'hypothèse de la variété dans de nombreux champs d'applications de la statistique moderne, surtout dans ceux confrontés à des données massives de grande dimension (BERENFELD 2022). Elle permettrait d'expliquer le succès des algorithmes de réduction de dimension. TENENBAUM, V. DE SILVA et LANGFORD 2000 s'appuient dessus pour proposer l'algorithme Isomap supposé traiter des données massives. On

1. Introduction – 1.3. Réponse au deuxième problème : la représentation du signal

retrouve des idées similaires chez MCINNES, HEALY et MELVILLE 2020. Le succès des méthodes de réduction de dimension, notamment de réduction de dimension non-linéaire, participe à l'importance de l'hypothèse dans bien des domaines d'application de la statistique moderne, sans être pour autant prouvé (FEFFERMAN, MITTER et NARAYANAN 2016).

Celle-ci permettrait également d'expliquer en partie les succès des réseaux de neurones, dans d'autres champs comme la reconnaissance d'image (POPE, C. ZHU, ABDELKADER et al. 2021), mais aussi pour la reconnaissance de parole et le traitement du signal, via l'apprentissage de représentation (Y. BENGIO, COURVILLE et VINCENT 2013). Les représentations hiérarchiques que le réseau apprend au fil de la succession de couches, dont on parlait dans la partie 1.2.2, correspondent à un système de coordonnées de la dimension intrinsèque de la variété. L'idée de l'existence d'une structure de basse dimension qui provient des contraintes physiques est assez classique en analyse de la parole (ERRITY et MCKENNA 2006), ainsi que dans d'autres champs de l'acoustique (COHEN, LINDENBAUM et GANNOT 2023). Selon les problèmes, on raffine d'ailleurs aujourd'hui l'hypothèse : on admet une régularité moindre pour la variété \mathcal{M} en la considérant immergée (plongement sans l'hypothèse d'homéomorphisme) (TINARRAGE 2020), ou on teste l'hypothèse d'union de variétés pour que différentes régions du support des données aient différentes quantités de facteurs de variation (B. C. BROWN, CATERINI, ROSS et al. 2023).

1.3.3. Lien avec la TDA

On a donc nos données dans un espace ambiant qui est un sous-ensemble fini d'un espace euclidien. Elles sont des échantillons d'un objet géométrique régulier, une sous-variété, de dimension inférieure. La dimension intrinsèque de nos données, qui s'exprime par la dimension de cette variété, est conditionnée par les contraintes physiques de production de ces données, dans le cas l'appareil buco-phonatoire de l'enfant et les paramètres permettant de le modifier. Il serait intéressant d'estimer des propriétés de \mathcal{M} à partir de X : cela réduirait la dimension de nos données tout en ouvrant possiblement des pistes de réflexion sur l'évolution des contraintes physiques portant sur la production vocale des enfants. La TDA nous permet cela.

La TDA s'intéresse à la forme des données (Frédéric CHAZAL et MICHEL 2021). Elle s'appuie sur les outils de la topologie algébrique (CARLSSON 2009; CARLSSON, ISHKHANOV, Vin DE SILVA et al. 2008). Avec la TDA, on ajoute une teinte probabiliste à la topologie computationnelle afin de décrire un espace sous-jacent à partir d'un échantillon aléatoire (Partha NIYOGI, SMALE et WEINBERGER 2008; MICHEL 2015).

On crée ainsi de nouveaux descripteurs des données, avec une information topologique sur la forme de l'espace sous-jacent en couvrant la variété de simplexes, représentée par l'échantillon que l'on a, sur lequel on calcule les homologies simpliciales persistantes (TROFIMOV, CHERNIAVSKII, TULCHINSKII et al. 2023). On retrouve alors le point antérieur qui motive l'utilisation de la TDA dans ce travail. On fait l'hypothèse de la variété. Ainsi, les données que l'on a dans un espace ambiant de grande dimension sont concentrées près d'une variété de basse dimension. La TDA consistant

1. Introduction – 1.3. Réponse au deuxième problème : la représentation du signal

à décrire numériquement les propriétés topologiques multi-échelles des distributions de données par l'analyse de ses échantillons, elle est le bon moyen d'inférer des descripteurs de la forme de la variété sous-jacente des données à partir des échantillons que l'on a. La variété sous-jacente en question étant déterminée par des contraintes physiques, comme dans le contexte de la production de parole, en apprendre plus sur la forme des données permet de mieux caractériser les contraintes en question. Ainsi, prendre en compte des caractéristiques topologiques des productions vocales et leur évolution au fil du temps, permet de mieux caractériser l'évolution des productions vocales des bébés au cours de leur première année de vie. Si pour le premier problème, nous étions principalement intéressés à prédire un *pattern* particulier de la manière la plus efficace possible, tâche pour laquelle l'utilisation d'un réseau de neurone était particulièrement adapté, nous nous intéressons dans le deuxième problème à une façon de comprendre la structure latente sous-jacente derrière ce *pattern*.

Les méthodes de la TDA capturent naturellement les propriétés des variétés des données sur des échelles de distance multiples et sont une bonne balance entre approches globales et locales. De plus, les descripteurs topologiques sont intéressants et ont de bonnes propriétés théoriques (COHEN-STEINER, H. EDELSBRUNNER et HARER 2007; Frederic CHAZAL, Vin DE SILVA, GLISSE et al. 2013).

1.3.4. Revue de littérature

La TDA a déjà été utilisée dans plusieurs champs, montrant son intérêt. L'approche topologique a prouvé notamment son utilité dans l'analyse de l'image, pour la détection (PATRANGENARU, BUBENIK, PAIGE et al. 2019), la segmentation (PARIS et F. DURAND 2007) ou l'analyse de forme (CARRIÈRE, OUDOT et OVSJANIKOV 2015). En sciences naturelles, elle est utile pour analyser la structure des protéines (K. XIA et WEI 2014; MAROULAS, MICUCCI et NASRIN 2022), la propriété chimique des molécules (KRISHNAPRIYAN, MONTOYA, HARANCZYK et al. 2021), ou pour étudier la structure atomique d'alliage particulier (MAROULAS, NASRIN et OBALLE 2020). Elle est aussi précieuse pour comparer des représentations, par exemple des graphes (BARBAROSSA et SARDELLITTI 2020a; BARBAROSSA et SARDELLITTI 2020b) ou des représentations apprises par des réseaux (BARANNIKOV, TROFIMOV, BALABIN et al. 2022). En médecine, elle est la première fois utilisée pour aider à la détection du cancer (NICOLAU, LEVINE et CARLSSON 2011). Elle est ensuite fortement utilisée pour l'analyse de données cérébrales, pour comparer les fonctions cérébrales entre sujets sains et sujets pathologiques (H. LEE, CHUNG, KANG, B.-N. KIM et al. 2011; H. LEE, CHUNG, KANG et D. S. LEE 2014; GRACIA-TABUENCA, DÍAZ-PATIÑO, ARELIO et al. 2020), ainsi que pour étudier le signal cérébral recueilli, que ce soit par fMRI (SALCH, REGALSKI, ABDALLAH et al. 2021) ou EEG (NASRIN, OBALLE, BOOTHE et al. 2019; MAROULAS, MIKE et OBALLE 2019; Xiaoqi XU, DROUGARD et R. N. ROY 2021).

Plus proche du champ de cette thèse, la TDA a déjà montré son intérêt pour traiter diverses séries temporelles (PEREIRA et DE MELLO 2015; SEVERSKY, S. DAVIS et BERGER 2016) et financières (GIDEA et KATZ 2018; YEN et CHEONG 2021). Elle commence à montrer son intérêt en traitement du langage en permettant de s'intéresser au

plongement de mots (HAIM MEIROM et BOBROWSKI 2022). Surtout, les outils de la topologie ont déjà fait leur preuve dans des questions de traitement du son. En effet, plusieurs études ont montré son intérêt pour la classification musicale (J.-Y. LIU, JENG et Y.-H. YANG 2016; SANDERSON, SHUGERMAN, MOLNAR et al. 2017; BERGOMI et BARATÈ 2020), la classification de sons environnementaux (Y. CAO, S. ZHANG, YAN et al. 2019) ou la détection sonore (FIREAIZEN, RON et BOBROWSKI 2022).

Si on est emmené à penser que la TDA est intéressante pour notre problème, étant donné les motivations présentées et les preuves de son intérêt dans d'autres champs, certains proches du nôtre, il n'y a en revanche pas encore d'études d'applications de la TDA au traitement de vocalisations humaines. Le Chapitre 5, la deuxième contribution de cette thèse, s'appuie à démontrer l'intérêt par une preuve empirique de l'approche topologique pour l'analyse et la classification de productions vocales humaines. On étudie aussi les différents choix qui peuvent être faits, notamment de représentation du signal sonore et de vectorisation des diagrammes de persistance, et on quantifie les conséquences de ces choix. Le Chapitre 4 introduit plus en détail l'étude, notamment les objets théoriques de la TDA.

1.4. Réponse au troisième problème : la classification des vocalisations

Nous avons extrait les vocalisations des enregistrements continus. Nous avons analysé la meilleure façon de représenter nos données pour notre problème. Il nous reste donc à résoudre le troisième et dernier problème, celui de la classification des vocalisations. Comme on veut proposer une catégorisation plus fine des productions vocales sur la première année de vie, on procède à une classification non-supervisée de ces productions.

La quantification de l'incertitude, une incertitude possiblement importante pour un sujet comme celui du développement du langage sur la première année de vie avec un échantillon d'enfant réduit, doit être prise en compte. On décide pour cela de passer par une approche bayésienne non-paramétrique. Ce choix nous permet d'avoir une proposition de partition basée sur un modèle, plus riche d'un point de vue informatif et offrant des possibilités de prolongement plus importantes. Le côté non-paramétrique nous laisse estimer la complexité du modèle directement sur l'échantillon et ainsi déterminer le nombre de classes sur la première année de vie.

1.4.1. Présentation du paradigme bayésien

L'approche bayésienne est une approche par modélisation, qui utilise le langage de la théorie des probabilités pour construire un modèle et représenter toutes les formes d'incertitude dans le modèle. Un modèle spécifie complètement la distribution conditionnelle de ce qui est observé (les données), étant donné les quantités non-observées (les paramètres). Une distribution a priori spécifie la distribution de toutes les quantités non-observées. Une distribution a posteriori renverse l'ordre de

conditionnement et donne la distribution des quantités non-observées sachant les quantités observées.

On définit donc la probabilité des données comme une fonction des paramètres du modèle, c'est la vraisemblance $p(x|\theta)$. On spécifie ainsi comment les données sont produites. Il manque à donner un peu plus d'informations sur ces paramètres. Dans le paradigme bayésien, on considère les paramètres du modèle comme aléatoire. Plus qu'aléatoire, ils sont incertains. En cela, la statistique bayésienne permet de modéliser clairement notre incertitude, en utilisant le langage des probabilités. C'est l'a priori $p(\theta)$, qui revient à poser une distribution sur les paramètres, donnant donc une information sur leur espace de définition. C'est aussi le moment où l'on peut intégrer un savoir expert.

Cela emmène à un traitement séquentiel des données. On combine les données observées (la vraisemblance) avec l'état initial de nos connaissances (l'a priori) et on met à jour nos connaissances en calculant l'a posteriori. L'a posteriori, qui inverse l'ordre de conditionnement de la vraisemblance, est calculé par la célèbre formule de Bayes,

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)},$$

et donne donc la probabilité des paramètres sachant les données. Le modèle, une représentation compacte des données que l'on peut observer, permet donc d'intégrer nos connaissances dans l'a priori et les met à jour dans l'a posteriori en les confrontant aux données. Celui-ci peut alors être utilisé pour faire des prédictions.

L'approche bayésienne est particulièrement adaptée pour la tâche que l'on doit faire. Grâce à l'a priori et aux choix de modélisation, on peut intégrer un savoir expert et des hypothèses développementales. De plus, nos données étant les vocalisations des bébés, enregistrés chez eux, on s'attend à une certaine hétérogénéité. Chaque enfant est unique, faire de l'inférence au niveau de la population risquerait de nous faire perdre cette diversité et la structure latente des données. L'approche bayésienne hiérarchique permet d'éviter cela, en traitant chaque enfant individuellement. Enfin, l'inférence bayésienne se base sur l'a posteriori de paramètres. La prédiction est elle-même probabiliste et donne une information sur l'incertitude liée à cette prédiction. Cela est d'autant plus important pour des questions comme la nôtre, où on attend une variabilité importante et où la taille de l'échantillon est réduite.

Les paramètres du modèle permettent de capturer une information sur les données. Dans un modèle paramétrique, ils sont en nombre fini. Si l'on revient sur le problème de classification de vocalisations, on fait l'hypothèse que les vocalisations sont tirées dans une distribution de mélange, où chaque composante de mélange correspond à une classe. On apprend notre modèle sur nos données, attribuant chaque vocalisation à une classe du mélange. On fait ainsi la partition de notre ensemble, regroupant les vocalisations les plus similaires.

Pour des tâches de classification non-supervisée, il est courant d'utiliser des modèles de mélange, même quand la dimension est importante (BOUYEYRON, GIRARD et SCHMID 2007). On peut utiliser l'approche bayésienne (GAIFFAS et MICHEL 2014),

même en grande dimension (KOCK, KLEIN et NOTT 2022). Mais à chaque fois, l'estimation de K , le nombre de classes du modèle, repose sur une euristique. On passe par de la sélection de modèle, on compare différentes hypothèses. L'approche bayésienne non-paramétrique permet d'intégrer l'estimation de la complexité au modèle. Au lieu d'estimer plusieurs modèles de complexités différentes, on estime un seul modèle où K est un paramètre comme les autres. La méthode bayésienne non-paramétrique permet d'adapter la complexité du modèle, donc K , aux données conjointement à l'estimation des autres paramètres.

1.4.2. Le non-paramétrique

Alors que sur un modèle bayésien classique, nous fixons a priori la dimension de l'espace des variables latentes, un modèle non-paramétrique ou semi-paramétrique est un modèle dans lequel au moins un des paramètres est de dimension infinie. En ce sens, on peut aussi parler de modèle de dimension infinie quand on parle de modèle bayésien non-paramétrique. Dans le cadre d'un modèle de mélange, il n'est alors pas nécessaire de spécifier le nombre d'éléments de mélange. Le nombre de classes sera directement estimé sur les données. On a $K \rightarrow \infty$ et, à la vue de l'échantillon, seulement un nombre fini de dimensions est sélectionné pour expliquer les données. La partition nous donne le nombre de classes.

Les modèles non-paramétriques sont ainsi plus flexibles que les paramétriques. Dans l'approche paramétrique, on fait une hypothèse sur le processus qui produit les données,

$$X|\theta \sim p_\theta,$$

pour $\theta \in \Theta \subset \mathbb{R}^d$. On pose ensuite un a priori sur θ , $\theta \sim \pi$. On peut réécrire notre modèle comme $X|p \sim p$ pour $p \sim \Pi$, où Π est la distribution a priori sur l'ensemble de toutes les densités possibles, ayant la propriété $\Pi(\{p_\theta : \theta \in \Theta\}) = 1$. Aussi, la modélisation paramétrique revient à mettre un a priori qui met une probabilité 1 à un petit sous-ensemble de toutes les densités possibles (GHOSAL et VAN DER VAART 2017). C'est *in fine* un très fort a priori que la méthode non-paramétrique permet de dépasser.

Ce passage du paramétrique au non-paramétrique, de la dimension finie à la dimension infinie, se joue sur l'a priori $p(\theta)$. Dans le premier cas, on spécifie les variables latentes et leur cardinalité. Dans le second cas, la structure latente peut grandir avec les données. Choisir un modèle de mélange non-paramétrique nous permet ainsi de faire de la classification sans avoir à spécifier le nombre de classes a priori. Cette valeur est inférée sur les données. Concrètement, cela est fait en choisissant un a priori adapté qui peut évoluer. Pour cela, on utilisera des objets particuliers comme a priori, des processus stochastiques. Le processus de Dirichlet est l'a priori adapté qui permet cette croissance des classes sur les données.

De plus, cet a priori se retrouve quand on calcule la distribution prédictive a posteriori. Le nombre de classes peut donc augmenter à l'ajout de nouvelles observations. De nouveaux éléments de mélange, de nouvelles classes, sont créées par le modèle si les données à prédire ne rentrent dans aucune des classes vues jusqu'alors. Dans

notre cas particulier de classification non-supervisée, les vocalisations sur lesquelles on apprend notre modèle sont toutes assignées à une classe. Une fois l'apprentissage terminé, on peut utiliser le modèle pour faire de la prédiction et assigner de nouvelles vocalisations aux différentes classes détectées. La flexibilité du modèle non-paramétrique offre la possibilité d'assigner un nouveau point à une nouvelle classe, non-vue jusqu'à présent. En cela, l'approche est plus robuste.

Cette approche est particulièrement intéressante pour notre problème. Certes, on a une information a priori dans la littérature sur les types de vocalisation produite au cours de la première année. Cette information pourra d'ailleurs être intégrée grâce à l'a priori. Néanmoins, les enregistrements massifs, et notamment de vocalisation spontanée, n'ont pas eu l'occasion d'être beaucoup traité jusqu'à présent. On a donc une incertitude quant au nombre de classes à spécifier. En cela, avoir un modèle dont la complexité s'adapte aux données, sans que cela ne passe par une inflation de paramètres (C. RASMUSSEN et GHAHRAMANI 2000), nous paraît comme étant le meilleur choix possible, au lieu de passer par une comparaison de multiples modèles.

De plus, les productions vont évoluer au fil du temps et des mois, et des productions qui n'étaient pas produites à un moment le seront plus tard, au fil de l'apprentissage de l'enfant. La modélisation Bayésienne non-paramétrique permet de prendre en compte cela, en laissant augmenter le nombre d'éléments de mélange à l'ajout de nouvelles données observées. Une approche bayésienne non-paramétrique est donc le choix le plus adapté pour la question à traiter.

1.4.3. Revue de littérature

L'approche bayésienne non-paramétrique a été utilisée pour répondre à de nombreux problèmes (Peter MÜLLER et QUINTANA 2004) : elle a prouvé son utilité pour modéliser des fonctions inconnues, faire de l'estimation de densité, de la classification, de la modélisation de séries temporelles, représenter les structures hiérarchiques dans les données (GHAHRAMANI 2013; MACÉACHERN 2016). C'est un champ relativement nouveau qui se développe fortement et qui peut potentiellement être appliqué à de nombreux sujets (JARA 2017). Nous concernant, on s'intéresse plus en détail aux processus de Dirichlet, qui se révèlent particulièrement utiles pour les tâches de classification non-supervisée.

Les processus de Dirichlet ont été introduits afin de fournir les outils nécessaires au traitement de problèmes non-paramétriques à l'analyse bayésienne (FERGUSON 1973). Ils ont pu être appliqués comme a priori sur des modèles de mélange (LO 1984), et les solutions algorithmiques existent afin de les utiliser pour estimer directement le nombre de classes sur les données (ESCOBAR 1994). Les processus de Dirichlet ont depuis été généralisés, par les processus Pitman-Yor (PITMAN et YOR 1997), puis par les a priori de type Gibbs (DE BLASI, FAVARO, LIJOI et al. 2015), qui généralisent les deux.

Les algorithmes permettant l'estimation de ces modèles existent, même quand le modèle n'est pas conjugué, que ce soit par MCMC (R. M. NEAL 2000; PAPASPILIOPOULOS et G. O. ROBERTS 2008; KALLI, GRIFFIN et WALKER 2011; ARBEL, DE BLASI et PRÜNSTER

2019; CANALE, CORRADIN et NIPOTI 2021; M. M. ZHANG, WILLIAMSON et PEREZ-CRUZ 2022) ou par inférence variationnelle (D. M. BLEI et Michael I. JORDAN 2004; D. M. BLEI et Michael I. JORDAN 2006; D. M. BLEI, KUCUKELBIR et MCAULIFFE 2017). Ils peuvent alors montrer leur efficacité à modéliser des problèmes dans différents champs, comme en psychologie (GERSHMAN et D. M. BLEI 2012; Y. LI, SCHOFIELD et GÖNEN 2019).

Au moment où le challenge de l'inférence bayésienne est de se mettre à l'échelle (ANGELINO, Matthew James JOHNSON et Ryan P. ADAMS 2016), l'approche non-paramétrique a les outils pour y répondre. Les modèles de mélange avec processus de Dirichlet peuvent ainsi être utilisés pour faire de la classification non-supervisée en grande dimension (GUAN, DY, NIU et al. 2010; MANSINGHKA, SHAFTO, JONAS et al. 2016; MEGUELATI, FONTEZ, HILGERT et al. 2019a; MEGUELATI, FONTEZ, HILGERT et al. 2019b).

La possibilité d'avoir des modèles de mélange non-paramétriques hiérarchiques (TEH, Michael I JORDAN, BEAL et al. 2006) agrandit encore son champ des possibles. L'ajout d'une hiérarchie dans le modèle permet de modéliser la distribution de sujet dans des corpus de documents (D. M. BLEI, A. Y. NG et Michael I. JORDAN 2003), et même analyser du flux de données (MCINERNEY, RANGANATH et D. BLEI 2015; CAI, MITZENMACHER et Ryan P ADAMS 2018), du trafic maritime (GLOAGUEN, CHAPEL, FRIGUET et al. 2023).

Les processus de Dirichlet sont aussi utiles pour modéliser des variables temporelles, que ce soit en combinaison avec des modèles markoviens (FOX, E. B. SUDDERTH, Michael I. JORDAN et al. 2010) ou via le processus de Dirichlet dépendant (CAMPBELL, M. LIU, KULIS et al. 2013). L'introduction de covariables (MACEACHERN 2001) facilite l'inférence sur des cas individuels, permettant d'avoir des distributions prédictives selon les différents niveaux des covariables. Les modèles de régression complètement non-paramétriques (QUINTANA, W. O. JOHNSON, WAETJEN et al. 2016; QUINTANA, Peter MÜLLER, JARA et al. 2022) permettent ainsi de modéliser longitudinalement la distribution de la réponse selon les covariables. On retrouve alors son intérêt en sciences sociales (KUNIHAMA, HALPERN et HERRING 2019) mais aussi et surtout dans le domaine de la santé. Au-delà des classifications des hôpitaux (GUGLIELMI, IEVA, PAGANONI et al. 2014), les modèles avec processus de Dirichlet peuvent être utiles pour l'analyse de données médicales (Xiao LI, GUINDANI, C. S. NG et al. 2021). Les extensions présentées précédemment permettent de voir l'évolution de variables de santé selon différentes dynamiques (LEHMAN, M. J. JOHNSON, NEMATI et al. 2015), entre différentes cohortes (MOLINARI, CREMASCHI, DE IORIO et al. 2022). La possibilité d'intégrer des covariables, de voir l'évolution de variables longitudinales, en faisant une partition des données, permet de développer des modèles pour la santé individualisée et déterminer les meilleurs traitements selon les caractéristiques individuelles du patient (PEDONE 2022; PEDONE, ARGIENTO et STINGO 2023). Grâce à l'approche par modèle, on a en plus les probabilités de réponses spécifiques à la classe et on peut donc identifier les patients qui bénéficieraient le plus d'un traitement spécialisé.

En cela, une approche bayésienne non-paramétrique nous semble être le bon choix de modélisation pour notre problème. La tâche principale est de faire une

1. Introduction – 1.4. Réponse au troisième problème : la classification des vocalisations

classification des vocalisations du bébé au cours de la première année de vie. Un modèle de mélange avec un processus de Dirichlet permettra de faire une partition des données et de proposer une classification des productions vocales, bien que la dimension soit importante. Si nous souhaitons de plus modéliser l'évolution des vocalisations du bébé au cours de la première année et distinguer les différentes classes de vocalisations au fil du temps, on pourra prendre en compte cette dépendance. Comme nous avons enregistré plusieurs bébés, il existe une hiérarchie dans notre ensemble. Les vocalisations sont propres aux individus, mais on s'attend à ce que les classes soient partagées. L'ajout d'une hiérarchie dans le modèle sera possible. Enfin, nous avons une information sur les familles des enfants ainsi que sur les conditions d'accouchement. Il nous sera possible d'intégrer ces covariables au modèle et voir leurs conséquences sur le développement des productions vocales. Bien sûr, chacune de ces étapes est en soi un challenge, et toutes ne seront d'ailleurs pas traitées dans cette thèse, mais ce choix de modélisation permet d'ouvrir la voie et rend possible les extensions futures.

Le Chapitre 7 présente notre contribution permettant de répondre à ce troisième problème. Il s'appuie donc sur la modélisation des vocalisations par un modèle de mélange Bayésien non-paramétrique avec un processus de Dirichlet, qui est introduit plus en détail dans le Chapitre 6 précédent.

Construction de la base d'enregistrement

Une des contributions majeures de cette thèse est la construction de la base d'enregistrements de nouveaux-nés sur une période d'un an. Plusieurs étapes ont été nécessaires avant de pouvoir commencer les enregistrements. Nous avons en premier lieu soumis pour avis au comité d'éthique de l'université d'Aix-Marseille le projet de recherche. Après l'émission d'un avis favorable, nous avons procédé à une analyse d'impact (PIA) et nous sommes déclarés en conformité avec le référentiel de méthodologie de référence MR-004 à la CNIL. Si les premières familles ont été recrutées dans le cercle proche, nous avons par la suite établi une charte de sous-traitance et le registre RGPD avec la maternité de l'hôpital Saint-Joseph, afin de recruter davantage de familles, avant ou juste après la naissance. Le projet a été présenté aux équipes médicales, une page internet et une adresse mail ont été créées pour faciliter la présentation et la prise de contact, un flyer a été imprimé et régulièrement déposé à la maternité afin de diffuser l'information et permettre le recrutement. Les parents intéressés ont ainsi pu nous contacter et, après avoir signé le formulaire de recueil de consentement, lu la notice d'information et posé les questions restantes, commencé les enregistrements quelques jours après la naissance.

Nous avons réalisé ceux-ci à domicile. Nous avons fourni un enregistreur portatif type Zoom H6, sur lequel était connectée la capsule XYH-6, composée de deux microphones unidirectionnels. Ce type de microphone est plus sensible aux signaux provenant directement de l'avant, nous demandions donc de diriger l'enregistreur vers la source, l'enfant. Chaque enregistrement correspond à un fichier WAV individuel au format 44.1 kHz / 16 bit, pour lequel nous avons l'information relative au jour, l'heure, la minute et la seconde du début de l'enregistrement. L'enregistreur étant mobile, nous avons demandé aux parents de faire autant d'enregistrement que possible, en déplaçant l'enregistreur pour suivre l'enfant, dans les moments de repos (sieste, nuit), mais aussi dans les moments d'éveils quand cela était possible et que les parents n'étaient pas en interaction directe avec l'enfant. Les parents étaient maîtres des moments d'enregistrement : ils se lançaient quand les parents appuyaient sur le bouton *play* et s'arrêtaient quand ils appuyaient sur le bouton *stop*. L'enregistreur n'était pas connecté à un quelconque appareil ou serveur, les enregistrements se faisaient en local, sur une carte SD. On laissait donc cet enregistreur pour trois jours et deux nuits, tous les mois pendant les 12 premiers mois, ce qui permet donc de récupérer les enregistrements à chaque mois.

Néanmoins, cette procédure oblige aussi à faire énormément d'aller-retour, pour aller récupérer les enregistreurs dans une famille et les déposer dans d'autres. Nous n'avions initialement que trois enregistreurs, qui devaient donc tourner entre les différentes familles. Nous avons déposé une demande de financement à l'ILCB afin d'acheter plus de matériel, pour pouvoir augmenter le nombre de familles enregistrées simultanément. La demande acceptée, nous avons pu laisser les micros au domicile des familles. En cela, le dispositif a quelque peu changé : les premières familles n'avaient le micro que trois jours par mois alors que les suivantes ont eu le micro en continu. Ces dernières ont ainsi pu faire des enregistrements beaucoup plus

1. Introduction – 1.4. Réponse au troisième problème : la classification des vocalisations

réguliers. Nous les appelions tous les mois pour nous assurer que les enregistrements se passaient bien, répondre à des questions éventuelles et passer changer la carte SD quand celle-ci était pleine.

En plus des enregistrements, nous avons récupéré des informations supplémentaires sur les familles, les conditions d'accouchement et l'enfant : la date de naissance de l'enfant, celle des parents et de la fratrie s'il y en avait, ainsi que leur genre, la CSP de chaque parent, le type d'habitation (appartement, maison), la surface, si l'enfant avait une chambre séparée, et si oui à partir de quand, le moment de naissance de l'enfant en semaine aménorrhée, le poids et la taille à la naissance, les conditions d'accouchement (césarienne, péridurale), la durée de l'accouchement, la garde de l'enfant (maison, crèche, nounou, famille), si l'enfant utilise une tétine et s'il est allaité.

Des enfants sont toujours en cours d'enregistrement au moment où ce manuscrit est écrit. Une fois que ces enfants auront fêté leur premier anniversaire, nous aurons constitué une base d'enregistrement continu d'un an de 16 bébés.

2. Réseau de neurones

Sommaire

2.1. Remise en contexte et résumé de l'étude	39
2.2. Construction d'un réseau de neurones	41
2.2.1. Architecture d'un réseau de neurones	41
2.2.2. Les fonctions d'activation	43
2.2.2.1. Les couches cachées	43
2.2.2.2. La couche de sortie	44
2.2.3. Apprentissage de représentation	45
2.2.3.1. Réseau convolutif	45
2.2.3.2. Transfert d'apprentissage	47
2.2.4. La fonction objectif	50
2.3. Apprentissage	50
2.3.1. La descente de gradient	51
2.3.1.1. Le gradient	51
2.3.1.2. Le gradient stochastique	51
2.3.2. La rétro-propagation	53
2.3.3. L'algorithme de descente de gradient	54
2.4. La régularisation, le sur-apprentissage	56
2.4.1. Data augmentation	56
2.4.2. Batch normalization	57
2.4.3. Dropout	57
2.4.4. Contrainte sur la norme des paramètres	58
2.4.5. Early-stopping	58
2.5. Apprentissage des hyper-paramètres	59
2.5.1. Le choix des hyper-paramètres	59
2.5.2. L'optimisation bayésienne	60

2.1. Remise en contexte et résumé de l'étude

Comme nous le disions dans l'introduction, nous avons enregistré des enfants durant un an. Ces enregistrements sont faits dans un contexte naturel, par les parents, à domicile. Le micro est placé à proximité de l'enfant et enregistre jusqu'à ce que l'un des parents l'éteigne. Il en résulte des enregistrements de durées longues et variables (de quelques secondes à plusieurs heures), bruités (les sons de la maison et des personnes présentes se retrouvent dans les enregistrements) et dans lesquels

seulement quelques segments nous intéressent (les bébés ne vocalisent pas tout le long de l'enregistrement). On a besoin de détecter, dans ces enregistrements audios longs, les moments où les enfants vocalisent, de manière automatique et en perdant le moins de signal possible.

Ce problème se rapproche d'un problème que l'on rencontre en bioacoustique. Les moyens d'enregistrement et de stockages étant de plus en plus accessibles, enregistrer en continu des environnements naturels est facilité. Ces enregistrements continus permettent sans doute de construire des ensembles de vocalisations d'espèce avec une richesse et une variabilité qu'il est difficile d'avoir avec les bases de données actuelles, mais il est d'abord nécessaire de segmenter ce signal, d'extraire les vocalisations afin de construire ces bases de données. La segmentation manuelle de ces enregistrements est longue et laborieuse, demande beaucoup de travail et est sujette aux erreurs.

Le prochain Chapitre, qui est la première contribution de la thèse, propose une méthode afin d'automatiser cette tâche. Le Chapitre présent sert d'introduction à celle-ci, que l'on retrouve au Chapitre 3 suivant.

Notre contribution consiste à proposer un *pipeline* qui apprend un réseau de neurones à partir d'une base d'entraînement d'une heure de vocalisation. Le réseau ainsi appris peut ensuite être utilisé pour détecter et classifier automatiquement et en même temps les vocalisations de l'espèce d'intérêt dans des enregistrements continus massifs. Notre méthodologie permet d'apprendre un réseau de neurones capable de traiter des données réelles et bruitées avec un haut niveau de précision, malgré une base d'entraînement réduite.

Notre méthodologie est adaptable à d'autres espèces que l'humain. Nous l'avons utilisé pour une tâche de détection et classification similaire mais pour un enregistrement continu d'une autre espèce, le babouin de Guinée. Nous avons appris un réseau de neurones grâce à notre *pipeline* à partir d'une base d'entraînement initiale d'un peu plus d'une heure. Nous avons ensuite traité un mois d'enregistrement en continu d'un groupe de singe dans son environnement afin d'en extraire les vocalisations. Cette nouvelle base de données de vocalisations de babouins est librement accessible <https://zenodo.org/record/8239697>. Le code pour reproduire la méthodologie est également accessible <https://gitlab.com/papers4375727/detection-and-classification-of-vocal-productions>.

On introduit ici les différents outils mobilisés dans la méthodologie. On commence par présenter comment s'organise un réseau de neurone et son architecture. Puis, on revoit l'apprentissage et les algorithmes que l'on utilise pour apprendre un réseau de neurones. On voit ensuite l'ensemble des stratégies de régularisation et on termine par le choix des hyper-paramètres du modèle. On s'intéresse aux réseaux appelés à propagation avant (*feedforward*), quand les connexions de la fonction vont seulement dans un sens, de x à y . Dans un souci de clarté, nous ne présentons pas dans ce Chapitre les autres architectures de réseaux de neurones, récurrent ou génératif, car nous ne les utilisons pas dans notre contribution du Chapitre 3

Ce Chapitre et les définitions présentées se basent sur GOODFELLOW, Y. BENGIO et COURVILLE 2016; MURPHY 2022; MURPHY 2023.

2.2. Construction d'un réseau de neurones

On peut décomposer la construction d'un réseau de neurones en cinq éléments : son architecture, les fonctions d'activation des couches, la forme des sorties, la fonction objectif et l'algorithme d'optimisation.

2.2.1. Architecture d'un réseau de neurones

Un réseau de neurones est une fonction, plus précisément une composée de fonctions. Il se compose en couches et chaque couche est composée de plusieurs nœuds. Le nombre de couches définit la profondeur du réseau, le nombre de nœud sa largeur. Ainsi, pour $y = f(\mathbf{x}; \boldsymbol{\theta})$ la fonction de notre modèle, on définit le réseau de neurones comme

$$f(\mathbf{x}) = f^{(n)}(f^{(n-1)}(\dots f^{(1)}(\mathbf{x}) \dots))$$

où $f^{(i)}$ correspond à la $i^{\text{ème}}$ couche.

On appelle la première couche la couche d'entrée, de la dimension de \mathbf{x} , les données d'entrée. Les couches intermédiaires sont les couches cachées. Le rôle des couches cachées, et finalement de l'apprentissage, est de projeter \mathbf{x} dans un espace latent dans lequel il sera ensuite aisé de prédire y (Y. BENGIO, COURVILLE et VINCENT 2013). Cet espace latent est une représentation de \mathbf{x} permettant de déterminer y . La combinaison de couches cachées revient à une extraction de caractéristiques de \mathbf{x} , mais de manière automatique, au sein de la fonction. La profondeur et la largeur du réseau, ainsi que les connections entre les couches, déterminent son architecture.

La dernière couche est la couche de sortie et produit \hat{y} , la prédiction du modèle pour y . Elle détermine aussi l'architecture du réseau. Par exemple, si le travail est de prédire s'il y a une vocalisation ou pas dans \mathbf{x} , la couche de sortie sera de dimension 1, $\hat{y} = 0$ s'il n'y a pas de vocalisation, 1 sinon.

Il est aussi possible, selon la sortie que l'on souhaite produire, de construire des architectures plus complexes à partir de la dernière couche. C'est par exemple le cas dans le Chapitre 3, où l'architecture est duale à partir de l'espace latent afin de produire deux sorties, la prédiction de la présence ou non d'une vocalisation et la prédiction de la classe de la vocalisation. La figure 2.1 représente un réseau de neurone schématique avec cette architecture. Les réseaux qui ne sont pas *feedforward* ont des architectures encore différentes.

Chaque couche du réseau est une fonction qui calcule une transformation des données. Cette transformation est une transformation linéaire des données reçues. Disons que la couche i reçoit la sortie \mathbf{x} de la couche $i - 1$, elle lui applique une transformation linéaire et produit

$$\mathbf{z} = f(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{W}\mathbf{x} + \mathbf{b}.$$

$\boldsymbol{\theta} = (\mathbf{W}, \mathbf{b})$ sont les paramètres de la fonction f . Chaque couche applique une transformation linéaire à son entrée, selon sa propre largeur. On appelle ces couches des couches linéaires ou des couches *fully connected*.

2. Réseau de neurones – 2.2. Construction d'un réseau de neurones

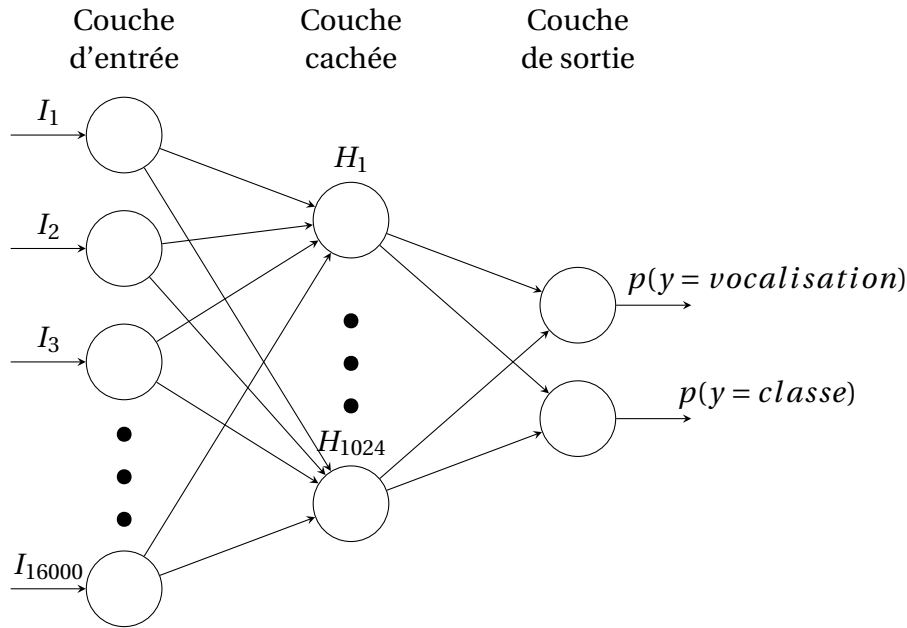


FIGURE 2.1. – Schéma de l'architecture d'un réseau de neurones. Une couche d'entrée prend les données. Une succession de couches cachées applique des transformations non-linéaires aux données, les projetant dans un espace latent. À partir de cette représentation, la couche finale produit \hat{y} . La profondeur, la largeur, les connexions entre les couches déterminent l'architecture du réseau. Ici, l'architecture est composée de deux modules à partir de la représentation apprise, permettant de produire deux sorties, la probabilité que l'entrée, de dimension 16000, contienne une vocalisation, et la probabilité de la classe de la vocalisation.

À cela s'ajoute une fonction d'activation supplémentaire $g(\mathbf{z})$. En effet, après avoir effectué cette transformation, on applique une nouvelle fonction. C'est la fonction d'activation de la couche. Ainsi, à chaque couche, on applique une transformation linéaire aux données puis une fonction d'activation à \mathbf{z} . Au total, une couche est donc une fonction $f(\mathbf{x}) = g(\mathbf{W}^T \mathbf{x} + \mathbf{b})$.

Les paramètres \mathbf{W} et \mathbf{b} de chaque couche sont les paramètres $\boldsymbol{\theta}$ du modèle. Ce sont eux que l'on apprend et que l'on modifie par descente de gradient comme on l'explique plus en détail dans la partie 2.3. Le choix de l'initialisation de ces paramètres est important pour l'apprentissage. En effet, le point de départ détermine pour partie la convergence du réseau de neurones et ses résultats de généralisation, ainsi que sa robustesse aux autres choix de modélisation faits comme le choix des hyper-paramètres (ARPIT, CAMPOS et Y. BENGIO 2019; GLOROT et Y. BENGIO 2010, Mai; HANIN et ROLNICK 2018; HENDRYCKS et GIMPEL 2017; SKORSKI, TEMPERONI et THEOBALD 2021). Les fonctions d'activation g étant souvent non-linéaires, les réseaux de neurones sont des fonctions non-linéaires. Leur optimisation est un problème d'optimisation non-convexe, pour lesquels les paramètres d'initialisation, le point

de départ, sont très importants pour la convergence. Le choix d'initialisation dépend aussi d'autres éléments de l'architecture du réseau, notamment les fonctions d'activation des couches cachées. Le schéma d'initialisation proposé par HE, X. ZHANG, REN et al. 2015, que l'on applique dans le Chapitre 3,

$$\theta_\ell \sim \mathcal{N}\left(0, \frac{2}{n_\ell}\right),$$

où n_ℓ est le nombre de noeuds de la couche ℓ , est censé permettre une meilleure convergence quand les fonctions d'activation des couches sont des ReLU paramétriques (S. K. KUMAR 2017).

2.2.2. Les fonctions d'activation

2.2.2.1. Les couches cachées

L'apprentissage d'un réseau de neurones permet d'estimer une fonction f^* . Plus la composée est grande, plus le modèle est profond, et plus on applique de transformations sur les données, permettant d'apprendre un panel de fonction plus important. Le choix des fonctions d'activation des couches cachées va permettre d'apprendre un certain type de fonction. En choisissant une fonction d'activation non-linéaire, le modèle est alors capable d'apprendre une fonction f^* non-linéaire.

La tâche du réseau de neurones est d'apprendre une représentation des données qui permette de résoudre le problème $y = f^*(\mathbf{x})$. On ne donne pas explicitement la marche à suivre pour construire une représentation de \mathbf{x} , on construit un réseau pouvant apprendre une gamme de fonction assez large, dont une permettrait de résoudre le problème que l'on pose. Si on se contente de fonctions linéaires, le réseau ne sera capable que d'apprendre une représentation linéaire des données, ce qui ne permettrait pas de résoudre bien des problèmes. En effet, la plupart des fonctions que l'on souhaite approximer, y compris celle nous permettant de segmenter le signal, ne sont pas des fonctions linéaires. Pour cela, on ajoute généralement une transformation non-linéaire à la transformation affine initiale.

Ainsi, on ajoute une fonction d'activation non-linéaire $g(\mathbf{z})$ à chaque couche. Dans le cas du Chapitre 3, la fonction g correspond à la fonction ReLU paramétrique (HE, X. ZHANG, REN et al. 2015),

$$f(\mathbf{z}) = \begin{cases} z & \text{si } z > 0 \\ az & \text{si } z \leq 0. \end{cases} \quad (2.1)$$

C'est une extension de la fonction ReLU classique, qui permet un meilleur apprentissage que d'autres fonctions d'activation non-linéaire (GLOROT, BORDES et Y. BENGIO 2011), mais n'annulant pas complètement le neurone quand son entrée est négative. Cela évite des problèmes de disparition du gradient, qui peut arriver avec la fonction ReLU classique.

En effet, le choix de la fonction d'activation des couches va aussi jouer sur l'apprentissage. Comme on va le voir plus en détail dans la partie 2.3, l'apprentissage d'un

réseau de neurones se fait par descente de gradient, qui est rétro-propagé à chaque étape de l'apprentissage. Le choix de la fonction d'activation, comme d'autres choix dans la création du réseau, tels que l'initialisation des paramètres, le choix des hyper-paramètres, les stratégies de régularisation, faciliteront ou au contraire compliqueront l'apprentissage d'un problème qui n'est bien souvent pas convexe.

2.2.2.2. La couche de sortie

La fonction d'activation de la couche de sortie se détermine selon le problème que l'on souhaite résoudre et donc selon le type de prédiction à faire. Le modèle a transformé les données \mathbf{x} dans les couches précédentes, la dernière couche du réseau doit faire la transformation finale à partir de la représentation apprise \mathbf{h} pour produire \hat{y} .

La forme de la couche de sortie dépend donc de la fonction que l'on souhaite apprendre et de la forme des données que l'on souhaite prédire. Si les y sont continus, la fonction d'activation de la couche finale produisant \hat{y} peut-être une fonction linéaire. En revanche, quand la fonction que l'on souhaite apprendre doit remplir une tâche de classification, comme c'est le cas pour de la détection de vocalisation, on choisit une fonction d'activation qui permet de mimer une distribution catégorielle.

Dans le cas où $K = 1$, comme c'est le cas par exemple quand on veut prédire si \mathbf{x} contient une vocalisation ou non, y suit une distribution de Bernoulli. Le modèle doit donc prédire $\hat{y} = \mathbb{P}(y = 1|\mathbf{x})$, avec $0 \leq \hat{y} \leq 1$. La fonction d'activation de la dernière couche doit donc permettre de sortir un seul nombre compris entre 0 et 1. La fonction sigmoïde

$$\sigma(\mathbf{h}) = \frac{1}{1 + \exp(-\mathbf{h})}$$

permet de s'assurer que la sortie est bien comprise entre 0 et 1, et donc d'interpréter \hat{y} comme la réalisation d'une Bernoulli. La dernière couche prend la représentation \mathbf{h} que les couches cachées du réseau ont apprises, applique la transformation linéaire de la couche puis sa fonction d'activation, permettant de produire

$$\hat{y} = \sigma(\mathbf{w}^T \mathbf{h} + b).$$

Dans le cas où $y = f^*$ est un problème de classification avec plusieurs classes, le modèle $f(\mathbf{x}; \boldsymbol{\theta})$ doit produire une variable \hat{y} discrète de K valeurs, K correspondant au nombre de classe du problème de classification, avec $\hat{y}_k = \mathbb{P}(y = k|\mathbf{x})$. Pour que l'on puisse bien interpréter la sortie du modèle comme une distribution de probabilité, il convient non seulement que $0 \leq \hat{y}_k \leq 1, \forall k$, mais aussi que la somme des \hat{y}_k fasse 1. Utiliser la fonction *softmax* comme fonction d'activation de la dernière couche du modèle permet de traiter ces problèmes de classification, en représentant facilement la sortie du modèle comme une distribution de probabilité sur K classes. La dernière couche applique la transformation linéaire classique, prédisant une log probabilité non-normalisée. Si n est la dernière couche du modèle, $z_k^n = \log \tilde{\mathbb{P}}(y = k|\mathbf{x})$.

On applique ensuite la fonction d'activation *softmax*,

$$\text{softmax}(z_k) = \frac{\exp(z_k)}{\sum_j \exp(z_j)}.$$

Ce qui nous permet bien d'interpréter la sortie \hat{y} du modèle comme $\mathbb{P}(\mathbf{y}|\mathbf{x})$.

2.2.3. Apprentissage de représentation

Toute la tâche de l'apprentissage d'un réseau de neurones est d'apprendre automatiquement une représentation \mathbf{h} de \mathbf{x} permettant de résoudre $f^* = y$. On crée pour cela une fonction $f(\mathbf{x}; \boldsymbol{\theta})$, une composée de plusieurs transformations non-linéaires des données, permettant d'approximer f^* . Cette succession de transformation au fil des couches du réseau permet d'apprendre un espace latent de \mathbf{x} . Idéalement, dans cet espace, notre problème devient linéairement séparable et on peut donc facilement prédire $y = f(\mathbf{x}; \boldsymbol{\theta})$. En cela, l'apprentissage d'un réseau de neurone passe par l'apprentissage de représentation (Y. BENGIO, COURVILLE et VINCENT 2013).

Pour cette question d'apprentissage de représentation, un certain type de couche a été particulièrement important et à l'origine de la vague des réseaux de neurones, les réseaux convolutifs.

2.2.3.1. Réseau convolutif

Un réseau de neurones est capable d'avoir une grande flexibilité pour apprendre $f(\mathbf{x}; \boldsymbol{\theta})$, une fonction qui approxime f^* . En particulier, au fil des couches de f , les données \mathbf{x} sont projetés dans un espace latent différent de leur espace ambiant. Alors que dans l'espace initial des données, prédire y est complexe, cette prédiction est beaucoup plus simple avec la représentation \mathbf{h} que le réseau a appris automatiquement par les transformations successives. La succession des couches est équivalente à un travail d'extraction de caractéristiques de \mathbf{x} , mais fait directement au sein de la fonction, lors de l'apprentissage. Dans cet apprentissage de représentation, un type de couche est particulièrement important, les couches convolutives.

Introduit par Y. LECUN, BOSER, DENKER et al. 1989, le réseau de neurone convolutif permet un apprentissage de représentation hiérarchique tout en facilitant grandement l'apprentissage d'un réseau de neurone par descente de gradient (LECUN et Y. BENGIO 1995; Yann LECUN, Y. BENGIO et G. HINTON 2015). Au lieu que les couches fassent une opération matricielle sur l'ensemble de leurs paramètres, elles font une opération de convolution.

Dans une couche linéaire classique, l'opération matricielle se fait entre tous les points de l'entrée \mathbf{x} et l'ensemble des paramètres de la couche, de la dimension de sa largeur. Par l'opération de convolution, on parcourt les données avec une fenêtre glissante, le noyau de convolution. L'opération de convolution permet de calculer une moyenne pondérée entre les données \mathbf{x} et une fonction de pondération w . Le noyau de convolution est généralement de plus petite dimension que les données d'entrée, permettant de faire une opération matricielle sur un patch des données. Le nombre

2. Réseau de neurones – 2.2. Construction d'un réseau de neurones

de paramètres est ainsi réduit. Le noyau, aussi appelé filtre, avance ensuite dans les données. On fait un calcul matriciel utilisant la même matrice de pondération, le même noyau, pour chaque patch local des données.

L'introduction de l'opération de convolution dans les réseaux de neurones est passé par la question de la reconnaissance d'image (Y. LECUN, BOSER, DENKER et al. 1989). Les données sont alors de dimension 2, $\mathbf{X} \in \mathbb{R}^{H \times W}$. On définit le noyau de convolution aussi en dimension 2 $\mathbf{W} \in \mathbb{R}^{h \times w}$. L'opération de convolution entre les deux produits $\mathbf{Z} = \mathbf{X} * \mathbf{W}$, où

$$Z_{i,j} = \sum_{u=0}^{h-1} \sum_{v=0}^{w-1} x_{i+u, j+v} w_{u,v}. \quad (2.2)$$

Formellement, cette opération n'est pas une opération de convolution comme on la retrouve définie dans d'autres branches des mathématiques. C'est l'opération de corrélation croisée. Néanmoins, c'est le terme convolution qui s'est imposé dans la littérature *deep* et c'est de cette opération dont on parle quand on parle de couche convolutive.

L'opération de convolution peut se faire pour d'autres dimensions, supérieure comme inférieure. On retrouve ainsi une opération de convolution de dimension 1, par exemple pour traiter du son brut, comme le fait le réseau WaveNet (OORD, DIELEMAN, ZEN et al. 2016; ABDOLI, CARDINAL et LAMEIRAS KOERICH 2019). L'opération est la même, mais le noyau fait sa convolution dans une dimension un au lieu de deux. À noter, quand \mathbf{x} est sonore, on lui applique souvent une transformation, on calcule son spectrogramme. Cette représentation de \mathbf{x} est alors en dimension deux et les noyaux de convolution qu'on lui applique sont également en dimension 2.

Concrètement, l'opération de convolution consiste à comparer un patch local des données, de taille $h \times w$ et centré en (i, j) , au filtre \mathbf{w} . La sortie de l'opération mesure à quel point le patch est similaire au filtre. Le même filtre passe sur l'intégralité des données, permettant ainsi de mesurer à chaque localisation de \mathbf{x} la similitude avec \mathbf{w} . On peut multiplier le nombre de noyaux, fixer D noyaux de convolution. On ajoute juste une dimension à l'opération. Chaque noyau apprend à repérer un élément dans les données, que l'on va chercher à repérer à chaque localisation des données. Concaténer des couches de convolution permet d'apprendre une représentation hiérarchique, les noyaux des premières couches apprenant à repérer des caractéristiques de bas niveaux, les couches plus élevées des caractéristiques de plus haut niveau.

Par rapport à une couche connectée classique, l'opération matricielle se fait entre une partie des données et le noyau de convolution. Les mêmes paramètres sont utilisés pour les différentes parties. Cela a plusieurs avantages, principalement trois :

- des paramètres clairsemés. Le noyau est généralement plus petit que les données d'entrée, ce qui permet d'enregistrer moins de paramètres en mémoire;
- un partage des paramètres. Puisqu'on utilise et fait passer le même noyau à toutes les localisations de \mathbf{x} , on apprend seulement un ensemble de paramètres, ceux du noyau, qui vont voir tous les points de \mathbf{x} , au lieu d'un paramètre par point;
- une équivariance par translation. Elle est la conséquence des paramètres clairse-

2. Réseau de neurones – 2.2. Construction d'un réseau de neurones

més et de leur partage. Comme la fonction apprise par le noyau est équivariante, si son entrée change, sa sortie change de la même façon. Chaque noyau apprend ainsi une représentation, la même représentation qui est calculée sur chaque localisation des données.

L'opération de convolution permet de gagner en efficacité de calcul, de réduire le besoin de mémoire, et donc de multiplier les noyaux de convolution. Chaque noyau apprend une représentation particulière, qui cherchera la présence d'un élément discriminant de y . Chaque couche apprend une représentation des données et la concaténation des couches permet d'avoir une représentation hiérarchique, du plus bas niveau au plus haut niveau (Yann LECUN, Y. BENGIO et G. HINTON 2015).

Après l'opération de convolution, on retrouve l'application d'une fonction d'activation, comme pour les couches connectées classiques. On ajoute souvent une opération particulière après cela, une opération de *pooling*. Celle-ci consiste à prendre une statistique résumant un ensemble de valeurs voisines dans les représentations apprises par la couche convolutive, que ce soit par la valeur maximale, la valeur moyenne, ou autre. Cela permet de réduire encore la dimension de la représentation apprise et d'aider à rendre celle-ci invariante par translation. La réduction de la dimension permet aussi de réduire la quantité de paramètres à conserver en mémoire

Les réseaux convolutifs sont devenus depuis ImageNet (KRIZHEVSKY, SUTSKEVER et Geoffrey E HINTON 2012) l'état de l'art en reconnaissance d'image. Ils se sont imposés dans de plus en plus de champ d'application, avec des adaptations ou légères modifications, permettant de proposer de nouvelles architectures et d'améliorer encore les résultats et les temps de calcul, *e.g.* OORD, DIELEMAN, ZEN et al. 2016 en génération de parole, avec la convolution causale dans Wavenet, SZEGEDY, WEI LIU, YANGQING JIA et al. 2015 qui introduisent la convolution 1×1 pour réduire le coût computationnel et ajouter une non-linéarité supplémentaire ou HOWARD, M. ZHU, B. CHEN et al. 2017 la convolution séparable pour réduire le nombre de paramètres à apprendre. Si l'architecture des modèles et la façon d'intégrer l'opération change, les bases pour la faire restent les mêmes.

Leur capacité à apprendre une bonne représentation des données, couplé à leur bonne propriété pour l'apprentissage et aux innovations permanentes facilitant leur apprentissage, fait que l'on retrouve des couches convolutives dans la plupart des réseaux de neurones qui sont l'état de l'art actuel.

2.2.3.2. Transfert d'apprentissage

Ainsi, l'un des intérêts du *deep-learning* est sa capacité à apprendre une représentation hiérarchique des données, de projeter les données brutes dans un espace latent adapté à la tâche, et cela automatiquement, pendant le processus d'apprentissage. On évite ainsi l'extraction de caractéristiques, qui nécessite des compétences d'ingénierie spécifiques au domaine. Cependant, la quantité de données étiquetées nécessaires pour obtenir de bons résultats doit être importante afin de réussir à apprendre une représentation permettant de généraliser au problème que l'on pose. On a noté dans la partie 2.2.1 l'importance de l'initialisation des paramètres du modèle dans la conver-

gence et la capacité de généralisation du modèle. En outre, l'apprentissage de zéro d'un modèle profond nécessite des capacités de calcul qui dépassent souvent les contraintes budgétaires fixées. Dans cette situation, le transfert d'apprentissage est une bonne solution. Au lieu d'initialiser aléatoirement les paramètres et de faire un apprentissage de zéro, on utilise des paramètres pré-entraînés.

Grâce à l'apprentissage par transfert, les connaissances acquises sur un domaine source \mathcal{D}_S , pour une tâche spécifique \mathcal{T}_S , peuvent être utilisées pour une tâche cible \mathcal{T}_T ou un domaine \mathcal{D}_T (PAN et Q. YANG 2010). Le modèle source peut être appris à partir de données provenant d'une distribution différente (TAN, SUN, KONG et al. 2018), mais le transfert fonctionnera d'autant mieux que les données sources seront liées aux données cibles.

Pour classifier des vocalisations d'une espèce par exemple, pour laquelle on a relativement peu d'exemples étiquetés, on peut regarder si un modèle n'a pas déjà été entraîné sur un ensemble de données d'enregistrements audios beaucoup plus important. YamNet est par exemple un modèle librement et facilement accessible¹. Basé sur l'architecture de MobileNet (HOWARD, M. ZHU, B. CHEN et al. 2017), il a été entraîné sur plus de 2 millions d'enregistrements audio de 10 secondes issus du corpus AudioSet (GEMMEKE, ELLIS, FREEDMAN et al. 2017) pour détecter 521 classes d'événements, parmi lesquelles on trouve des sons humains, animaux et environnementaux. La distribution qui a généré les données d'entraînement de YamNet est certainement liée à la distribution qui nous intéresse, et sa distribution empirique est suffisamment riche pour caractériser le paysage sonore des écosystèmes que nous traitons. Par conséquent, l'espace latent de YamNet, la source, et les données que nous voulons traiter, la cible, sont susceptibles d'être les mêmes. De plus, l'espace des étiquettes \mathcal{Y} de la source (c'est-à-dire les 521 étiquettes) et de la cible sont également probablement liés, car nous essayons essentiellement de détecter des événements acoustiques dans un flux de sons. Le cardinal de \mathcal{Y}_S étant important, on peut supposer que \mathcal{Y}_T en est un sous-ensemble.

Ainsi, au lieu d'apprendre un modèle de zéro avec une forte probabilité de ne pas réussir à apprendre une représentation permettant de généraliser à des données non-vues par le modèle, on extrait les premières couches de YamNet et on les utilise comme les premières couches de notre réseau. Elles extraient l'information des fichiers audios jusqu'à l'espace latent à partir duquel il devient simple de prédire y .

De la même manière que des méthodes de pré-apprentissage permettent de mettre à profit des données non-étiquetées pour ne pas commencer l'apprentissage supervisé de zéro, et ainsi faciliter l'apprentissage de représentation, le transfert d'apprentissage facilite et accélère grandement la convergence du modèle. Les couches transférées peuvent être conservées gelées, si on considère que \mathcal{D}_S et \mathcal{D}_T sont assez liés et que la représentation ainsi apprise permet de répondre au problème \mathcal{T}_T . Si cette hypothèse est vraie, l'apprentissage est alors fortement facilité et raccourci. La figure 2.2 illustre la modèle que l'on présente dans notre première contribution au Chapitre 3, qui transfère ses premières couches du modèle YamNet, en les gardant fixes. On peut

1. <https://tfhub.dev/google/yamnet/1>

2. Réseau de neurones – 2.2. Construction d'un réseau de neurones

aussi procéder à du réglage fin, *fine tuning*, utiliser les paramètres transférés comme point de départ de l'apprentissage et laisser le gradient les faire évoluer pour coller davantage aux données.

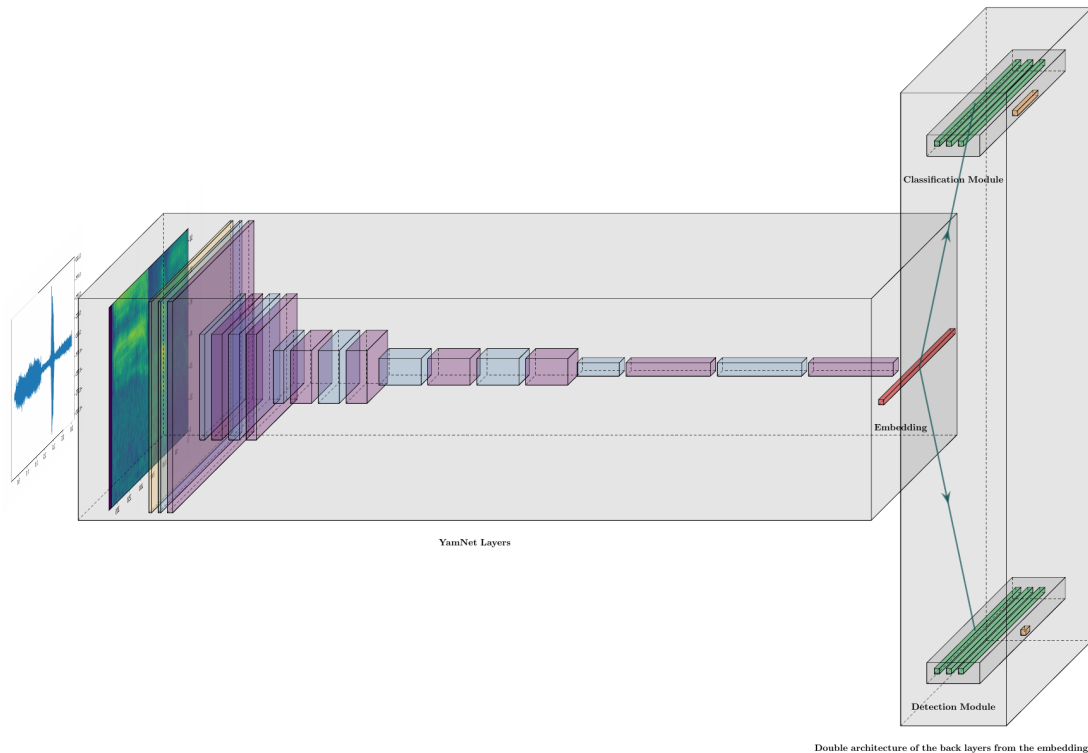


FIGURE 2.2. – L'architecture du réseau de notre première contribution (BONAFOS, PUDLO, FREYERMUTH et al. 2023). Nous transférons les couches avant du modèle de YamNet, qui calcule pour une fenêtre d'une seconde d'audio son spectrogramme de 96×64 , suivi d'une concaténation de couches de convolution *depth-wise* (en bleu) et *step-wise* (en violet), opérations de convolution introduites par HOWARD, M. ZHU, B. CHEN et al. 2017. Elles se terminent par une couche de *pooling* pour créer une représentation de dimension 1024. A partir de cette représentation, nous créons une double architecture : un module pour détecter s'il y a une vocalisation dans l'image ou non, un module pour classifier la vocalisation. Les couches avant sont totalement gelées pendant l'apprentissage, tandis que nous apprenons l'arrière de l'architecture sur les données. Une fois entraîné, le modèle peut traiter des données audio non étiquetées pour extraire des segments de vocalisation et prédire leur classe simultanément.

2.2.4. La fonction objectif

Il reste le dernier élément du réseau de neurones à déterminer, celui sur lequel va se concentrer tout l'apprentissage : la fonction objectif du modèle. C'est la fonction que l'on va minimiser en suivant le gradient de notre modèle.

En général, on prend l'entropie croisée entre la distribution des données et la distribution du modèle. Selon comment on a choisi de représenter la sortie du modèle \hat{y} , la forme de la fonction d'entropie croisée change. Notre modèle définit une distribution $p_{\text{modèle}}(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$. C'est la distribution des données selon notre modèle. On veut chercher à rendre cette distribution aussi proche que possible de la vraie distribution des données, que l'on ne connaît pas, mais que l'on représente par l'ensemble de données étiquetées que l'on a, $\hat{p}_{\text{données}}(\mathbf{y}|\mathbf{x})$. Par le principe du maximum de vraisemblance, on prend l'entropie croisée entre ces deux distributions comme fonction objectif,

$$J(\boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{x},\mathbf{y}\sim\hat{p}_{\text{données}}}\log p_{\text{modèle}}(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}). \quad (2.3)$$

C'est la fonction de sortie qui détermine \hat{y} et donc $p_{\text{modèle}}$. Dans le cas de la détection de vocalisation, $p_{\text{modèle}}$ est une distribution catégorielle. Si on réécrit l'équation (2.3) en discret, on a donc

$$J(\boldsymbol{\theta}) = -\sum_{(\mathbf{x}_i,\mathbf{y}_i)} \hat{p}_{\text{données}}(\mathbf{y}_i|\mathbf{x}_i)\log p_{\text{modèle}}(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}).$$

Si on prend le cas $K = 1$, y suit une distribution de Bernoulli : il est nul s'il n'y a pas de vocalisations dans \mathbf{x} , égal à 1 sinon. Pour prédire \hat{y} , on choisit une fonction qui permet de mimer une sortie comprise entre 0 et 1, la fonction sigmoïde, comme on l'a vu dans la partie 2.2.2.2. Quand y est nul, la fonction s'annule. Quand $y = 1$, si $\hat{y} = 1$, la fonction s'annule aussi. En revanche, plus \hat{y} tend vers 0, plus le log tend vers l'infini ce qui fait augmenter la perte avec le signe négatif devant. La logique est la même pour $K > 2$, où y_k vaut 1 quand les données \mathbf{x} en question contiennent la vocalisation k , zéro sinon.

Comme c'était le cas pour l'initialisation des paramètres ou le choix des fonctions d'activation des couches, le choix de la fonction objectif permet de faciliter la descente de gradient. Le gradient de la fonction objectif doit prendre des valeurs assez importantes et prédictibles pour être un bon guide lors de l'apprentissage. Les fonctions qui saturent, *i.e.*, qui deviennent plates, ne sont pas adaptées, puisqu'elles induisent un gradient très petit. Cela arrive souvent avec les fonctions d'activation des couches cachées, qui sont des fonctions qui saturent, provoquant donc une disparition du gradient, ce qui bloque et empêche l'apprentissage.

2.3. Apprentissage

Si le réseau de neurone fonctionne bien et est très rapide pour la prédiction en *feedforward*, son apprentissage est plus long et peut-être compliqué. Théoriquement, l'apprentissage est assez simple et s'appuie sur une descente de gradient.

2.3.1. La descente de gradient

2.3.1.1. Le gradient

Quand un modèle prend en entrée \mathbf{x} , les données se propagent le long du réseau pour produire une sortie \hat{y} . Durant l'entraînement, on utilise cette sortie pour calculer la fonction de perte $J(\boldsymbol{\theta})$. On est en train d'apprendre le modèle : on s'attend à ce qu'il y ait une différence importante entre \hat{y} et y . $p_{\text{modèle}}$ est encore loin de $p_{\text{données}}$. On va modifier les paramètres $\boldsymbol{\theta}$ de notre réseau de neurones f afin de faire de meilleures prédictions au second passage des données. L'information qui nous permet de modifier la valeur des paramètres se trouve dans le gradient de la fonction de perte par rapport aux paramètres $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$.

Les différents algorithmes d'optimisation de réseaux de neurones se basent sur le gradient. Comme on le présentait dans la partie 1.2.2, l'apprentissage consiste à optimiser un critère pour en optimiser un autre non-accessible. On minimise la fonction de perte $J(\boldsymbol{\theta})$ que l'on a défini. Son gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ contient toutes les dérivées partielles de J par rapport à $\boldsymbol{\theta}$. La dérivée de J au point $\boldsymbol{\theta}$ donnant la pente de J à ce point, elle indique comment J change pour un changement de $\boldsymbol{\theta}$. L'idée de la descente de gradient est donc d'utiliser les signes du gradient pour aller dans la direction opposée. En changeant $\boldsymbol{\theta}$ légèrement dans le sens opposé du signe de sa dérivée, on réduit J .

Pour minimiser J , on cherche la direction dans laquelle J décroît le plus vite. On définit les nouveaux points

$$\boldsymbol{\theta}' = \boldsymbol{\theta} - \epsilon \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

où ϵ est le taux d'apprentissage, un scalaire positif qui détermine la taille du pas à faire. La descente a convergé quand tous les éléments du gradient sont nuls (ou proche de zéro).

On se base donc sur le gradient pour minimiser $J(\boldsymbol{\theta})$. En pratique néanmoins, ce n'est pas à proprement parler le gradient de la fonction de perte que l'on calcule pour mettre à jour les paramètres du réseau de neurones. On calcule une estimation du gradient. La descente de gradient que l'on fait est stochastique.

2.3.1.2. Le gradient stochastique

Lors de l'apprentissage, on ne calcule pas l'espérance sur l'ensemble des données, mais sur un sous-ensemble de celui-ci, tiré au hasard, pour lequel on calcule ensuite la moyenne. On a un retour moins que linéaire à utiliser plus d'exemples pour calculer le gradient sur l'erreur standard de l'estimation de sa moyenne. Calculer un gradient sur 10 000 exemples prendrait 100 fois plus de temps que calculer un gradient sur 100 exemples, mais cela réduirait l'erreur standard de la moyenne d'un facteur 10. On a en pratique une convergence plus rapide en calculant rapidement des estimations approchées du gradient plutôt qu'en calculant longtemps le gradient exact.

Le calcul du gradient est donc une estimation statistique du gradient, pas son calcul précis. Le calcul du gradient ne se fait pas de manière déterministe. On passe par une méthode stochastique. La descente de gradient stochastique est une méthode

online : on tire les exemples d'un flux plutôt que de calculer directement l'intégralité du stock, le flux en question étant ici la distribution empirique. On tire des lots dans nos données étiquetées pour lesquelles on calcule le gradient moyen.

On obtient une estimation du gradient en prenant le gradient moyen d'un lot de m exemples tirés de la distribution des données. L'algorithme de descente de gradient stochastique, sous sa forme la plus classique, est présentée dans l'algorithme 1.

Algorithm 1 Mise à jour des paramètres θ à l'itération k par descente de gradient stochastique

Require: Taux d'apprentissage ϵ_k .

Require: Paramètres initiaux θ .

while Le critère d'arrêt n'est pas rempli **do**

 Tirer un lot de m exemples de l'ensemble d'entraînement.

 Estimer le gradient : $\hat{\mathbf{g}} \leftarrow \frac{1}{m} \nabla_{\theta} \sum_i L(f(\mathbf{x}^{(i)}; \theta), \mathbf{y}^{(i)})$.

 Faire la mise à jour : $\theta \leftarrow \theta - \epsilon \hat{\mathbf{g}}$.

end while

Le taux d'apprentissage est un paramètre primordial de l'algorithme d'optimisation par descente de gradient. Un taux d'apprentissage trop important provoquerait des oscillations très fortes de la courbe d'apprentissage. Si celui-ci est trop bas, l'apprentissage sera long. Un taux trop haut ou trop bas présente aussi le risque de sortir d'un minimum dans un cas, de rester bloqué à un niveau trop haut dans l'autre. Le choix de ce paramètre important peut se révéler assez complexe. Dans notre contribution du Chapitre 3, on l'apprend comme les autres hyper-paramètres par optimisation bayésienne, que l'on présente dans la partie 2.5.2.

Le taux d'apprentissage n'est pas nécessairement constant tout du long de l'apprentissage. Il peut en effet varier, et c'est d'ailleurs une pratique assez courante de le réduire graduellement au fil de l'apprentissage. On considère que l'on se rapproche d'un point intéressant au fil de l'apprentissage, que le modèle converge, et qu'il n'est plus nécessaire de faire des bonds aussi importants à chaque étape. On décroît alors l'importance des pas faits à chaque itération.

Cela est d'autant plus vrai que la descente de gradient stochastique introduit du bruit. Le gradient calculé sur un échantillon est un estimateur du gradient, et contient une erreur. Cette erreur est toujours présente, même quand on s'approche du point de convergence d'un minimum (alors qu'un vrai gradient serait nul à l'approche d'un minimum local). Ainsi, comme c'est le cas dans le Chapitre 3, le taux d'apprentissage décroît au fil de l'apprentissage.

À noter, on peut avoir une estimation du gradient moyen non biaisé si les lots sont sélectionnés de manière indépendante. La descente de gradient stochastique suit l'erreur de généralisation tant que les exemples ne sont pas répétés et indépendants. En général, les données sont mélangées puis passées plusieurs fois pour l'apprentissage. Lors de la première passe, chaque lot permet de calculer une estimation non biaisée de l'erreur de généralisation. À la seconde passe, cette estimation est biaisée parce qu'on ré-échantillonne des valeurs déjà utilisées au lieu de prendre de nouveaux

échantillons de la distribution des données. Quand les jeux de données à disposition deviennent énormes, le sur-apprentissage n'est plus un souci. Le facteur limitant est la capacité de calcul. Il devient possible de faire du vrai *online*, de faire des passes incomplètes sur les données, d'utiliser chaque exemple seulement une fois, et donc d'avoir à chaque fois une estimation non biaisée. De plus, comme le temps de calcul du gradient n'augmente pas avec le nombre d'exemples, on peut faire de l'apprentissage avec des ensembles d'entraînement massifs et avoir une convergence avant la fin de l'ensemble d'entraînement. Ces considérations sont d'ordre général et ne concernent pas le cas de cette thèse, où les données étiquetées à disposition ne permettent pas de se poser ce genre de question.

2.3.2. La rétro-propagation

La technique de descente de gradient n'est pas nouvelle. Elle date de Cauchy, en 1847. Afin de pouvoir être guidé par le gradient, il convient d'abord de le calculer. L'algorithme de rétro-propagation du gradient permet de calculer rapidement le gradient et rend donc son utilisation intéressante pour apprendre un réseau de neurones. L'utilisation de la rétro-propagation pour entraîner un réseau de neurones a été découverte de manière indépendante dans différentes équipes (LECUN 1985; RUMELHART, Geoffrey E. HINTON et R. J. WILLIAMS 1986; WERBOS 1988).

L'algorithme s'appuie sur le théorème de dérivation des fonctions composées. Celui-ci permet de calculer la dérivée de fonctions composées. Un réseau de neurones étant une fonction composée, on utilise ce théorème pour déterminer les règles de dérivation du réseau. L'algorithme de rétro-propagation est un algorithme qui calcule la dérivée dans un ordre particulièrement efficace.

En utilisant le théorème de dérivation des fonctions composées, on peut facilement écrire l'expression analytique pour le gradient d'un scalaire par rapport à chaque nœud ayant produit le scalaire. Ainsi, pour une sortie $J(\theta)$, on peut calculer le gradient de J par rapport à chaque nœud du réseau de neurone, *i.e.*, chaque θ , et donc le taux de variation de J selon le taux de variation de chaque $\theta_{\ell,i}$. Néanmoins, calculer cette expression demande de prendre en considération certains points. Une application naïve du théorème de dérivation des fonctions composées la rendrait impraticable, avec une demande de mémoire ou de calcul trop importante. L'algorithme de rétro-propagation propose une façon efficace de le faire. Il est fait pour éviter une explosion de sous-expressions identiques, chose qui arrive dans le cas d'un réseau de neurones. Il garde en mémoire certaines sous-expressions qui sont réutilisées afin de calculer le gradient par rapport à chaque $\theta_{\ell,i}$ plus rapidement.

L'algorithme de rétro-propagation peut être vu comme un algorithme qui remplit une table. Il enregistre des résultats intermédiaires qui seront réutilisés. Chaque nœud a un emplacement dans la table pour enregistrer le gradient de ce nœud. En enregistrant chaque nœud dans l'ordre, la rétro-propagation évite d'avoir à recalculer plusieurs fois la même expression. L'algorithme de rétro-propagation est un exemple d'algorithme avec une stratégie de programmation dynamique. Cela permet de rendre le calcul de la dérivée d'un réseau de neurone possible d'un point de vue calculatoire.

Ce n'est pas la seule façon, ni même la façon optimale selon comment l'algorithme est implémenté (*e.g.*, algorithmes qui utilisent des règles algébriques de substitution, ou préférer recalculer certaines expressions plutôt que les conserver en mémoire, justement pour en économiser), mais c'est une méthode pratique et efficace (BAYDIN, PEARLMUTTER, RADUL et al. 2017; MARGOSSIAN 2019).

L'algorithme de rétro-propagation est la première pierre sur laquelle se construit l'apprentissage d'un réseau de neurone. Il permet seulement de calculer le gradient du réseau de neurones. Une fois ce gradient calculé, c'est l'algorithme de descente de gradient stochastique qui permet de faire l'apprentissage à proprement parler, en utilisant le gradient.

2.3.3. L'algorithme de descente de gradient

On peut choisir d'apprendre notre modèle en suivant l'algorithme de descente de gradient classique présenté à l'algorithme 1, mais de nombreux algorithmes, basés à chaque fois sur la descente de gradient stochastique classique, ont été développés afin de faciliter l'apprentissage, l'accélérer, le rendre plus robuste au bruit, le rendre adaptatif, utiliser des méthodes de second ordre (DUCHI, HAZAN et SINGER 2011; ZEILER 2012; Y. A. LECUN, BOTTOU, ORR et al. 2012; DOZAT 2016; KINGMA et BA 2017).

On en présente ici un, l'algorithme NAdam, qui consiste en l'algorithme Adam (KINGMA et BA 2017) amélioré par l'ajout d'un momentum de Nesterov (DOZAT 2016). On choisit de présenter celui-ci car il permet d'avoir une vision représentative des extensions qui existent par ailleurs (taux d'apprentissage adaptatifs, ajout d'un momentum) et est le choix que l'on fait dans le Chapitre 3 pour l'apprentissage de nos modèles. Il n'y a pas à notre connaissance de preuves d'une supériorité théorique d'un algorithme sur les autres, mais des résultats empiriques semblent donner l'avantage à notre choix (DOGO, AFOLABI, NWULU et al. 2018).

L'algorithme Adam est un algorithme adaptatif, *i.e.*, le taux d'apprentissage s'adapte à chaque paramètre pour prendre en compte le fait que la perte soit très sensible à certaines dimensions de l'espace des paramètres et pas à d'autres. L'idée d'un algorithme adaptatif est alors d'ajuster le taux d'apprentissage à la dimension, et donc d'avoir autant de taux d'apprentissage que de paramètres. Adam ajoute à cette propriété adaptative un momentum. Physiquement, le momentum peut-être vu comme une impulsion, une lancée. L'idée du momentum est d'accumuler une somme décroissante des mises à jour précédentes dans un vecteur de momentum \mathbf{m} , pondéré par un facteur de décroissance μ . On remplace ensuite le gradient par ce vecteur lors de la mise à jour. Dans l'algorithme 1, cela revient à ajouter

$$\mathbf{m} \leftarrow \mu \mathbf{m} + \epsilon \hat{\mathbf{g}}$$

et à changer la mise à jour par

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \mathbf{m}.$$

Là encore, l'ajout d'un momentum induit une différence selon les dimensions de l'espace des paramètres. Il permet d'aller plus vite dans les dimensions où les mises à

jour vont toujours dans la même direction, et plus doucement dans les dimensions où la direction de la mise à jour du gradient change souvent d'une étape à une autre.

L'algorithme Adam utilise donc des taux d'apprentissage adaptatifs et ajoute un momentum. Alors que le momentum est généralement déterminé comme une somme décroissante des mises à jour passées, l'algorithme Adam le définit plutôt comme une moyenne décroissante des gradients passés, \mathbf{m} . De plus, la variance du gradient est également estimée, \mathbf{v} . Celle-ci permet de corriger les taux d'apprentissage adaptatif. On ajoute donc l'estimation des moments d'ordre 1 et 2 du gradient. Cela explique le nom de l'algorithme Adam : adaptative moment estimation. Pour garder l'idée du momentum, et donc une somme décroissante, deux paramètres β_1 et β_2 contrôlent le taux de décroissance de respectivement le moment d'ordre 1 et le moment d'ordre 2. Ces valeurs sont initialisées à zéro et à chaque étape, on les calcule en ajoutant la valeur passée pondérée par les paramètres β_i . Cette initialisation à zéro induit un biais, qui est corrigé à chaque étape en divisant chaque estimation par $1 - \beta_i^t$, t correspondant à l'itération. La correction est donc de moins en moins importante. Si on note $\hat{\mathbf{m}}$ et $\hat{\mathbf{v}}$ les estimations corrigées du biais des moments d'ordre 1 et 2, avec momentum (*i.e.*, en prenant en compte les valeurs précédentes), on a une mise à jour des paramètres qui est

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \frac{\hat{\mathbf{m}}}{\sqrt{\hat{\mathbf{v}} + \iota}},$$

où ι est un paramètre de stabilité numérique.

L'algorithme NAdam est juste une extension de l'algorithme Adam. On améliore le calcul du momentum en utilisant la méthode de Nesterov. Concrètement, le momentum de Nesterov recalcule le gradient à partir du gradient calculé à l'itération du moment alors que le momentum classique utilise le gradient passé. Cela permet des mises à jour d'encore meilleures qualités. L'algorithme 2 reproduit l'algorithme NAdam.

Algorithm 2 Algorithme NAdam

Require: Taux d'apprentissage ϵ .

Require: Taux de décroissance $\beta_1, \beta_2 \in [0, 1)$.

Require: Paramètres initiaux $\boldsymbol{\theta}$.

$\mathbf{m} \leftarrow \mathbf{0}$

▷ Initialisation du vecteur de premier moment

$\mathbf{v} \leftarrow \mathbf{0}$

▷ Initialisation du vecteur de deuxième moment

while Le critère d'arrêt n'est pas rempli **do**

$\hat{\mathbf{g}} \leftarrow \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$

$\mathbf{m} \leftarrow \beta_1 \mathbf{m} + (1 - \beta_1) \hat{\mathbf{g}}$

$\mathbf{v} \leftarrow \beta_2 \mathbf{v} + (1 - \beta_2) \hat{\mathbf{g}} \odot \hat{\mathbf{g}}$

$\hat{\mathbf{m}} \leftarrow \mathbf{m} / (1 - \beta_1^t)$

$\hat{\mathbf{v}} \leftarrow \mathbf{v} / (1 - \beta_2^t)$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \frac{\epsilon}{\sqrt{\hat{\mathbf{v}} + \iota}} \odot (\beta_1 \hat{\mathbf{m}} + \frac{(1 - \beta_1) \hat{\mathbf{g}}}{1 - \beta_1^t})$

end while

2.4. La régularisation, le sur-apprentissage

La partie 2.2 a présenté ce qu'est un réseau de neurones. La partie 2.3 a présenté son apprentissage. Mais l'apprentissage d'un réseau de neurone ne se limite pas à son algorithme. Comme on a pu le noter à différents endroits, celui-ci peut-être assez complexe et différentes stratégies sont utiles et nécessaires pour que cet apprentissage soit un succès. L'architecture du réseau, la façon d'initialiser les paramètres, le recours à du transfert d'apprentissage sont autant de facteurs qui conditionnent la réussite de l'apprentissage, pour l'ensemble d'entraînement à disposition. De plus, un réseau de neurones a la capacité d'apprendre presque n'importe quel ensemble de données (C. ZHANG, S. BENGIO, HARDT et al. 2017). Si le risque de sur-apprentissage est limité dans le contexte de données réellement massives, ce n'est pas le cas quand les quantités de données d'entraînement sont plus modestes. Mettre en place des stratégies de régularisation est alors nécessaire afin de se prémunir contre ce phénomène et apprendre bel et bien un réseau capable de généraliser sur des exemples non-connus. En pratique, le meilleur modèle, celui qui généralise le mieux, est bien souvent un modèle large qui a été bien régularisé pour la tâche en question. Ainsi, on peut avoir des modèles sur-paramétrés, avec la capacité de sur-apprendre les données, qui ont malgré tout de bonnes capacités de généralisation (ZHOU, VEITCH, AUSTERN et al. 2018).

2.4.1. Data augmentation

Une des clés du succès des réseaux de neurones étant les ensembles d'entraînement massifs à disposition, la première idée qui vient pour améliorer la capacité de généralisation d'un réseau de neurones est d'augmenter les données d'entraînement. L'idée est d'appliquer une transformation aux données originales qui les modifie légèrement, afin d'ajouter de la variance à celles-ci, pour rendre le modèle plus robuste. On déforme les données originales tout en gardant la sémantique des étiquettes, pour que l'apprentissage reste possible. Cela va jouer sur la variance et le biais des estimateurs (HUANG, ORBANZ et AUSTERN 2022).

Les méthodes d'augmentation des données, de création de nouveaux exemples à partir des existants, ont d'abord montré leur intérêt dans les questions de reconnaissance d'image. Dans ce cas, des translations de pixels, des rotations d'image aident le modèle à apprendre une représentation invariante des données utiles pour augmenter les capacités de généralisation.

Des méthodes d'augmentation de données adaptées à des données sonores existent et permettent d'augmenter artificiellement les ensembles de données (MCFEE, HUMPHREY et Juan P BELLO 2015), avec de bons résultats sur l'amélioration des capacités de généralisation (SALAMON et Juan Pablo BELLO 2017). Les méthodes consistent à modifier la tonalité de l'enregistrement initial, de l'accélérer ou le ralentir, ou encore d'ajouter du bruit de fond, qui peut être du bruit blanc généré aléatoirement ou de vrais sons venant du paysage sonore de l'enregistrement original.

2.4.2. Batch normalization

La normalisation des lots (IOFFE et SZEGEDY 2015) n'a pas pour but initial de régulariser le modèle. En effet, la stratégie est originellement prévue pour faciliter l'apprentissage. En centrant et réduisant les lots, et ce à l'entrée de chaque couche, *i.e.*, en incorporant l'opération à l'architecture même du réseau, l'apprentissage d'un réseau de neurones profond est facilité et sa durée réduite. L'opération permet d'éviter de propager un signal qui prendrait des valeurs trop hautes ou trop basses, ce qui a un effet négatif lors de la descente de gradient. Surtout, la normalisation est prise en compte lors de la rétro-propagation du gradient, afin d'éviter que le gradient n'induisse une augmentation de l'écart-type ou de la moyenne. L'incorporation de la normalisation au sein de l'architecture permet d'éviter d'avoir à incorporer des pénalités supplémentaires au sein de la fonction de coût. On a ainsi chaque couche de variance unitaire et de moyenne nulle.

En plus de stabiliser l'apprentissage en améliorant l'optimisation, la normalisation des lots à chaque couche permet aussi d'ajouter du bruit. En effet, l'estimation des statistiques pour faire la normalisation, la moyenne et la variance, sont bruitées car calculées sur un lot. Le bruit induit par l'opération a un effet de régularisation sur les données.

2.4.3. Dropout

Le *dropout* (SRIVASTAVA, G. HINTON, KRIZHEVSKY et al. 2014) consiste à annuler de manière aléatoire une partie des nœuds du modèle à chaque itération lors de l'apprentissage du modèle. L'application de cette stratégie ajoute un bruit très bénéfique à l'apprentissage. De plus, le *dropout* peut être perçu comme l'utilisation d'une méthode d'ensemble sur un réseau de neurones.

En effet, chaque itération de l'apprentissage avec *dropout* revient à échantillonner un réseau aminci et l'entraîner. Cela est refait à chaque itération, puisque ce sont de nouveaux nœuds qui sont annulés. Entraîner un réseau de neurones avec du *dropout* peut-être vu comme entraîner une collection de réseaux amincis, avec un partage des paramètres, mais sans un coût computationnel. Une fois l'apprentissage terminé, le modèle est utilisé normalement.

Si on reprend l'opération que fait la couche ℓ d'un réseau de neurones qui produit la sortie \mathbf{h}^ℓ

$$\mathbf{h}^\ell = f(\mathbf{W}^{\ell T} \mathbf{h}^{\ell-1} + \mathbf{b}^\ell)$$

où f est une fonction d'activation. La régularisation *dropout* consiste à introduire une variable $r \sim \text{Bernoulli}(p)$ et à modifier le processus de propagation

$$\begin{aligned} \tilde{\mathbf{h}} &= \mathbf{r} \odot \mathbf{h}, \\ \mathbf{h}^\ell &= f(\mathbf{W}^{\ell T} \tilde{\mathbf{h}}^{\ell-1} + \mathbf{b}^\ell). \end{aligned} \tag{2.4}$$

Pour chaque couche ℓ , \mathbf{r}^ℓ est un vecteur de variables aléatoires de Bernoulli indépendante, *i.e.*, un vecteur de la taille de la largeur de la couche ℓ d'éléments égaux à 0

ou 1, égaux à 1 avec la probabilité p . Ce vecteur est tiré et multiplié par élément à la sortie de la couche ℓ \mathbf{h}^ℓ , ce qui produit une sortie amincie $\tilde{\mathbf{h}}^\ell$ qui est utilisée comme entrée de la couche suivante.

La procédure est appliquée à chaque couche, ce qui revient à échantillonner un sous-réseau du réseau global. Lors de l'apprentissage, les dérivées de la fonction de perte sont rétro-propagées le long du sous-réseau. La technique permet d'éviter des co-adaptations entre les unités cachées. Toutes les unités doivent participer à la construction d'une représentation et à la prédiction même si certaines sont absentes. Cela évite la construction de dépendances entre les différentes unités du réseau, qui sont fragiles en pratique.

Dans le cas de l'utilisation de la régularisation *dropout*, le paramètre p devient un hyper-paramètre du modèle.

2.4.4. Contrainte sur la norme des paramètres

Les stratégies de régularisation peuvent fonctionner de concert. C'est par exemple le cas du *dropout* et de l'ajout d'une contrainte sur la norme des paramètres, qui semble améliorer encore les effets (SRIVASTAVA, G. HINTON, KRIZHEVSKY et al. 2014).

L'ajout d'une contrainte sur la norme des paramètres revient à contraindre la norme des paramètres d'une couche à rester inférieure à une constante c . Si \mathbf{w}^ℓ est le vecteur des poids d'une couche, le réseau de neurones est optimisé sous la contrainte

$$\|\mathbf{w}^\ell\|_2 \leq c.$$

On parle de régularisation max-norm car elle implique que la valeur maximum que la norme de tout poids peut prendre est c .

Cette pénalisation est largement utilisée et empêche les poids de prendre trop d'amplitudes. Si l'utilisation du *dropout* seul, comme l'ajout d'une contrainte sur la norme des paramètres, permet déjà d'améliorer les résultats, la conjugaison des deux semble particulièrement bénéfique.

2.4.5. Early-stopping

Le *early-stopping* arrête l'apprentissage alors que la fonction objectif est toujours en train de décroître. Les réseaux de neurones ont souvent une très grande capacité et peuvent facilement sur-apprendre, d'autant plus quand les ensembles de données ne sont pas énormes. L'erreur de validation, qui nous intéresse car mesurant l'erreur de généralisation, peut et va sans doute remonter pendant l'apprentissage alors que l'erreur d'apprentissage continue à baisser. C'est le sur-apprentissage que le *early-stopping* permet d'éviter.

En pratique, l'utilisation de cette méthode de régularisation est très simple. L'apprentissage est fixé pour un nombre d'itérations T . Si à l'itération E , l'erreur de validation ne décroît plus, on garde en mémoire les paramètres du modèle à l'étape E . Si l'erreur de validation continue de ne pas décroître pendant une période de patience de P

itérations, on stoppe en avance l'apprentissage et on renvoie comme modèle le réseau avec les paramètres de l'itération E . L'apprentissage aura été stoppé plus tôt, à l'itération $E + P < T$, alors même que la fonction objectif continuait à décroître. On récupère ainsi les paramètres du réseau ayant permis d'avoir la meilleure erreur de généralisation lors de cet apprentissage.

2.5. Apprentissage des hyper-paramètres

L'ensemble de la construction et de l'apprentissage d'un réseau de neurones a été vu. Il reste le dernier point à aborder avant de pouvoir lancer l'apprentissage d'un réseau de neurones, la détermination de ses hyper-paramètres.

2.5.1. Le choix des hyper-paramètres

Les paramètres du modèle θ sont les éléments que l'on apprend, qui permettent de construire automatiquement une représentation hiérarchique du signal et de prédire \hat{y} . D'autres paramètres restent à la discrétion de l'utilisateur, ce sont les hyper-paramètres du modèle. Contrairement à d'autres modèles qui n'ont qu'assez peu de paramètres de réglage, les réseaux de neurones ont un nombre important de paramètres à fixer a priori et leurs valeurs impactent significativement l'apprentissage. Un réseau de neurones demande à être méticuleusement réglé pour être performant.

Comme on le disait dans la partie 2.3, le taux d'apprentissage est par exemple un hyper-paramètre dont le rôle dans le processus d'apprentissage est primordial. Son choix relève souvent plus d'un art que d'une science et conditionne fortement le succès ou non de l'apprentissage. On a également vu dans la partie 2.3.3 que l'algorithme de descente de gradient peut introduire de nouveaux paramètres. C'est le cas avec l'algorithme NAdam, qui introduit les paramètres β_1 et β_2 (ainsi que ι , mais celui-ci étant là pour une question de stabilité numérique, on peut le laisser fixe). KINGMA et BA 2017 conseille d'ailleurs des valeurs pour ces paramètres de pondération ($\beta_1 = 0.9, \beta_2 = 0.999$).

L'architecture même du réseau peut-être vu comme un hyper-paramètre. Le nombre de couches, leur largeur, la taille des noyaux de convolution, ces éléments sont des choix que l'on fait a priori qui ont une conséquence sur la fonction f que l'on peut apprendre, son temps d'apprentissage et ses capacités de généralisation. Pareillement, les stratégies de régularisation présentées dans la partie 2.4 sont déterminées selon certaines valeurs, *e.g.*, la probabilité p de *drop-out*, la norme c des paramètres.

Si la valeur des hyper-paramètres n'est pas adaptée par l'algorithme d'apprentissage, on peut mettre en place des procédures d'apprentissage imbriquées où les hyper-paramètres deviennent les paramètres d'un autre algorithme d'apprentissage et sont appris. L'apprentissage des hyper-paramètres se fait alors sur l'ensemble de validation. Le réseau est appris sur l'ensemble d'entraînement, son erreur de généralisation est estimée sur l'ensemble de validation, pour une combinaison donnée d'hyper-paramètres. Une nouvelle combinaison d'hyper-paramètres induit l'apprentissage

d'un nouveau réseau, dont son erreur de généralisation sera à nouveau estimée sur l'ensemble de validation. La combinaison ayant permis de minimiser l'erreur de généralisation du modèle est l'ensemble des hyper-paramètres choisis, qui ont été ainsi appris sur l'ensemble de validation. La véritable capacité de généralisation du modèle est finalement testée sur l'ensemble de test.

Diverses stratégies existent afin d'apprendre ces valeurs. Le *grid search* consiste à spécifier un nombre de valeurs possible pour chaque hyper-paramètre et tester chaque combinaison. On construit une grille, où chaque élément est une combinaison d'hyper-paramètre, et on procède à l'apprentissage d'un modèle pour chaque élément de la grille. Cette stratégie est computationnellement très gourmande. Il est assez compliqué de procéder à un *grid search* quand les hyper-paramètres commencent à s'accumuler et peuvent prendre un nombre important de valeurs, comme c'est souvent le cas pour des réseaux de neurones. Le *random search* permet une convergence plus rapide en ajoutant une part d'aléatoire dans la sélection des hyper-paramètres testés. Néanmoins, on ne prend pas non plus en considération l'information contenue par les combinaisons passées. L'optimisation des hyper-paramètres basée sur un modèle d'apprentissage permet d'intégrer cette information et ainsi réduire le nombre d'apprentissages de modèle, ce qui est à chaque fois très coûteux dans le cas des réseaux de neurones.

2.5.2. L'optimisation bayésienne

L'optimisation bayésienne (BROCHU, CORA et DE FREITAS 2010; SNOEK, LAROCHELLE et Ryan P ADAMS 2012; SHAHRIARI, SWERSKY, Ziyu WANG et al. 2016) est une optimisation séquentielle des hyper-paramètres basée sur un modèle d'apprentissage. Cette stratégie est construite sur un modèle de régression bayésien qui estime l'espérance de l'erreur sur l'ensemble de validation pour chaque hyper-paramètre ainsi que l'incertitude liée à cette estimation. Le processus d'optimisation qui se construit sur ce modèle procède ensuite à un arbitrage entre exploitation et exploration, *i.e.*, rechercher et sélectionner des valeurs dans l'espace des paramètres pour lesquelles l'incertitude est importante, ce qui pourrait emmener à des améliorations importantes (ou le contraire), ou bien proposer des valeurs pour lesquelles le modèle est confiant sur la possibilité de réduire l'erreur (des valeurs généralement assez proches de valeurs déjà utilisées auparavant).

L'objectif est de trouver les arguments qui optimisent une fonction coûteuse à calculer. En l'occurrence, on veut minimiser l'erreur de validation d'un réseau de neurones, avec comme arguments les hyper-paramètres du modèle. La fonction est coûteuse à calculer car un calcul équivaut à l'apprentissage d'un réseau de neurones. La stratégie se divise en deux blocs. Premièrement, un modèle probabiliste, qui modélise la fonction que l'on optimise. L'a priori porte sur la forme que peut prendre cette fonction. Les données que l'on observe sont les résultats du réseau pour une combinaison possible d'hyper-paramètres. L'a posteriori permet de mettre à jour nos a priori étant donné les observations, *i.e.*, la forme de la fonction selon les résultats du réseau pour différentes combinaisons d'hyper-paramètres. Deuxièmement, une fonction d'ac-

quisition, qui permet de sélectionner au mieux la combinaison d'hyper-paramètres pour faire un apprentissage. Une optimisation bayésienne est un processus itératif : le modèle probabiliste se substitue à la vraie fonction. La fonction d'acquisition, en se basant sur l'a posteriori du modèle probabiliste, sélectionne le point de l'espace des hyper-paramètres du réseau de neurones qui est le plus intéressant dans un arbitrage exploitation-exploration. Le modèle est entraîné avec cette combinaison d'hyper-paramètres, permettant de mettre à jour l'a posteriori du modèle probabiliste, et donc de sélectionner une nouvelle combinaison en prenant en compte les itérations passées.

Concernant le premier bloc, le modèle probabiliste s'appuie sur un processus gaussien pour représenter la fonction g qui modélise l'erreur de généralisation selon les hyper-paramètres du réseau. Soit $\mathcal{D}_n = \{(\boldsymbol{\lambda}_i, y_i) : i = 1, \dots, n\}$. $\boldsymbol{\lambda}_i$ représente une combinaison d'hyper-paramètres. $y_i = g(\boldsymbol{\lambda}_i) + \epsilon_i$ représente l'erreur de généralisation du réseau pour la combinaison $\boldsymbol{\lambda}_i$ (i.e., on apprend le réseau avec les hyper-paramètres $\boldsymbol{\lambda}_i$ sur l'ensemble d'apprentissage et on calcule son erreur sur l'ensemble de validation une fois l'entraînement terminé). g étant une fonction, on le définit comme un processus stochastique gaussien (C. E. RASMUSSEN et C.K. I. WILLIAMS 2006)

$$p(g(\boldsymbol{\lambda})|\mathcal{D}_n) = \mathcal{N}(g|\mu_n(\boldsymbol{\lambda}), \sigma_n^2(\boldsymbol{\lambda})). \quad (2.5)$$

Avec les processus Gaussiens, on entre dans le monde des modèles non-paramétriques, que l'on voit plus en détail dans le Chapitre 6.

Concernant le deuxième bloc, une fonction d'acquisition $\alpha(\boldsymbol{\lambda}; \mathcal{D}_n)$ calcule l'utilité d'estimer g pour $\boldsymbol{\lambda}$, afin de déterminer la prochaine valeur de $g(\boldsymbol{\lambda})$ à calculer. On observe $y_{n+1} = g(\boldsymbol{\lambda}_{n+1}) + \epsilon_{n+1}$ pour mettre à jour nos croyances sur g , et on répète. Le rôle de α est d'arbitrer entre explorer davantage l'espace des paramètres pour potentiellement trouver une combinaison de paramètres permettant de réduire drastiquement l'erreur, ou exploiter les endroits de cet espace pour lesquels la probabilité d'amélioration est importante. Différentes fonctions d'acquisition sont possibles (KUSHNER 1964; JONES 2001; SRINIVAS, KRAUSE, KAKADE et al. 2010). La fonction d'acquisition UCB, pour *Upper Confidence Bound*, utilisée dans le Chapitre 3, est définie comme

$$\alpha_{\text{UCB}}(\boldsymbol{\lambda}; \mathcal{D}_n) = \mu_n(\boldsymbol{\lambda}) + \beta \sigma_n(\boldsymbol{\lambda}). \quad (2.6)$$

Le paramètre β étant le paramètre d'arbitrage entre l'exploration et l'exploitation.

La procédure permet ainsi d'apprendre les meilleurs hyper-paramètres adaptés à la tâche. Les "nouveaux" hyper-paramètres sont alors ceux de l'algorithme d'optimisation Bayésienne, le choix de la fonction moyenne et de la fonction de covariance du processus, ainsi que le choix du paramètre d'arbitrage d'exploration et d'exploitation de la fonction d'acquisition. Cela laisse donc beaucoup moins de paramètres à déterminer.

Ce Chapitre a revu les points importants de la définition d'un réseau de neurones, l'algorithme sur lequel repose son apprentissage ainsi que les différentes stratégies utilisées en pratique. Cela permet d'introduire le Chapitre 3 suivant, la première

2. Réseau de neurones – 2.5. Apprentissage des hyper-paramètres

contribution de cette thèse, qui utilise un réseau de neurones pour répondre au premier problème posé, à savoir la détection de vocalisations.

3. Detection and classification of vocal productions in large scale audio recordings

Sommaire

3.1. Introduction	64
3.2. Methodology	66
3.2.1. Data	66
3.2.2. Network architecture	67
3.2.3. Fit on the data	69
3.2.4. Vocalization delineation and classification	71
3.3. Experimental Validation	71
3.3.1. From audio recordings to data banks for our method	72
3.3.2. Performance of our deep learning architecture	73
3.3.3. New large-scale databases of vocalizations	74
3.4. Conclusion and Discussion	75
3.5. Supplementary materials	77

Ce chapitre est la première contribution de la thèse. Il a donné lieu à l’écriture d’un article, avec le co-autorat de Pierre Pudlo, Jean-Marc Freyermuth, Thierry Legou, Joël Fagot, Samuel Tronçon et Arnaud Rey. Celui-ci est actuellement soumis à la revue *Machine Learning*, une prépublication est librement accessible sur ArXiv <https://arxiv.org/abs/2302.07640>, ainsi que le code sur <https://gitlab.com/papers4375727/detection-and-classification-of-vocal-productions>. Ce travail a permis la production de deux bases de vocalisations, une de singe, rendue elle aussi accessible sur <https://zenodo.org/record/8239697>, une de bébé, qui reste confidentielle pour respecter des données privées.

Abstract

We propose an automatic data processing pipeline to extract vocal productions from large-scale natural audio recordings and classify these vocal productions. The pipeline is based on a deep neural network and addresses both issues simultaneously thanks to a dual architecture. The front layers are transferred from YamNet, enabling successful learning despite a reduced training set, that we enhance by data augmentation and re-sampling. The back of the architecture is organized into two modules, whose precise

number of layers is automatically determined on the training data through Bayesian optimization. Our end-to-end methodology can handle noisy recordings made under different recording conditions. We test it on two different natural audio data sets, one from a group of Guinea baboons recorded from a primate research center and one from human babies recorded at home. The pipeline trains a model on 72 and 77 minutes of labeled audio recordings, with an accuracy of 94.58% and 99.76%. It is then used to process 443 and 174 hours of natural continuous recordings and it creates two new databases of 38.8 and 35.2 hours, respectively. We discuss the strengths and limitations of this approach that can be applied to any massive audio recording.

Keywords

detection, classification, neural network, transfer learning, vocalization

3.1. Introduction

The manual process of continuous audio recordings to extract and label vocalizations is a complex, tedious and error-prone task. Databases obtained manually are the result of a large and time-consuming task. With an automatic method, we can quickly and cheaply build new massive databases. Many continuous-time audio recordings represent significant amounts of data, within which the events of interest are infrequent or even rare. Yet these continuous recordings have their own merit : recording an ecosystem without the presence of an experimenter, then automatically extracting vocalizations from it, makes it possible to create new, richer databases. Having a larger number of vocalizations with greater variability would provide domain experts with much more important and relevant information to refine or even challenge repertoire definitions. In this article, we propose a methodology entirely based on a deep neural network to address the dual challenge of (1) detecting vocalization periods and (2) performing supervised classification of these vocalizations. We need a general workflow, adaptable to find vocalizations of different species, produced in different conditions and different ecosystems. The workflow should be user-accessible, relatively fast and cheap to implement and run. It should require neither massive computational resources nor massive labeled data.

This dual challenge faces numerous issues. Firstly, the vocalization data are necessarily scarce due to manual processing to obtain them, or the limitations of publicly available databases. Secondly, the audio data are captured in uncontrolled environmental conditions, and we have to contend with a variety of background sounds. Thirdly, recording conditions may vary, including different microphone positions or orientations, the use of multiple microphones during recording, and the subject's position relative to the microphone. Fourthly, it is necessary to address the issue of the digital representation of such audio data. Finally, we aim to control the computational cost to maintain reasonable resource usage and facilitate the wider adoption of the proposed procedure. Traditional detection methods (1) filter the signal, (2) extract a feature vector and (3) use a classification model on this vector (DIETRICH, PALM, RIEDE et al. 2004; X. XIA, TOGNERI, SOHEL et al. 2018; T. T. T. NGUYEN, T. T. NGUYEN,

3. Detection and classification of vocal productions in large scale audio recordings – 3.1. Introduction

LIEW et al. 2018; STRISCIUGLIO, VENTO et PETKOV 2019). Neural network avoids these steps learning a hierarchical representation of data automatically (Yann LECUN, Y. BENGIO et G. HINTON 2015), in an end-to-end manner. It becomes state-of-the-art in bioacoustics and event detection problems (STOWELL, STYLIANOU, WOOD et al. 2018), detecting animal vocal productions when these are in low proportions in massive audio recordings (BERGLER, SCHRÖTER, CHENG et al. 2019), and despite the noise present in the recordings (OIKARINEN, SRINIVASAN, MEISNER et al. 2019). But the limitations remain significant for use on many detection problems (STOWELL 2022). These approaches require training databases (and the computational resources that go with them) which are often not available.

The objective of the methodology we propose is to detect variable-duration vocalization segments within a continuous recording and label them according to a pre-labeled training dataset. Our methodology is based on a single deep neural network that handles all the tasks we aim to achieve using raw PCM audio files and a pre-labeled training database. It is described in the second part of this article. We have taken great care in this method to avoid biases. To be able to delineate vocalization segments in the recordings, the machine learning method should be trained on a base that is as representative as possible of the soundscape, including, e.g. biological, geophysical, anthropogenic sounds (PIJANOWSKI, FARINA, GAGE et al. 2011; BERGLER, SCHRÖTER, CHENG et al. 2019). We detail the data required for training and its enrichment, the proposed network architecture, its optimization through data-driven adjustments, the cost function, as well as the transfer learning from a pre-trained network. Transfer learning has already proved its worth for sound event detection (CHOI, György FAZEKAS, M. SANDLER et al. 2017; HERSHEY, CHAUDHURI, ELLIS et al. 2017; PALANISAMY, SINGHANIA et YAO 2020; AYTAR, VONDRICK et TORRALBA 2016). Indeed, we have adopted the representation provided by the YamNet model (*TensorFlow Hub* 2022). This trained deep neural network is based on a MobileNet architecture (HOWARD, M. ZHU, B. CHEN et al. 2017) and has been trained on the massive AudioSet database (GEMMEKE, ELLIS, FREEDMAN et al. 2017). Starting from a mel spectrogram of dimension 96×64 , YamNet learns a 1024-dimensional representation of the audio signal, that is then used to solve a classification problem with 521 different classes. The massive AudioSet database, composed of more than 2 billions of 10-seconds audio records, has been used in various bioacoustic tasks (STOWELL 2022) and for different sound event detection tasks (TENA, CLARIÀ et SOLSONA 2022; PATIL et WANI 2023). Additionally, the MobilNet architecture that was implemented to set YamNet, with deep-wise and piece-wise convolution layers, has been designed to be resource-efficient. Relying on YamNet in the middle of our network will ease the resort to our method without placing a heavy burden on computational resources. Moreover, since we keep the weights of YamNet as they are, the fitting on the learning database is simplified.

The third part of this article provides a numerical evaluation of the proposed method using two distinct studies. The first study focuses on recordings of baboons in their habitat at a primate research center. The training database, whether for vocalization detection or classification, is derived from manual labeling of a few recordings

captured using the same setup. The second study involves recordings of human infants in their natural environment using recording devices provided to parents. Here, the training database relies on publicly available data concerning infants, which may not possess the same acoustic properties as they originate from other recording conditions. While not an exhaustive evaluation of the proposed method, we draw important conclusions regarding the method’s qualities at the end of this part and in the conclusion section.

3.2. Methodology

This section presents the method we propose for processing these sound recordings. The recordings are divided into short overlapping frames of 1 second each. The details of this segmentation and the training databases are provided in Section 3.2.1. Our methodology for processing these frames using a single neural network is described in Section 3.2.2. To keep the structure of our network relatively simple, we completely disregard the temporal correlation between successive frames in the continuous recordings as well as in the training databases. While other choices are possible, such as using recurrent networks, training them on medium to large databases can be computationally expensive. Furthermore, for sound production detection tasks, the usefulness of using recurrent layers is limited (STOWELL 2022). The adjustment of the network’s layers that need to be calibrated on data is described in Section 3.2.3. Our objective also requires a step of reconciling predictions on successive frames of the continuous recording, which is also provided in this Section. This reconciliation step allows us to obtain variable-duration vocalization segments and a single labeling per segment.

3.2.1. Data

To create a suitable training dataset for our dual problem, we need to build two banks of sound recordings. The first bank of audio recordings should consist of labeled vocalization recordings (short or long) without any silence. It aims to address the problem of labeling the detected vocalizations in continuous recordings. These recordings can be manually extracted from the continuous sound recordings and labeled by an expert. Alternatively, they can be sourced from publicly available databases. The second bank of audio recordings aims to delineate the vocalization segments from the soundscape in our continuous recordings. The soundscape is the acoustic expression of an ecosystem (PIJANOWSKI, FARINA, GAGE et al. 2011) or an environment. Hence we highly recommend to build the second bank of background sounds without vocalizations by hand-picking them from the continuous audio recordings. To be as representative as possible of the diversity of noise, they should be picked at various places throughout the whole continuous recordings. Since vocalizations are more seldom than background noises, they should be relatively easy to isolate by hand.

Moreover, once those two banks are compiled, we strongly encourage the resort to relevant data augmentation techniques to enrich these data. We used the ready-to-use library of MCFEE, HUMPHREY et Juan P BELLO 2015, which allows us to multiply by 15 the number of labeled recordings we have in the first bank. From each original file, we shift the pitch by 5 values linearly spaced within $(-4, 4)$, we stretch the speed by 5 values logarithmically spaced within $(0.81, 1.23)$, and we add 5 background noise from the second bank.

To standardize the audio recordings, we assume that they are all available as a single-channel audio signal, sampled at 16 kHz in PCM format, for example as WAV files. The pulse modulation signal is translated to be on a scale between -1 and $+1$. Through windowing, the signal is divided into frames of one second each, with an overlap of 80%. This divides a recording of T seconds into $5T$ frames. The position of each frame along the continuous recordings should be saved. The same process is also applied to the recordings of the two banks, yet saving their position is useless. Note that a time window of one second is consistent with the problem with which we are dealing. It allows us to quickly discard noise segments from the data to be analyzed. One second seems to be a good compromise, sufficient to encompass most vocalizations but not too large to be easily processed.

Whether it is the frames obtained from the decomposition of the continuous recording to be analyzed or from the two banks, we use the following conventions in the notations : x represents a one-second audio signal, $y \in \{0, 1\}$ is a vocalization indicator, and $z \in \{1, 2, \dots, K\}$ is the label of the vocalization. Note that if x is not a vocalization, $y = 0$ and z takes an arbitrary value. At this stage, we can aggregate the frames coming from the two banks as a single table of triplets (x_i, y_i, z_i) , where $i = 1, 2, \dots$. As usual in machine learning, the table should be divided in three parts : a training dataset ($\approx 60\%$), a validation and a test dataset ($\approx 20\%$ each).

To serve as input to the YamNet deep neural network (*TensorFlow Hub* 2022), these one-second frames need to be represented in the time-frequency domain using a log-mel spectrogram. The log-mel spectrogram mimics the sensitivity of the human ear to frequency and amplitude differences. This transformation is a frequent pre-processing step of audio signal to input them into deep learning models. The following steps are required to obtain an input of dimension 96×64 : (1) apply the Short-Time Fourier Transform (STFT) with a window size of 25 ms, a hop size of 10 ms, and a Hann window function, (2) segment the spectrum into 64 mel bins spanning the frequency range of 125 - 7500 Hz, and (3) apply a logarithmic scaling. This process can be accomplished using one layer of a neural network. We have made this choice in our code for the sake of efficiency in our pipeline.

3.2.2. Network architecture

We rely on the YamNet network to move from the mel spectrogram of dimension 96×64 to a representation of the audio frame in dimension 1024 that is more relevant for the dual problem at hand. The first layer that computes the mel spectrogram, as well as the layers transferred from YamNet are not fitted on the data described in

3. Detection and classification of vocal productions in large scale audio recordings –
3.2. Methodology

TABLEAU 3.1. – Network Architecture

	Type	Input Size
YamNet	Log-scaled Mel Spectrogram	16000×1
	Convolution	$96 \times 64 \times 1$
	\vdots	\vdots
	Average Pooling	$3 \times 2 \times 1024$
	Fully Connected Batch-Normalization Drop-out	$1 \times 1 \times 1024$
Detection	Fully Connected Batch-Normalization Drop-out	$1 \times 1 \times m_1^d$
	\vdots	\vdots
	Fully Connected Batch-Normalization Drop-out	$1 \times 1 \times m_{(\ell^d-1)}^d$
	Sigmoid	$1 \times 1 \times m_{\ell^d}^d$
	Fully Connected Batch-Normalization Drop-out	$1 \times 1 \times 1024$
Classification	Fully Connected Batch-Normalization Drop-out	$1 \times 1 \times m_1^c$
	\vdots	\vdots
	Fully Connected Batch-Normalization Drop-out	$1 \times 1 \times m_{(\ell^c-1)}^c$
	Softmax	$1 \times 1 \times m_{\ell^c}^c$

$\ell^d, \ell^c \in [1, 6]$, learned on the data.
 $m_i^d \in [32, 1024]$, learned on the data, for $i = 1, \dots, \ell^d$
 $m_i^c \in [32, 1024]$, learned on the data, for $i = 1, \dots, \ell^c$

Section 3.2.1.

The dual problem we are trying to solve begins with the prediction of (y, z) given the observation x of a one-second frame. From a mathematical point of view, the problem can be reduced to the prediction of $K + 1$ -classes' problem by predicting $t = yz \in \{0, 1, 2, \dots, K\}$: we simply add a non-vocalization class (labeled by 0) to the classes of vocalization. Yet we expect that the difference between a silence or a noise from the soundscape and a vocalization is much larger than the subtle difference between two classes of vocalizations. Resolving the detection problem is thus a simpler learning task compared to the classification of vocalizations. We might expect a high error rate in the subsequent classification of detected vocalizations. Yet it is crucial to maintain strict control over the error rate in vocalization detection since our primary focus is on finding the seldom vocalization in continuous recordings. Furthermore, using the 1024-dimensional representation of YamNet, the detection problem should not be approached in the same way as the classification problem : the relevant coordinates

and/or the manner in which to use these coordinates to solve both problems are likely to be different. Based on these considerations and supported by preliminary numerical results (not presented in this article), we have chosen to treat the detection and classification problems separately while developing a single neural network.

Instead of directly predicting (y, z) , we construct a network that outputs estimates $\hat{p}(x)$ and $\hat{q}(x)$ of the posterior probability vectors :

$$\begin{aligned} p_k(x) &= \mathbb{P}(y = k|x), & k = 0, 1 \\ q_k(x) &= \mathbb{P}(z = k|x), & k = 1, \dots, K. \end{aligned}$$

The value of z indicates the class of the vocalization and is only meaningful if x is such a recording. Hence, our loss function is

$$L\left((y, z), (\hat{p}, \hat{q})\right) = -(1 - y) \log \hat{p}_0 - y \log \hat{p}_1 - y \left(\sum_{k=1}^K \mathbf{1}\{z = k\} \log \hat{q}_k \right) \quad (3.1)$$

where the second cross-entropy term plays a role only if $y = 1$, i.e., only if it is a vocalisation.

For all those reasons, we add two separated modules on top of the last layer of YamNet. They aim at estimating $p(x)$ and $q(x)$ respectively. Both need to be trained on data and are composed of fully connected layers with Parametric Rectified Linear Unit activation function (HE, X. ZHANG, REN et al. 2015). The weights of the two modules are initialized following the initialization scheme proposed by HE, X. ZHANG, REN et al. 2015, which should facilitate the convergence of the model (S. K. KUMAR 2017). The number of layers of each module is between 1 and 6, the number of nodes by layer between 32 and 1024. We rely on a regularization strategy to avoid problem of over-fitting : a batch normalization (IOFFE et SZEGEDY 2015) is computed after each layer, as well as drop-out and a max constraint on the norm of the weights (SRIVASTAVA, G. HINTON, KRIZHEVSKY et al. 2014). The activation function of the last layer of each module is either a sigmoid or a softmax function to get the desired posterior probabilities. The resulting architecture is described in Table 3.1.

3.2.3. Fit on the data

Gradient based algorithm show the advantage of our architecture. To learn from our training dataset, we minimize the loss in Equation (3.1) with the NAdam algorithm (DOZAT 2016; KINGMA et BA 2017). Many layers of our network are frozen and transferred. Yet the layers of the two modules should be fitted to the data. When the input data is a non-vocalization frame, the loss reduces to the cross-entropy of the detection module's output $\hat{p}(x)$. Its gradient with respect to the network weights to be adjusted is therefore zero on the weights of the classification module. On the other hand, when the input data is a classified vocalization frame, the loss is the sum of two cross-entropies, each computed on the output of one of the modules, either $\hat{p}(x)$ or $\hat{q}(x)$. In this case, the gradient decomposes into the sum of two vectors : the first

3. Detection and classification of vocal productions in large scale audio recordings –
3.2. Methodology

TABLEAU 3.2. – Hyper-parameters of the model

Hyper-parameters	Research Space
$p_{\text{drop-out}}$ (dropout within both modules)	[0.1, 0.9]
c_{norm} (batch-normalization within both modules)	[[0, 8]
α (learning-rate)	$[1e-10, 1e-2]$
β_1 (decay-rate of the moving average of the gradient)	[0, 0.9]
β_2 (decay-rate of the moving average of the squared gradient)	[0.99, 0.9999]
ℓ^d (number of hidden fully connected layers)	[[1, 6]
ℓ^c (number of hidden fully connected layers of the classification module)	[[1, 6]
m_i , for $i = 1, \dots, \ell^d$ and ℓ^c (number of nodes per layer)	[32, 1024]

vector only affects the weights of the detection module, and is obtained by taking the gradient of the detection cross-entropy; the second vector only affects the weights of the classification module, and is derived from the gradient of the classification cross-entropy. Thus, the adjustment of both modules can be done simultaneously, without the improvement of one module degrading the performance of the other module. However, if we had relied on a single module to solve the single classification problem with $(K + 1)$ classes by adding a non-vocalization class to the K vocalization classes, this observation would no longer hold. As the detection problem is simpler than the vocalization classification problem (see Section 3.2.2), the single network struggles to adjust to the dual problem. The two-module solution is a way to avoid bias due to the fact that detection is simpler than classification of vocalizations.

To avoid other bias, we need to take care of the way training frames enter in the NAdam algorithm, and deal with possible unbalanced training data. The solution we propose is as follows. Each batch is composed of N_{batch} triplets (x_i, y_i, z_i) drawn from the training dataset with replacement. Each triplet is drawn as follow. First, we draw y^\dagger uniformly over $\{0, 1\}$. If $y^\dagger = 0$, we draw the triplet at random among the non-vocalization triplets, namely $\{(x_i, y_i, z_i) : y_i = y^\dagger\}$. Otherwise, $y^\dagger = 1$, we draw z^\dagger uniformly over $\{1, \dots, K\}$ and pick a triplet at random among the vocalization triplets of class z^\dagger , namely $\{(x_i, y_i, z_i) : y_i = y^\dagger, z_i = z^\dagger\}$. Thus, on average, half of each batch is composed of non-vocalization frames; and among the vocalization frames, the K classes are balanced. Since the data seen by NAdam is an infinite flow of triplets drawn at random from the training dataset, we define an epoch as a set of $N_{\text{epoch}} = (K + 1)N_{\text{max}}/N_{\text{batch}}$ batches, where N_{max} is the size of the set of the largest class of vocalization and N_{batch} the size of the batch. Along the NAdam algorithm, the learning rate is decreased by a factor of 0.2 if the validation loss has not decreased after 5 epochs. And the whole algorithm is stopped after 20 epochs without validation error decrease.

The hyperparameters of the two modules are calibrated on the validation dataset using a Bayesian optimization scheme (BROCHU, CORA et DE FREITAS 2010; SNOEK, LAROCHELLE et Ryan P ADAMS 2012; SHAHRIARI, SWERSKY, Ziyu WANG et al. 2016) which minimizes the validation computed with the loss given in Equation (3.1). The parameters of the NAdam algorithm, as well as the parameters that defines the precise architecture of the two modules are calibrated with this method, are given in Table 3.2.

3.2.4. Vocalization delineation and classification

The use of the trained network to delineate a vocalization period on a continuous audio recording as well as its classification needs to be explained. We need a conciliation procedure that reintroduce the temporal dependency of our overlapping frames that was lost by the network. More precisely, we need a rule that allows errors in the detection of vocalization frames, and in the classification of detected vocalization frames. To this aim, we design the procedure given below with the following rules. First, a vocalization period can include period of time of length less than one second where vocalizations have not been detected by the network. Second, to aggregate the class predictions, we use a majority vote.

As before, the recording to be analyzed of length T^* seconds should be divided into frames of 1s with an overlap of 80%. Let us denote x_t^* , $t = 1, \dots, 5T^*$ the t -th frame of this recording. Using the trained network, we can compute for each t the maximum a posteriori :

$$\hat{y}_t^* = \begin{cases} 1 & \text{if } \hat{p}_1(x_t^*) > 0.5, \\ 0 & \text{otherwise} \end{cases} \quad \hat{z}_t^* = \operatorname{argmax}_k \hat{q}_k(x_t^*).$$

In order to delineate a vocalization segment based on the predicted values \hat{y}_t^* , $t = 1, \dots$, we introduce an equivalence relation on the set $\mathcal{T}_1 = \{t : \hat{y}_t^* = 1\}$ as follows. We say that $t \sim t'$ if and only if there exists an increasing sequence $t_0, \dots, t_N \in \mathcal{T}_1$ such that $t_0 = t$, $t_N = t'$ and $t_{i+1} - t_i \leq 5$. It means that the audio segment starting with the frame at position t and ending with the frame at position t' is composed of frames that are predicted as vocalizations, except on small time periods that last less than one second. The equivalence classes of this relation are easy to determine along time. The position of the starts and ends of the vocalization segments along the continuous recording are then given by the starts and the ends of the equivalence classes of this equivalence relation.

Once a vocalization segment is delineated, we have to predict its class. To this aim, we use the predicted classes \hat{z}_t^* of the frames that compose the segment. Considering our loss function given in Equation (3.1), the predicted classes are reliable only on the frames that have been detected as vocalizations. Among these reliable predictions, a majority vote allows us to determine the class of the vocalization segment.

3.3. Experimental Validation

The proposed pipeline has been tested in two different studies : a first study dealing with baboon vocalizations and a second one dealing with human baby vocalizations. The first study is a bioacoustic problem that aims at collecting vocalizations of Guinea baboons (*Papio papio*). The second study is a developmental psycho-acoustic problem and is focused on the vocalizations of human infants between 0 and 12 months of age. For each study, the output is a large-scale database of labeled vocalizations.

The data of the two studies are described in Section 3.3.1.

3.3.1. From audio recordings to data banks for our method

Both studies aim at capturing vocalizations that are not provoked by the experimental setup, using continuous and possibly daylong recordings. However, the recording conditions are quite different. These two studies provide a first example of the diversity of situations to which our method can adapt, as each of them contains different sound events other than vocalizations, as well as different background noises that interfere with the vocalizations.

In the baboon study, we recorded continuously during approximately one month a group of 25 Guinea baboons (*Papio papio*) from the CNRS primatology center of Rousset-sur-Arc (France). The group lives in semi-liberty in a large rectangular enclosure outdoors. Ethical agreement (# 02054.02) was obtained from the CEEA-14 for experimental animal research to conduct audio recordings of the baboons' vocalizations. Two microphones are placed at two corners of the enclosure. In addition to the baboons' vocalizations, the sound environment is composed of climatic events (wind, rain), the presence of other nearby animal species (sheep, birds), and human activities (people around the enclosure, cars on the nearby highway, planes, etc.). One month of recording leads to a tremendous amount of data : after removing night recordings when baboons are at sleep inside a room (from 9 pm to 7 am), there is a total of 460 files representing 443 hours of recording (i.e., 1 595 018.24 seconds).

In the human study, we collected recordings from two human babies at home from birth to their first birthday, at a rate of three days per month. An ethical agreement (# 2019-12-12-005) was obtained from the ethics committee of Aix-Marseille University as well as a declaration of conformity from the CNIL (# 2222631 v 0) for experimental research on humans in order to make audio recordings of human baby vocalizations. All parents gave their informed consent for inclusion before their inclusion in the study. The records were done by the parents at different moments of the days and nights. The parents were instructed to start and stop themselves the recordings. Although less noisy than the baboon environment, the recordings are composed of a lot of heterogeneous sources of sounds : TV, radio, domestic works, parents, other children. In total, the records represent 174.15 hours (626 940 seconds) for the two children.

Both studies had some uncontrolled recording conditions. In the baboon study, the microphone had a fixed position and the signal source is mobile. The monkeys move and vocalize from various location into different direction. In the human study, the microphone is constantly changing position. With each new recording, the parents place the microphone in a position that may be different from the source, the baby.

In addition to the continuous sound recordings to be analyzed, our method is based on two recording banks for each study : a first bank of non vocalization recordings, and a second bank of labeled vocalization recordings. In both studies, the first bank that helps to delineate vocalizations was extracted from the continuous recordings, as proposed in Section 3.2.1. In the baboon study, we listened to a total amount of 7 hours of these recordings from which we removed the vocalizations. In the human study, we used the same method on a total amount of 5 hours of recordings. In both cases, we took care to get a bank as representative as possible of the various sound

3. Detection and classification of vocal productions in large scale audio recordings – 3.3. Experimental Validation

events and noises : the excerpts we listened to were chosen at different dates and times of the day and came from different families. For the baboon study, it represents 355.62 minutes. For the baby study, it represents 104.56 minutes.

The bank of labeled vocalization recordings were constructed differently in both studies. In the baboon study, we had from a previous study (BOË, BERTHOMMIER, LEGOU et al. 2017) a total amount of 72.49 minutes of labeled vocalization, divided into 6 classes : bark, copulation grunt, grunt, scream, wahoo, and yak. These baboon vocalizations came from the same group of baboons and from the same experimental setup. In contrast, the baby vocalizations came from a public database (Meg CYCHOSZ, SEIDL, Erika BERGELSON et al. 2019), based on daylong audio recordings of 49 children (1–36 months) from five different languages and cultural backgrounds that were annotated by citizen scientists. This public database gave us a bank of labeled vocalizations that represents a total amount of 77.03 minutes of recordings, divided into 5 classes : canonical, crying, junk, laughing, non-canonical. In both studies, the classes in the vocalization bank were unbalanced. More details on the composition of the two banks in both studies are given in the Supplementary Materials (Table S3.1 and S3.3)

3.3.2. Performance of our deep learning architecture

Before presenting the final results of the proposed methodology, we start by analyzing the outputs of our network on the 1-second frames from the test datasets of both studies : Table 3.3 provides different evaluations of the network’s performance on our dual problem, calculated using standard metrics. Detailed confusion matrices for the detection problem as well as the classification problem in both studies are given in Supplementary Materials (Figures S3.1 and S3.2). At first glance, the results are relatively similar, suggesting that our pipeline can be used successfully in different types of studies. In particular, the performances in the detection problem measured with the precision and recall metrics indicate that we succeeded in our primary goal of detecting correctly the vocalization frames.

As can be seen in Table 3.3, the detection problem is resolved more satisfactorily in the study on human infants than in the study on baboons. We can propose two explanations for this discrepancy. Firstly, the quality of the recordings of baby vocalizations is much better than that of baboons : they were made in a much quieter place, at quieter times of the day, and with a microphone likely closer to the source, whereas baboons move in a noisier and larger environment. Particularly, on windy days (Mistral), the recordings of the baboon study are of very poor quality. Secondly, the 1024-dimensional representation provided by YamNet is likely more suited to detect human vocalizations than baboon vocalizations. Indeed, this representation was learned on the Audioset database to solve a classification problem with 521 classes. This massive database contains numerous antropophonic sounds distributed across several classes. But vocalizations of animals are much rarer and distributed across coarser classes.

On the other hand, we can see in Table 3.3 that the classification problem is resolved more satisfactorily in the baboon study even if it is a 6-classes problem whereas the

3. Detection and classification of vocal productions in large scale audio recordings – 3.3. Experimental Validation

TABLEAU 3.3. – Performance of the deep learning in the baboon and human studies

		Data set	
		Baboon	Baby
Loss (from equation 3.1)		0.12	0.07
Detection	Cross-entropy	0.04	0.01
	Accuracy	94.58	99.76
	AUC	0.94	0.99
	Precision	82.68	99.33
	Recall	90.28	99.74
Classification	Cross-entropy	0.23	0.26
	Accuracy	48.92	39.96

baby vocalizations are divided in 5 classes only. Even though YamNet was not originally designed to understand the differences between baboon vocalizations, the results of our methodology are relatively good at classifying them. Moreover, the distinction between the classes of infant vocalizations is likely more subtle than the differences between the classes of baboon vocalizations. The study on infants marks the beginning of an investigation into language development. In the early months, the infants are still in the process of learning this classification for themselves, and the differences between the vocalizations produced can sometimes be subtle. A more closer look at the confusion matrices given in Supplementary Materials shows that the major part of errors in the baboon study comes from rather similar classes : “copulation grunt” which is a specific class in our study can be seen as a specific type of “grunt” and a “bark” vocalization share common features with a “yahoo”.

We processed the continuous audio recordings for both problems on a laptop with a single GPU, on which Tensorflow (DEVELOPERS 2022) was able to train and run the deep network. The 460 files representing 443 hours of continuous recording of the baboon have been loaded, segmented, and classified in 9 hours 28 minutes. The 261 files representing 174 hours of continuous recording for the two human babies have been loaded, segmented, and classified in 9 hours 44 minutes.

3.3.3. New large-scale databases of vocalizations

Once the model has been trained on the labeled data, we can use it on the massive continuous data to extract the moments of vocalization and create two new large-scale data sets. We can measure the amount of data extracted and the time to do it.

Two new databases are constructed from the continuous recordings processed by the model through our pipeline. The new human baby database represents 35.20 hours of records. The new baboon database represents 38.75 hours of records. Table S3.2 and S3.4 respectively, summarizes the distribution for each class, for each data.

We have made the data extracted by our model from continuous baboons recordings freely accessible on <https://zenodo.org/record/8239697>. This includes the labeled database used for training, as well as the vocalizations detected during the

month and a csv file summarizing, for each vocalization, its duration, probability and vote for each class. A typical day was also made available, to give everyone the chance to test the model on this typical example (it was not possible to make more than one day of continuous recording available for legal reasons). The code is accessible on <https://gitlab.com/papers4375727/detection-and-classification-of-vocal-productions.git>, with an example with this day to reproduce the work. Results for baby recordings are not accessible for legal reasons.

3.4. Conclusion and Discussion

The goals of the pipeline were to quickly process and classify hundreds of hours of audio, with as few errors as possible, minimizing information loss, through an end-to-end pipeline with no engineering steps, so that it can be reused in different situations. In addition, the pipeline had to adapt to various environmental sound classification problems, with little labeled data for learning.

Our two-module architecture, together with the care taken with the training set, enables us to achieve high scores on precision and recall in the vocalization detection problem. This was the primary objective of our methodology, and we can consider that it has been achieved without mobilizing massive computing resources, thanks to the transfer of YamNet. Even on vocalizations with which YamNet is unfamiliar, such as those of baboons, the detection scores (precision and recall) remain very high, showing that our method is capable of attacking a certain diversity of species.

The two studies of Section 3.3 show that YamNet has sufficient generality to tackle a wide class of problems similar to the one addressed in this work. AudioSet is massive enough to learn a representation which distinguish between one type of signal and another, between a species' vocalization and the rest of its soundscape. The two studies show that our method is robust to a variety of complicated recording conditions, and generic enough for use in a variety of contexts, species,... The limits are certainly in the frequency range to which YamNet is sensitive. This range is similar to that of the human ear; it does not allow us to deal with species such as bats, for example, which emit sounds in the high treble or ultra-sounds. The representation transferred by YamNet is undoubtedly not the most appropriate for tackling a problem without adopting an anthropocentric stance, since it is based on a representation designed for the human ear (PRAT 2019).

Thanks to our two-module architecture, we can handle a second problem simultaneously with detection, such as the classification of vocalizations, without degrading detection scores. Classification scores are less satisfactory, but this problem, like many learning problems on vocalizations, requires the learning model to be able to distinguish more subtle differences. Probably, our training bases were too limited, representing a total of one hour of recording in which our classes were unbalanced.

The conciliation procedure we have introduced in Section 3.2.4 to reintroduce the temporal dependency is quite rough. We discarded other attempts that introduced too

many computational burden for the output provided. Yet, we have lost the uncertainty measure provided by the network through $\hat{p}(x)$ and $\hat{q}(x)$. It would be interesting to develop a probabilistic method capable of performing this reconciliation, without weighing down our pipeline numerically.

Declarations

Funding

This work was carried out in a collaboration between the CNRS, Aix-Marseille University and Résurgences R&D around the CIFRE PhD n°215582 with the support of the ANRT, within the Labex BLRI (ANR-11-LABX-0036) and the Institut Convergence ILCB (ANR-16-CONV-0002). It also benefited support from the French government, managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). This work was also supported by the CHUNKED ANR project (ANR-17-CE28-0013-02). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We are grateful to Rosic Ferry-Huiban and Myriam Sabatier for their help in labeling the baboon database.

Conflicts of interest

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

Ethical agreement (# 02054.02) was obtained from the CEEA-14 for experimental animal research to conduct audio recordings of the baboons' vocalizations. An ethical agreement (# 2019-12-12-005) was obtained from the ethics committee of Aix-Marseille University as well as a declaration of conformity from the CNIL (# 2222631 v 0) for experimental research on humans in order to make audio recordings of human baby vocalizations.

Consent to participate

Written informed consent was obtained from the parents.

Consent for publication

Parents consent regarding publishing analysis from their recordings.

Availability of data and materials

Data set created on the continuous baboons recordings is accessible on <https://zenodo.org/record/7963> with the training set used for this study and two hours as example. Data set created on the continuous baby recordings is not accessible for legal reasons. The labeled data for the babies came from a public database named Babblecor (Meg CYCHOSZ, SEIDL, Elika BERGELSON et al. 2019) which can be found on <https://osf.io/rz4tx/>

Code availability

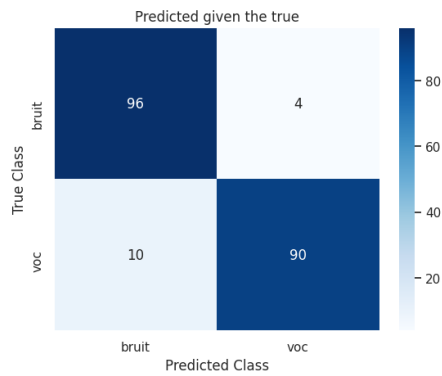
The pipeline proposed in this paper is available on GitLab at <https://gitlab.com/papers4375727/detection-and-classification-of-vocal-productions.git>. The repository contains the implementation of the algorithms described in the methods section, along with detailed documentation on how to use the code.

Authors' contributions

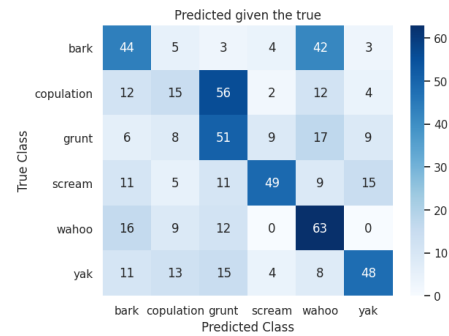
The authors confirm contribution to the paper as follows : study conception and design : G. Bonafos, P. Pudlo, J-M. Freyermuth, A. Rey, S. Tronçon; data collection : G. Bonafos, T. Legou, J. Fagot; analysis and interpretation of results : G. Bonafos, P. Pudlo; draft manuscript preparation : G. Bonafos, P. Pudlo, J-M. Freyermuth, A. Rey. All authors reviewed the results and approved the final version of the manuscript.

For the purpose of Open Access, a CC-BY¹ public copyright licence has been applied by the authors to the present document and will be applied to all subsequent versions up to the Author Accepted Manuscript arising from this submission.

3.5. Supplementary materials



(a) Detection of vocalizations against noise.

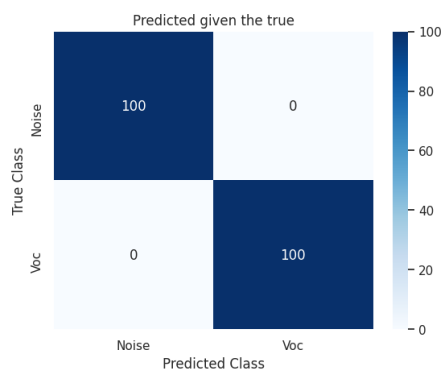


(b) Classification of the detected vocalizations.

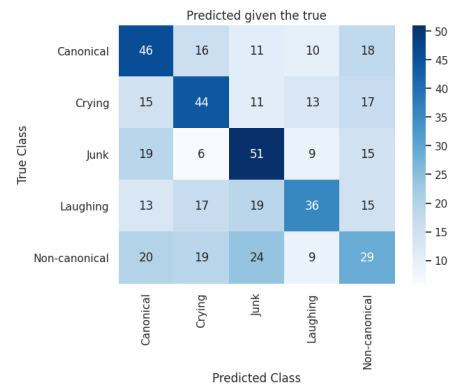
FIGURE S3.1. – Confusion matrices for the *baboon* data.

1. <https://creativecommons.org/licenses/by/4.0/>

3. Detection and classification of vocal productions in large scale audio recordings –
3.5. Supplementary materials



(a) Detection of vocalizations against noise.



(b) Classification of the detected vocalizations.

FIGURE S3.2. – Confusion matrices for the *human baby* data.

TABLEAU S3.1. – Total and per partition distribution of the baboon labeled data set.

	Classes					
	Bark	Copulation	Grunt	Scream	Wahoo	Yak
Number of records	269	68	502	414	97	60
Duration in seconds	192.1	132.1	2007.5	1695.8	146.4	175.5
Mean record time (secs)	1.3	3.0	6.7	7.0	2.4	4.9

(a) Initial sample size, number of records and duration for each vocalization class, for the baboon data set.

	Classes						
	Bark	Copulation	Grunt	Scream	Wahoo	Yak	Noise
Numbers of records	152	44	302	244	51	36	82
Duration in seconds	98.5	84.3	1252.9	951.6	105.3	99.1	13884.7
Mean record time (secs)	0.7	1.9	4.2	3.9	1.7	2.8	169.3

(b) Number of records and their duration per vocalization class for the training set of the baboon data set.

3. Detection and classification of vocal productions in large scale audio recordings –
3.5. Supplementary materials

TABLEAU S3.1. – Total and per partition distribution of the baboon labeled data set.

	Classes						
	Bark	Copulation	Grunt	Scream	Wahoo	Yak	Noise
Numbers of records	54	11	101	89	20	12	20
Duration in seconds	36.1	20.7	362.0	340.2	20.3	54.2	3379.0
Mean record time (secs)	0.2	0.5	1.2	1.4	0.3	1.5	169.0

(c) Number of records and their duration per vocalization class for the validation set of the baboon data set.

	Classes						
	Bark	Copulation	Grunt	Scream	Wahoo	Yak	Noise
Numbers of records	63	13	99	81	16	12	24
Duration in seconds	57.5	27.2	392.6	404.0	20.7	22.2	4073.4
Mean record time (secs)	0.4	0.6	1.3	1.7	0.3	0.6	169.7

(d) Number of records and their duration per vocalization class for the testing set of the baboon data set.

TABLEAU S3.2. – Seconds of vocalization for baboons, for each class, over the month.

Classes					
Bark	Copulation	Grunt	Scream	Wahoo	Yak
3 197	8 455	50 679	35 070	23 026	19 080

3. Detection and classification of vocal productions in large scale audio recordings –
3.5. Supplementary materials

TABLEAU S3.3. – Total and per partition distribution of the human baby labeled data set

	Classes				
	Canonical	Crying	Junk	Laughing	Non-Canonical
Numbers of records	1826	823	4974	241	5606
Duration in seconds	677.9	297.6	1665.1	87.1	1894.0
Mean record time (secs)	0.4	0.4	0.4	0.4	0.4

(a) Initial sample size, number of records and their duration per vocalization class for the *human baby* data set.

	Classes					
	Canonical	Crying	Junk	Laughing	Non-Canonical	Noise
Numbers of records	1057	490	2988	138	3407	48
Duration in seconds	390.4	176.5	996.7	49.9	1148.6	3748.5
Mean record time (secs)	0.4	0.4	0.4	0.4	0.4	78.1

(b) Number of records and their duration per vocalization class for the training subset of the *human baby* data set.

	Classes					
	Canonical	Crying	Junk	Laughing	Non-Canonical	Noise
Numbers of records	386	169	982	51	1106	16
Duration in seconds	142.8	62.0	331.4	18.6	372.1	753.4
Mean record time (secs)	0.4	0.4	0.4	0.4	0.4	47.1

(c) Number of records and their duration per vocalization class for the validation subset of the *human baby* data set.

3. Detection and classification of vocal productions in large scale audio recordings –
 3.5. Supplementary materials

TABLEAU S3.3. – Total and per partition distribution of the human baby labeled data set

	Classes					
	Canonical	Crying	Junk	Laughing	Non-Canonical	Noise
Numbers of records	383	164	1004	52	1093	17
Duration in seconds	144.6	59.2	337.0	18.6	373.3	1772.0
Mean record time (secs)	0.4	0.4	0.4	0.4	0.4	104.2

(d) Number of records and their duration per class for the testing set of the *human baby* data set.

TABLEAU S3.4. – Seconds of vocalization for human babies, for each class, over the year

Canonical	Classes			
	Crying	junk	Laughing	Non-canonical
3	248	11 893	966	113 627

4. Analyse Topologique du Signal

Sommaire

4.1. Remise en contexte et résumé de l'étude	82
4.2. La forme des données	84
4.2.1. Complexe simplicial géométrique	84
4.2.2. Complexe simplicial abstrait	85
4.2.3. Espace sous-jacent	86
4.3. Notions d'équivalence	86
4.3.1. Homéomorphisme	86
4.3.2. Équivalence homotopique	87
4.3.3. Théorème du Nerf	88
4.4. Homologie	90
4.4.1. Chaînes, cycles et bords	90
4.4.2. Groupe d'homologie	92
4.5. Persistance	93
4.5.1. Topologie et échelle	93
4.5.2. Filtration	94
4.5.3. Homologies persistantes	95
4.6. Diagramme de persistance	97
4.6.1. Définition d'un diagramme de persistance	97
4.6.2. Stabilité d'un diagramme de persistance	98
4.6.3. Consistance d'un diagramme de persistance	99

4.1. Remise en contexte et résumé de l'étude

Dans le Chapitre 3 précédent, nous avons présenté notre contribution permettant de traiter des données massives et de trouver le signal dans des enregistrements continus. Ce problème résolu, la question suivante porte sur la représentation adaptée de ce signal pour une tâche de classification.

La bonne représentation des données dépend du problème que l'on étudie. Il est nécessaire de la construire à la lumière de la question que l'on se pose, afin de représenter au mieux le signal et d'extraire l'information utile pour notre problème et le travail statistique que l'on souhaite faire. En l'occurrence, étant donné l'intérêt pour l'approche topologique que l'on a présenté dans la partie 1.3.3, on souhaite que cette représentation prenne en compte des caractéristiques topologiques du signal. Néanmoins, bien qu'en fort développement, la TDA n'a pas encore été utilisée à notre

4. Analyse Topologique du Signal – 4.1. Remise en contexte et résumé de l'étude

connaissance pour traiter des questions de développement du langage. L'application même de méthode topologique pour analyser des données sonores est assez limitée. Nous voulons donc d'abord quantifier expérimentalement le gain informatif obtenu par une représentation du signal vocal prenant en compte des caractéristiques topologiques. Nous voulons aussi avoir une idée et une intuition de l'information qualitative que des descripteurs topologiques du signal permettraient de mettre en lumière. Afin de s'assurer de la pertinence de l'approche topologique pour analyser un signal vocal humain, pour déterminer la meilleure façon de procéder et pour quantifier l'impact des différents choix que l'on fait, on étudie sur un problème contrôlé l'intérêt et les méthodes de la TDA. Ce travail est la deuxième contribution de la thèse.

Pour cela, nous enregistrons 20 adultes francophones (15 femmes et 5 hommes) prononcer 8 voyelles (ä, ã, ə, i, o, õ, u, y), sous 7 conditions (naturel, voix basse, voix haute, long, court, tonalité ascendante, tonalité descendante), 10 fois chacune. Cela nous permet d'avoir $20 \times 8 \times 7 \times 10 = 11200$ enregistrements pour lesquels nous avons trois informations : la voyelle prononcée, le sexe de l'émetteur et son identité, soit trois tâches de classification potentiels. Nous avons rendu la base de données produite pour ce papier librement accessible <https://zenodo.org/record/7961904>.

Pour chacun des enregistrements, on calcule les spectrogrammes, les zéros du spectrogramme et les plongements de Taken. Pour chacun de ces objets, on calcule les diagrammes de persistance. On vectorise ensuite ces diagrammes pour construire une représentation topologique de chaque enregistrement. Pour chacun des trois problèmes de classification, on estime une forêt aléatoire et on compare les résultats de classification selon la représentation du signal. On calcule aussi les MFCC de chaque enregistrement pour ajouter à la comparaison une représentation sans information topologique, ainsi qu'une représentation augmentée topologiquement (MFCC + descripteurs topologiques).

Ce travail sur des données vocales réelles mais contrôlées emmène une preuve expérimentale de la valeur ajoutée d'une représentation du signal prenant en compte une information topologique de celui-ci pour un exercice de classification, compare les différentes stratégies pour calculer des diagrammes de persistance d'un signal sonore (sur quel objet?) ainsi que les différentes méthodes pour les exploiter. On retrouve ce travail dans le Chapitre 4 suivant.

Afin d'introduire l'étude, ce Chapitre présente plus en détail le *pipeline* de la TDA, ses différents objets et son assise théorique.

Ce Chapitre et les définitions présentées sont basées sur DEY et Y. WANG 2022; H. EDELSBRUNNER et HARER 2009; Frédéric CHAZAL et MICHEL 2021; CARLSSON 2009; A. J. ZOMORODIAN 2005.

4.2. La forme des données

4.2.1. Complexe simplicial géométrique

Le point de départ de la TDA est de considérer que les données ont une forme et que cette forme contient une information pertinente sur le processus sous-jacent les ayant produites. Comme les données que nous avons à analyser sont supposées des échantillons d'un objet géométrique sous-jacent, on commence par y poser dessus la structure qui sera la forme que dessinent nos données. Cette structure combinatoire permettra de plus les calculs efficaces d'invariants permettant de la caractériser. Pour cela, on va s'appuyer sur les outils de la topologie.

Définition 4.2.1 (Topologie). *Pour un ensemble de points X on définit une topologie sur l'ensemble de points X , comme une collection \mathbb{T} de sous-ensembles de X , ses ouverts (open sets), tels que :*

- *l'ensemble X et l'ensemble vide \emptyset sont des éléments de $\mathbb{T} : X, \emptyset \in \mathbb{T}$;*
- *pour une famille d'éléments $\{T_1, \dots, T_n\} \subseteq \mathbb{T}$, possiblement infinie, leur union est dans $\mathbb{T} : \bigcup_i T_i \in \mathbb{T}$;*
- *pour une famille finie d'éléments $\{T_1, \dots, T_n\} \subseteq \mathbb{T}$, l'intersection des éléments est dans $\mathbb{T} : \bigcap_{i=1}^n T_i \in \mathbb{T}$.*

La paire (X, \mathbb{T}) est appelée un espace topologique. On omettra \mathbb{T} et on fera référence à X comme un espace topologique.

Ainsi, une topologie est simplement un système d'ensembles qui décrit la connectivité d'un ensemble. Un espace topologique est un ensemble de points muni d'un ensemble de relations de voisinage. Ce sont ces relations de voisinage qui vont nous permettre de définir rigoureusement la forme de X .

À noter, la notion d'espace topologique inclut la notion d'espace métrique. Aussi, si un espace est un espace métrique, c'est aussi un espace topologique (l'inverse n'étant pas nécessairement vrai, tout espace topologique n'est pas un espace métrique). En associant une topologie à notre espace, on associe des relations de voisinages entre les points. On s'affranchit des questions de métriques et de coordonnées en étudiant les propriétés géométriques d'un espace par sa connectivité.

Sur ces fondations, nous pouvons poser la première brique pour construire la forme des données : le simplexe.

Définition 4.2.2 (k -simplexe). *Pour $k \geq 0$, un k -simplexe σ est l'enveloppe convexe d'un ensemble de $k + 1$ points $S = \{v_0, \dots, v_k\}$ affinement indépendants. Les points dans S sont les vertexes du simplexe.*

Un k -simplexe est un sous-espace de \mathbb{R}^d de dimension k . C'est la généralisation à toutes les dimensions du triangle. Ainsi, un 0-simplexe est un point, un 1-simplexe un segment, un 2-simplexe un triangle, un 3-simplexe un tétraèdre. La dimension de S pour chacun de ces simplexes croît d'un en un. Un point suffit à créer un 0-simplexe, deux un 1-simplexe, trois un 2-simplexe, quatre un 3-simplexe.

Un simplexe est constitué de faces. Chaque sous-ensemble $\sigma' \subseteq \sigma$ est une face de σ .

Définition 4.2.3 (Face). Soit σ un k -simplexe défini par $S = \{v_0, v_1, \dots, v_k\}$. Un simplexe τ défini par $T \subseteq S$ est une face de σ et a σ comme co-face. La relation est notée par $\sigma \geq \tau$ et $\tau \leq \sigma$.

Un k -simplexe a $\binom{k+1}{\ell+1}$ faces de dimension ℓ et $\sum_{\ell=1}^k \binom{k+1}{\ell+1} = 2^{k+1}$ faces au total. Un simplexe est donc un objet combinatoire large mais simple. Surtout, les simplexes peuvent être associés ensemble afin construire un objet plus grand permettant de représenter des espaces : les complexes simpliciaux.

Définition 4.2.4 (Complexe simplicial). Un complexe simplicial K est un ensemble fini de simplexes tels que :

- $\sigma \in K, \tau \leq \sigma \Rightarrow \tau \in K$;
- $\sigma, \sigma' \in K \Rightarrow \sigma \cup \sigma' \leq \sigma, \sigma'$.

La dimension du complexe simplicial K est la dimension du plus grand simplexe dans K . Un complexe simplicial est une collection de simplexes correctement arrangés.

Ainsi, pour nous intéresser à la variété sous-jacente de nos données discrètes, on couvre la variété, représentée par les données, par des simplexes. Puis, on les utilise pour construire un complexe qui permet d'avoir une structure combinatoire qui trace la forme des données. En effet, à partir de cette structure, il nous est désormais possible de compter les différents éléments qui caractérisent cette forme, et notamment les éléments invariants.

4.2.2. Complexe simplicial abstrait

Nous venons de voir la définition géométrique d'un complexe simplicial. Mais celui-ci peut être défini sans faire référence à la géométrie, juste par la topologie, ce qui en fait d'ailleurs tout son intérêt computationnel.

Définition 4.2.5 (Complexe simplicial abstrait). Un complexe simplicial abstrait est un ensemble K associé à une collection \mathcal{S} de sous-ensembles de K appelés les simplexes (abstraites) tels que :

- Pour tout $v \in K, \{v\} \in \mathcal{S}$. On appelle les ensembles $\{v\}$ les vertexes de K .
- Si $\tau \subseteq \sigma \in \mathcal{S}$, alors $\tau \in \mathcal{S}$.

On dit que σ est un k -simplexe de dimension k si $|\sigma| = k + 1$. Si $\tau \subseteq \sigma$, τ est une face de σ et σ est une co-face de τ .

On peut lier cette définition abstraite du complexe simplicial basée sur la théorie des ensembles à la définition géométrique précédente du complexe simplicial.

Théorème 4.2.1 (Réalisation géométrique). Tout complexe simplicial abstrait de dimension d a une réalisation géométrique dans \mathbb{R}^{2d+1} .

4. Analyse Topologique du Signal – 4.3. Notions d'équivalence

On trouve la preuve chez H. EDELSBRUNNER et HARER 2009. Ainsi, un complexe simplicial abstrait est un espace topologique et un complexe géométrique est la réalisation géométrique de sa structure sous-jacente (Frédéric CHAZAL et MICHEL 2021). On peut donc associer à un complexe simplicial abstrait un complexe simplicial géométrique plongé dans un espace, dont les propriétés topologiques sont équivalentes. On peut décomposer tout complexe simplicial en son composant topologique ou géométrique. Le premier est un complexe abstrait, objet purement combinatoire, pratique pour être enregistré et manipulé. Le deuxième projette les vertexes du complexe dans l'espace dans lequel il se réalise.

Les complexes simpliciaux peuvent être utilisés pour représenter les variétés.

4.2.3. Espace sous-jacent

Définition 4.2.6 (Espace sous-jacent). *L'espace sous-jacent d'un complexe simplicial abstrait K , noté $|K|$, est l'union point à point de ses simplexes dans sa réalisation géométrique canonique; i.e., $|K| = \bigcup_{\sigma \in K} \sigma$.*

À noter, $|K|$ est un espace topologique. Il hérite de la topologie de l'espace ambiant dans lequel vivent ses simplexes.

Définition 4.2.7 (Triangulation). *Étant donné un complexe simplicial K et une variété M , on dit que K est une triangulation de M si son espace sous-jacent $|K|$ est homéomorphe à M .*

Une triangulation d'un espace topologique X est donc un complexe simplicial K tel que $|K| \simeq X$.

4.3. Notions d'équivalence

On a posé sur nos données une structure permettant de nous intéresser à leur forme, un complexe simplicial. Celui-ci est un objet abstrait facile de manipulation et utile pour les calculs, qui a une réalisation géométrique. Néanmoins, il nous reste à nous assurer que les invariants que l'on calcule sur cet espace topologique sont bien les mêmes que ceux de la variété sous-jacente qui nous intéresse. Pour cela, il convient de définir les notions d'équivalence qui existent entre ces différents objets, ainsi que le formalisme mathématiques que l'on utilise pour analyser de manière quantitative et non-ambiguë la connectivité d'un espace.

4.3.1. Homéomorphisme

Nous avons des espaces topologiques qui tracent la forme de nos données, les complexes simpliciaux. Afin de nous assurer que leurs invariants nous permettent bien de dire quelque chose sur la variété sous-jacente, on cherche une relation d'équivalence entre les deux objets.

Définition 4.3.1 (Équivalence). *Soit S un ensemble non-vide et \sim une relation entre éléments qui satisfait les propriétés suivantes, pour tout $a, b, c \in S$.*

- *Réflexivité. $a \sim a$.*
- *Symétrie. Si $a \sim b$, alors $b \sim a$.*
- *Transitivité. Si $a \sim b$ et $b \sim c$ alors $a \sim c$.*

Alors, la relation \sim est une relation d'équivalence sur S .

La notion d'homéomorphisme permet de définir une relation d'équivalence entre espaces topologiques.

Définition 4.3.2 (Homéomorphisme). *Un homéomorphisme $f : X \rightarrow Y$ est une application bijective dont l'inverse f^{-1} est aussi continue.*

Un homéomorphisme est la notion rigoureuse permettant d'assurer qu'une opération préserve la topologie du domaine. Deux espaces sont dits homéomorphiques s'il existe un homéomorphisme entre eux. Un homéomorphisme induit une relation d'équivalence entre espaces topologiques. Pour cela, deux espaces topologiques homéomorphiques sont dits topologiquement équivalents.

Si la notion d'homéomorphisme permet d'établir une équivalence topologique entre deux espaces, c'est une notion forte, trop forte en pratique. En cela, la notion d'équivalence homotopique est plus utilisée.

4.3.2. Équivalence homotopique

La notion d'équivalence homotopique est une autre notion de similarité entre espaces topologiques, mais plus faible que la notion d'homéomorphisme. L'homotopie préserve certaines formes de connectivité, comme le nombre d'éléments connectés, les cavités et tunnels, bref les trous de différentes dimensions d'un espace.

Pour définir l'équivalence homotopique, nous avons d'abord besoin du concept d'homotopie.

Définition 4.3.3 (Homotopie). *Une homotopie est une famille d'applications $f_t : X \rightarrow Y$, $t \in [0, 1]$, telle que l'application associée $F : X \times [0, 1] \rightarrow Y$ donnée par $F(x, t) = f_t(x)$ est continue. Deux applications $f_0, f_1 : X \rightarrow Y$ sont homotopiques via l'homotopie f_t .*

Alors que la notion d'homéomorphisme relie deux espaces topologiques, l'homotopie lie deux applications, créant un lien indirectement entre deux sous-espaces, $f_0(X) \subseteq Y$ et $f_1(X) \subseteq Y$. Cette relation n'est pas nécessairement une relation d'équivalence, mais la suivante.

Définition 4.3.4 (Équivalence homotopique). *Une application $f : X \rightarrow Y$ est appelée une équivalence homotopique s'il existe une application $g : Y \rightarrow X$ telle que $f \circ g \simeq 1$ et $g \circ f \simeq 1$. Alors, X et Y sont homotopiquement équivalents et ont le même type homotopique.*

Alors que des espaces homéomorphiques ont nécessairement la même dimension, ce n'est pas le cas d'espaces homotopiquement équivalents.

Deux espaces homotopiquement équivalents partagent les mêmes invariants topologiques. Cette notion d'équivalence est celle utilisée dans le théorème du Nerf, qui permet d'appuyer et prouver théoriquement le lien entre le complexe simplicial que l'on construit sur nos données et la variété sous-jacente.

4.3.3. Théorème du Nerf

On commence par définir un recouvrement.

Définition 4.3.5 (Recouvrement). *Le recouvrement d'un espace topologique X est la collection C d'ensembles ouverts tels que $X = \bigcup_{c \in C} c$. L'espace topologique X est dit compact si tous ses recouvrements C ont un sous-recouvrement fini, i.e., il existe $C' \subseteq C$ tel que $X = \bigcup_{c \in C'} c$ et C' sont finis.*

Le recouvrement d'un espace topologique permet de définir un complexe simplicial particulier, le nerf.

Définition 4.3.6 (Nerf). *Soit une collection finie d'ensemble F . On définit le nerf de F comme le complexe simplicial de toute sous-collection non-vide dont les ensembles ont une intersection commune non-vide,*

$$Nrv(F) = \{X \subseteq F \mid \bigcap X \neq \emptyset\}.$$

Le nerf est toujours un complexe simplicial abstrait, et peut être réalisé géométriquement dans un espace Euclidien. Cela a donc du sens de parler de son type topologique et de son type homotopique. Le nerf permet de lier les espaces topologiques aux complexes.

Théorème 4.3.1 (Théorème du Nerf). *Soit F une collection finie d'ensembles fermés et convexes dans un espace Euclidien. Le nerf de F et l'union des ensembles dans F ont le même type homotopique.*

Si on s'inscrit dans le cadre des espaces métriques, un cas particulier d'espace topologique, on peut introduire une autre version du théorème,

Théorème 4.3.2 (Théorème du Nerf pour un espace métrique). *Étant donné un recouvrement fini C d'un espace métrique M , l'espace sous-jacent $|Nrv(C)|$ est homotopiquement équivalent à M si toute intersection non-vide $\bigcap_{i=0}^k C_{\alpha_i}$ d'éléments du recouvrement est homotopiquement équivalents à un point, i.e., contractile.*

Ainsi, à partir d'un sous-ensemble P d'un espace métrique (M, d) , on peut construire un complexe simplicial abstrait dont les vertexes dans P utilisent le concept du nerf, le complexe Čech.

4. Analyse Topologique du Signal – 4.3. Notions d'équivalence

Définition 4.3.7 (Complexe Čech). Soit (M, d) un espace métrique et P un sous-ensemble fini de M . Étant donné un réel $r > 0$, le complexe Čech $\mathbb{C}^r(P)$ est défini comme le nerf de l'ensemble $\{B(p_i, r)\}$ où

$$B(p_i, r) = \{x \in M \mid d(p_i, x) \leq r\}$$

est la boule fermée géodésique de rayon r centrée en p_i .

Dans le cas où notre espace M est euclidien, les boules que l'on utilise dans la construction du complexe Čech sont convexes et leurs intersections sont donc contractiles. Par le théorème 4.3.2, le complexe, objet combinatoire, est alors homotopiquement équivalent à l'union des boules, un espace.

Le complexe Čech est associé à un autre complexe, le complexe Vietoris-Rips.

Définition 4.3.8 (Complexe Vietoris-Rips). Soit (P, d) un espace métrique fini. Étant donné un réel $r > 0$, le complexe Vietoris-Rips est le complexe simplicial abstrait $\mathbb{V}\mathbb{R}^r(P)$ pour lequel un simplexe $\sigma \in \mathbb{V}\mathbb{R}^r(P)$ si et seulement si $d(p, q) \leq 2r$ pour toute paire de vertices de σ .

Les deux complexes sont associés car ils s'imbriquent. En effet,

Proposition 4.3.1. Soit P un sous-ensemble fini d'un espace métrique (M, d) . On a

$$\mathbb{C}^r(P) \subseteq \mathbb{V}\mathbb{R}^r(P) \subseteq \mathbb{C}^{2r}(P).$$

On trouve une preuve chez DEY et Y. WANG 2022.

Les complexes Čech ou Vietoris-Rips sont en pratique souvent trop importants pour être utilisés. On préfère des complexes plus parcimonieux. Les complexes que l'on calcule dans le Chapitre 5 sont des complexes Alpha. Ceux-ci sont une famille de sous-complexes des complexes Delaunay.

Définition 4.3.9 (Complexe de Delaunay). Soit un ensemble fini de point $P \in \mathbb{R}^d$. Un k -simplexe est un simplexe de Delaunay si ses vertexes sont dans P et aucun point de P ne se trouve dans l'hypersphère circonscrite du simplexe. Un complexe de Delaunay de P , noté $Del(P)$, est un complexe simplicial géométrique avec ses vertexes dans P et dans lequel tout simplexe est Delaunay. $|Del(P)|$ coïncide avec l'enveloppe convexe de P .

Étant donné un ensemble de point P , on construit un complexe simplicial K dessus, dont l'espace sous-jacent $|K|$ est l'enveloppe convexe de P .

Les complexes Alpha reprennent cette structure mais sont paramétrisés par un réel $\alpha \geq 0$.

Définition 4.3.10 (Complexe Alpha). Pour un ensemble de point P et $\alpha \geq 0$, un complexe alpha est constitué par tous les simplexes dans $Del(P)$ qui ont une sphère circonscrite de rayon au plus α .

Le complexe Alpha peut également être défini comme un nerf. Pour chaque point $p \in P$, on note $B(p, \alpha)$ la boule fermée de rayon α centrée en p . On définit l'espace D_p^α comme

$$D_p^\alpha = \{x \in B(p, \alpha) \mid d(x, p) \leq d(x, q), \forall q \in P\}.$$

Le complexe Alpha, que l'on note $Del^\alpha(P)$, est le nerf de l'ensemble D_p^α . $Del^\alpha(P) \subseteq \mathbb{C}^\alpha(P)$.

En utilisant le lien qui existe entre le complexe de Delaunay et le diagramme de Voronoï, le théorème du Nerf permet là encore de prouver que l'union des boules et le complexe Alpha que l'on construirait sur un nuage de points ont le même type homotopique, $\bigcup_{p \in P} B(p, \alpha) \simeq |Del^\alpha(P)|$.

Ainsi, on considère que nos données sont représentatives de la variété de laquelle elles sont échantillonnées. On couvre cette variété de simplexe pour construire un complexe simplicial dessus. On s'intéresse à la forme des données en regardant les informations topologiques du complexe simplicial.

4.4. Homologie

Avec le complexe simplicial, on a tracé la forme de nos données, posé une structure combinatoire permettant le calcul. On veut maintenant comparer rigoureusement les formes des différents complexes simpliciaux. Pour extraire une information topologique de cet espace, on va le caractériser en s'intéressant à sa connectivité, la topologie étant la branche des mathématiques qui s'intéresse à l'information géométrique qualitative d'un espace par l'étude de sa connectivité (CARLSSON 2009). Pour cela, on va compter ses trous. Derrière cette idée prosaïque se cache le concept d'homologie, qui est une méthode pour quantifier efficacement la topologie d'un espace. Plus particulièrement, on s'intéresse aux homologies simpliciales, étant donné qu'on les calcule sur des complexes simpliciaux.

L'homologie est un formalisme mathématique permettant de parler de manière quantitative et non ambiguë de la façon dont un espace est connecté. La notion d'homologie est moins forte que celle d'homéomorphisme ou d'équivalence homotopique, capture moins d'information topologique, mais comparativement à ces formalismes, les homologies peuvent être calculées rapidement. De plus, si cela est possible en dimension 2, dès la dimension 4, il n'est plus possible de déterminer si deux variétés triangulées sont homéomorphiques ou homotopiquement équivalentes, alors qu'il reste possible de s'intéresser à leurs homologies. En cela, l'homologie permet d'avoir une information topologique quantitative rapidement. Elle permet de définir et classer les trous d'un espace de manière rigoureuse et ces trous permettent de distinguer et classer différents espaces.

4.4.1. Chaînes, cycles et bords

Pour définir les groupes homologiques, les trous de différentes dimensions de l'espace que l'on étudie, nous avons besoin d'abord de définir ce que seront les

chemins et les boucles. En effet, les groupes homologiques capturent les trous d'un espace topologique en s'intéressant à ce qui les entoure, à la connectivité de cet espace. On s'appuie pour cela sur un certain formalisme, qui commence par la définition des p -chaînes.

Définition 4.4.1 (p -chaîne). *Soit K un complexe simplicial et p une dimension. Une p -chaîne est une somme de p -simplexes de K . On note*

$$c = \sum a_i \sigma_i,$$

où les σ_i sont les p -simplexes et a_i les coefficients. Deux p -chaînes peuvent être ajoutées pour obtenir une troisième p -chaîne.

Une p -chaîne c est donc la somme de simplexes de dimension p d'un complexe K . L'ensemble des p -chaînes, muni de la loi d'addition, forme le groupe des p -chaînes, $C_p(K)$. Un complexe simplicial a un groupe de p -chaînes pour chaque entier p . Pour p plus petit que zéro ou plus grand que la dimension de K , C_p est constitué uniquement de l'élément neutre 0.

Comme nous le disions, l'homologie permet d'étudier la connectivité entre deux dimensions. On doit donc définir une structure permettant de relier les groupes de différentes dimensions. On fait ce lien avec l'opérateur de bord, un homomorphisme (une application entre groupes qui commute pour l'opération de composition) qui relie les groupes de p -chaînes des différentes dimensions.

Définition 4.4.2 (Opérateur de bord). *Soit un complexe simplicial K et $\sigma \in K, \sigma = \{v_0, \dots, v_p\}$. L'opérateur ou homomorphisme de bord $\partial_p : C_p(K) \rightarrow C_{p-1}(K)$ est*

$$\partial_p \sigma = \sum_i (-1)^i \{v_0, \dots, \hat{v}_i, \dots, v_n\},$$

où \hat{v}_i indique que v_i est omis.

Appliquer l'opérateur de bord sur une p -chaîne produit donc une $(p-1)$ -chaîne, et on écrit $\partial_p : C_p \rightarrow C_{p-1}$. Si on applique, pour une p -chaîne $c = \sum a_i \sigma_i$, qui est donc la somme de plusieurs simplexes, l'opérateur de bord ∂_p , on obtient le bord de la chaîne, qui est la somme des bords de ses simplexes, $\partial_p c = \sum a_i \partial_p \sigma_i$. Le résultat est lui-même une chaîne, mais de dimension $p-1$.

Cela permet de définir la séquence de groupes abéliens connectés par les opérateurs de bord, que l'on appelle un complexe de chaînes,

$$\dots \xrightarrow{\partial_{p+2}} C_{p+1} \xrightarrow{\partial_{p+1}} C_p \xrightarrow{\partial_p} C_{p-1} \xrightarrow{\partial_{p-1}} \dots$$

Les chaînes que l'on vient de définir peuvent être divisées en deux types, et cette division va nous permettre de définir les groupes homologiques. Les deux types de chaînes sont les cycles et les bords.

Définition 4.4.3 (p -cycle). *Un p -cycle est une p -chaîne avec un bord vide, $\partial c = 0$. Le groupe des p -cycles, que l'on note $Z_p = Z_p(K)$ est un sous-groupe du groupe des p -chaînes C_p . Le groupe des p -cycles est le noyau du $p^{\text{ième}}$ opérateur de bord, $Z_p = \ker \partial_p$.*

Définition 4.4.4 (*p*-bord). *Un p-bord est une p-chaîne qui est le bord d'une (p + 1)-chaîne, $c = \partial d$, avec $d \in C_{p+1}$. Le groupe des p-bords, que l'on note $B_p = B_p(K)$, est un sous-groupe du groupe des p-cycles Z_p . Le groupe des p-bords est l'image du (p + 1)^{ième} opérateur de bord, $B_p = \text{im } \partial_{p+1}$.*

Ainsi, un élément du noyau de ∂_p est un cycle, un élément de son image est un bord. Il convient d'ajouter la propriété permettant à l'homologie de fonctionner,

Proposition 4.4.1. $\partial_p \partial_{p+1} d = 0$ pour tout entier *p* et toute (p + 1)-chaîne *d*.

Le bord d'un bord est nécessairement zéro. La figure 4.1 illustre la connexion entre les groupes et les différentes dimensions via homomorphisme.

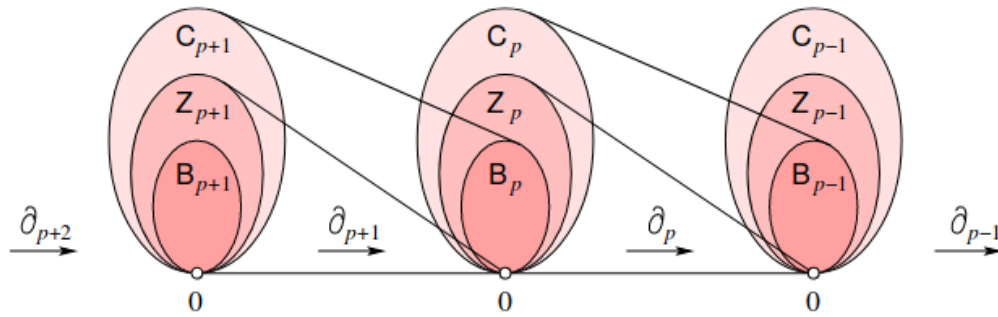


FIGURE 4.1. – Complexe de chaînes d'un complexe, consistant en une séquence linéaire de groupes de chaînes, de cycles et de bords, connectés par des homomorphismes, l'opérateur de bord. Les groupes des *p*-bords sont des sous-groupes des groupes des *p*-cycles, qui sont eux-mêmes des sous-groupes des *p*-chaînes. Schéma tiré de H. EDELSBRUNNER et HARER 2009.

4.4.2. Groupe d'homologie

La définition des groupes des *p*-cycles et des *p*-bords permet de définir les groupes d'homologie. Tout *p*-bord est un *p*-cycle, et donc B_p est un sous-groupe de Z_p (qui est lui-même un sous-groupe de C_p). Comme les bords forment un sous-groupe du groupe des cycles, on peut en prendre le quotient. Cela revient à faire une partition du groupe des cycles en différentes classes de cycles, selon leurs bords.

Définition 4.4.5 (Groupe d'homologie). *Le $p^{\text{ième}}$ groupe d'homologie est*

$$H_p = \frac{Z_p}{B_p} = \frac{\ker \partial_p}{\text{im } \partial_{p+1}}$$

Chaque élément de H_p est obtenu en ajoutant toutes les *p*-bords à un *p*-cycle $c \in Z_p$ donné, $c + B_p$. Les éléments de H_p sont les classes d'homologie du complexe K . Tout cycle dans la même classe d'homologie sont dits homologues.

Le groupe H_p est un espace vectoriel dont sa dimension est égale à son rang. On appelle celui-ci le $p^{\text{ième}}$ nombre de Betti, que l'on note β_p .

Définition 4.4.6 (Nombre de Betti). *Le $p^{\text{ième}}$ nombre de Betti β_p d'un complexe simplicial K est $\beta(H_p)$,*

$$\beta_p = \text{rank } H_p = \text{dim } H_p$$

Les nombres de Betti capturent ainsi le nombre d'éléments de chaque groupe. De manière informelle, le p -nombre de Betti renvoie au nombre de surfaces indépendantes de dimension p , *i.e.*, le nombre de trous de chaque dimension du complexe simplicial que l'on a construit sur nos données pour trianguler la variété sous-jacente. β_0 mesure le nombre de composantes connexes (le nombre de morceaux), β_1 le nombre de boucles, β_2 le nombre de cavités.

On a ainsi un complexe simplicial constitué de simplexes. En sommant les simplexes, on a une chaîne. Deux types de chaînes existent, les cycles et les bords. Un cycle est une boucle fermée sur la variété d'intérêt. Le point d'arrivée et le point de départ sont les mêmes. C'est une coupe de la variété qui produit une sous-variété fermée, de différentes dimensions possibles. Un bord est lui-même un cycle qui borde un cycle de plus grande dimension. En cela, les bords sont des cycles pleins. Enfin, le rapport des deux nous permet de définir les homologies. Une classe d'homologie est une classe d'équivalence de cycles modulo un bord. On divise le groupe des cycles en différents groupes. Les cycles de chacun de ces groupes sont liés par une classe d'équivalence, en fonction de leur bord. Une classe d'homologie est représentée par un cycle, mais vide, qui n'est le bord d'aucune sous-variété. Le cycle est dans ce cas un trou. On a ainsi un formalisme permettant de calculer rigoureusement et algorithmiquement efficacement les trous d'un espace, et donc une façon de caractériser ledit espace.

4.5. Persistence

4.5.1. Topologie et échelle

On a le formalisme mathématique permettant de nous intéresser à la forme de X . Comme on l'a vu, on triangule la variété sous-jacente de X en construisant un complexe simplicial dessus. Les homologies simpliciales sont des invariants que l'on peut calculer efficacement permettant de caractériser cet espace topologique. Le théorème du nerf 4.3.2 nous garantit par équivalence homotopique que l'on calcule l'homologie de la variété sous-jacente en calculant celle du complexe. Néanmoins, X est en pratique un ensemble de données discrètes. Les données que l'on étudie sont justement des échantillons finis et bruités d'une variété sous-jacente inconnue, posant donc la question de quel complexe simplicial construire.

En effet, comme X est bruité, une multitude de complexes simpliciaux pourrait être construits dessus, selon l'échelle que l'on choisit. Il s'ensuit des classes d'homologie présentes à certaines échelles mais absentes à d'autres, certaines porteuses d'information et d'autres non, porteuses de bruit topologique.

Si l'on reprend les complexes simpliciaux présentés dans la partie 4.3.3, ceux-ci se construisent en prenant une certaine valeur d'échelle. Le complexe Alpha par exemple est constitué par les simplexes du complexe de Delaunay qui ont une sphère circonscrite de rayon au plus α , $\alpha \geq 0$. Le complexe Vietoris-Rips prend également un paramètre réel $r > 0$. Comment choisir la bonne valeur permettant d'approximer au mieux la variété sous-jacente à X à l'aide d'un unique complexe simplicial?

La notion de persistance topologique, introduit chez BARANNIKOV 1994; EDELSBRUNNER, LETSCHER et ZOMORODIAN 2002, permet de ne pas choisir. En effet, au lieu de construire un complexe simplicial en sélectionnant une valeur d'échelle par une procédure nécessairement instable du fait des données bruitées et discrètes, sur lequel on calculerait ensuite l'homologie, on construit plutôt une séquence emboîtée de complexes simpliciaux. C'est le principe de filtration. En cela, la TDA permet la construction de descripteur topologique multi-échelle.

4.5.2. Filtration

On calcule la persistance d'un complexe simplicial K selon une filtration de ce complexe. On choisit un certain paramètre d'échelle pour décomposer le complexe simplicial construit sur les données en sous-niveaux. L'idée de la filtration est de ne pas construire un complexe simplicial, mais une séquence emboîtée de complexes simpliciaux.

Définition 4.5.1 (Filtration). *Soit un complexe simplicial K , composé de m simplexes. On introduit une fonction f monotone définie sur les simplexes de K , $f : K \rightarrow \mathbb{R}$. Comme f est monotone, l'ensemble de sous-niveaux $K(r) = f^{-1}(-\infty, r]$ est un sous-complexe de K pour tout $r \in \mathbb{R}$ et $K^{r_1} \subseteq K^{r_2}$ pour $r_1 \leq r_2$. Pour une suite $r_1 \leq r_2 \leq \dots \leq r_n$, on obtient une séquence emboîtée de $n + 1 \leq m + 1$ différents sous-complexes, connectés par inclusion, que l'on arrange dans une séquence croissante*

$$\mathcal{F}_f : \emptyset = K^0 \subseteq K^1 \subseteq \dots \subseteq K^n = K.$$

\mathcal{F}_f est une filtration. On appelle un complexe avec une filtration un complexe filtré.

On appelle cette séquence de complexe une filtration de f . Le paramètre $r \in \mathbb{R}$ est le paramètre de filtrage. Cela revient à ajouter des morceaux de simplexes à chaque étape de la filtration. On construit à chaque étape un complexe simplicial pour la valeur du paramètre de filtrage donné.

À partir du complexe filtré, la séquence de complexes, on peut s'intéresser à son évolution topologique, par la séquence des groupes homologiques de chaque élément de la séquence de complexes. On a une application d'inclusion pour $i \leq j$ de l'espace sous-jacent de K^i à l'espace sous-jacent de K^j . Du fait de la monotonie de f , on a un homomorphisme $f_p^{i,j} : H_p(K^i) \rightarrow H_p(K^j)$, pour chaque dimension p (H. EDELSBRUNNER et HARER 2009). Toute filtration simpliciale donne donc lieu à une

séquence de groupes homologiques connectés par des homomorphismes

$$0 = H_p(K^0) \rightarrow H_p(K^1) \rightarrow \dots \rightarrow H_p(K^i) \xrightarrow{f_p^{i,j}} H_p(K^j) \rightarrow \dots \rightarrow H_p(K^n) = H_p(K).$$

Plus que la séquence de complexes simpliciaux, c'est cette séquence de groupes homologiques qui nous intéresse, car elle nous permet de tracer l'évolution de caractéristiques topologiques de l'espace sous-jacent de K . De K^{i-1} à K^i , de nouvelles classes d'homologie apparaissent et d'anciennes disparaissent. On résume cela avec la persistance.

4.5.3. Homologies persistantes

Pour chaque élément de \mathcal{F}_f , on calcule H_p , pour tout p . On calcule donc l'homologie de chaque complexe simplicial de la séquence.

Définition 4.5.2 (Homologie d'une filtration). *Soit \mathcal{F}_f une filtration de K et K^ℓ le complexe filtré correspondant. Soit $Z_p^\ell = Z_p(K^\ell)$ et $B_p^\ell = B_p(K^\ell)$ respectivement les groupes des $p^{\text{ième}}$ cycles et bords de K^ℓ . Le $p^{\text{ième}}$ groupe d'homologie de K^ℓ est*

$$H_p^\ell = \frac{Z_p^\ell}{B_p^\ell}.$$

Le $p^{\text{ième}}$ nombre de Betti de K^ℓ est le rang de H_p^ℓ .

Le $p^{\text{ième}}$ nombre de Betti décrit alors la topologie d'un complexe simplicial croissant par une séquence d'entier. Si on espère que cette séquence contienne une information topologique sur l'espace initial, elle contient aussi de nombreux autres attributs topologiques capturés par l'homologie. Pour autant, ces attributs ne sont pas tous informatifs. Il est difficile de distinguer ceux qui sont des caractéristiques topologiques de l'espace original et ce qui sont du bruit topologique, conséquence de la filtration. Au lieu de prendre l'intégralité de l'homologie d'une filtration, on va s'intéresser aux groupes d'homologie persistante. La persistance est la mesure permettant de distinguer ce qui relève de caractéristique topologique de ce qui relève du bruit topologique.

L'idée de la persistance est qu'un attribut topologique significatif aura une durée de vie longue dans la filtration. L'attribut en question persiste pour caractériser un complexe qui croît. Comme la filtration est une séquence de complexes simpliciaux et que l'homologie capture les classes d'équivalence des cycles modulo les limites pour chaque complexe, on va capturer les cycles qui ne sont pas dans B_p et qui ne rentrent pas dans les j complexes qui suivent. Ces cycles persistent alors j étapes. Ils sont persistants.

Définition 4.5.3 (Groupe d'homologie persistante). *Soit \mathcal{F}_f une filtration. Le $p^{\text{ième}}$*

groupe d'homologie j -persistante de la filtration est

$$H_p^{i,j} = \frac{Z_p^i}{B_p^{i+j} \cap Z_p^i},$$

où Z_p^i est le groupe des $p^{\text{ième}}$ cycles de K^i , complexe dans la séquence du complexe filtré, et B_p^{i+j} est le groupe des $p^{\text{ième}}$ bords j complexes plus tard. Le $p^{\text{ième}}$ nombre de Betti persistant correspondant est le rang du groupe,

$$\beta_p^{i,j} = \text{rank } H_p^{i,j}.$$

Ce groupe est là encore bien défini car $B_p^{i+j} \cap Z_p^i$ est l'intersection de deux sous-groupes de C_p^{i+j} . Selon la valeur que l'on choisit pour j , on s'intéressera à des attributs topologiques plus ou moins persistants. Pour des valeurs faibles, on trouvera le bruit topologique du complexe. Plus celui-ci augmente, plus on aura à faire à des éléments qui caractérisent le complexe.

On peut également définir l'homologie persistante à partir des homomorphismes induits par la relation d'inclusion de la filtration $f_p^{i,j} : H_p(K^i) \rightarrow H_p(K^{i+j})$. Les $p^{\text{ième}}$ groupes d'homologie persistante en sont les images

$$H_p^{i,j} = \text{im } f_p^{i,j},$$

pour $0 \leq i \leq j \leq n$.

Au fil de la croissance du complexe simplicial par l'ajout de simplexe, de nouvelles classes d'homologie apparaissent, d'autres disparaissent en devenant triviale ou en fusionnant avec d'autres. Toutes ces classes, nées avant un seuil et mortes à un autre, sont regroupées dans les groupes d'homologie persistante. Un groupe d'homologie persistante $H_p^{i,j}$ est constitué par les classes d'homologies de K^i toujours vivante à K^j . Ces groupes existent pour chaque dimension p et chaque paire $i \leq j$. La durée de vie des groupes d'homologie persistante, au fil de la croissance du complexe simplicial par l'ajout de simplexe, permet de déterminer les attributs topologiques les plus persistants.

Définition 4.5.4 (Persistence). Soit γ une classe dans $H_p(K^i)$. On dit que γ est née à K^i si $\gamma \notin H_p^{i-1,i}$. γ meurt à K^j s'il est fusionné avec une classe plus ancienne du passage de K^{j-1} à K^j , i.e., $f_p^{i,j-1}(\gamma) \notin H_p^{i-1,j-1}$ mais $f_p^{i,j}(\gamma) \in H_p^{i-1,j}$. Si γ est né K^i et meurt à K^j , on appelle cette différence la persistance

$$\text{pers}(\gamma) = a_j - a_i.$$

On peut prendre la différence en indice $j - i$ et on a alors l'indice de persistance de la classe. Si γ est né à K^i et ne meurt jamais, alors sa persistance et son indice de persistance sont ∞ .

Si au fil de la filtration, un p -cycle vide (*i.e.*, sans bord, pas dans B_p) γ est créé au temps i par le simplexe σ , on dit que σ est le créateur de γ . Si γ est transformé en bord au moment j par le simplexe τ , on dit que τ détruit γ .

4.6. Diagramme de persistance

Une fois le calcul de l'homologie persistante de la filtration faite, on a l'information homologique de X calculée pour différentes échelles, pour chaque complexe simplicial de la séquence emboîtée pour une valeur croissante de r . On a la collection des classes de persistance, chacune représentée par une paire (a_j, a_i) , le moment de sa naissance et de sa mort. On enregistre et résume cela dans le diagramme de persistance, pour la valeur de r .

4.6.1. Définition d'un diagramme de persistance

Définition 4.6.1 (Diagramme de persistance). *Pour une filtration \mathcal{F}_f , soit $\mu_p^{i,j}$ le nombre de classes d'homologies de dimension p nées à K^i et mortes à K^j ,*

$$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j})$$

pour tout $i < j$ et tout p . Le diagramme de $p^{\text{ième}}$ persistance de la filtration \mathcal{F}_f est un multi-ensemble de points dans le plan réel étendu \mathbb{R}^2 , où chaque point (a_j, a_i) correspond à un couple appareillé de classes d'homologies, avec sa multiplicité $\mu_i^{i,j}$. Les points sur la diagonale $\Delta : \{(x, x)\}$ sont ajoutés au diagramme, chacun avec multiplicité infinie. On le note $\text{Dgm}_p(f)$.

La première différence dans la formule de $\mu_p^{i,j}$ compte le nombre de classes nées avant ou à K^i et qui meurent à K^j , alors que la seconde différence compte le nombre de classes nées avant ou à K^{i-1} et qui meurent à K^j . La différence des différences permet bien de compter le nombre de classes nées à K^i et mortes à K^j . Comme $\mu_p^{i,j}$ est défini pour $i < j$, tous les points sont au-dessus de la diagonale. L'ajout de la diagonale se fait pour des raisons techniques.

Certains points peuvent être confondus, d'autres peuvent avoir des coordonnées infinies. Ces derniers sont des points de persistance essentiels, correspondant à des classes d'homologie essentielles. Chaque classe γ ayant une persistance $\text{pers}(\gamma)$ est représentée par un point sur le diagramme de persistance à une distance euclidienne $\frac{\text{pers}(\gamma)}{\sqrt{2}}$ de la diagonale Δ .

Enfin, une importante propriété relie les nombres de Betti persistants et les diagrammes de persistance,

Proposition 4.6.1. *Soit une filtration \mathcal{F}_f . Pour toute paire d'indices $0 \leq k \leq \ell \leq n$ et*

toute dimension p , le $p^{\text{ième}}$ nombre de Betti persistant est

$$\beta_p^{k,\ell} = \sum_{i \leq k} \sum_{j > \ell} \mu_p^{i,j}.$$

On peut ainsi retrouver le nombre de Betti de chaque complexe de la séquence de la filtration à partir du diagramme de persistance. Le diagramme de persistance permet d'encoder tout l'historique de la filtration, encode toute l'information sur l'homologie à toutes les échelles et son évolution, d'où l'intérêt de son utilisation.

4.6.2. Stabilité d'un diagramme de persistance

De plus, il s'avère que les diagrammes de persistance sont stables à des perturbations de faible amplitude, ce qui motive d'autant plus leur utilisation pour l'analyse de données réelles et bruitées. On entend par stabilité le fait que si l'entrée est légèrement perturbée, la sortie ne change pas beaucoup. Il ne serait pas utile en pratique d'avoir une information qui change très fortement pour toute légère perturbation.

Pour pouvoir formuler la stabilité des diagrammes de persistance, on a d'abord besoin d'une notion de distance entre diagrammes. L'espace des diagrammes de persistance est équipé d'une métrique, la distance bottleneck.

Définition 4.6.2 (Distance bottleneck). Soit $\Pi = \{\pi\}$ l'ensemble de toutes les bijections $\pi : Dgm_p(\mathcal{F}_f) \rightarrow Dgm_p(\mathcal{F}_g)$. La distance bottleneck entre deux diagrammes est

$$d_b(Dgm_p(\mathcal{F}_f), Dgm_p(\mathcal{F}_g)) = \inf_{\pi \in \Pi} \sup_{x \in Dgm_p(\mathcal{F}_f)} \|x - \pi(x)\|_\infty.$$

On retrouve le premier résultat de stabilité chez COHEN-STEINER, H. EDELSBRUNNER et HARER 2007, qui permet de quantifier la notion de stabilité d'un diagramme de persistance,

Théorème 4.6.1. Soit X un espace triangulable et deux fonctions continues tame $f, g : X \rightarrow \mathbb{R}$. Les diagrammes de persistance satisfont

$$d_b(Dgm_p(\mathcal{F}_f), Dgm_p(\mathcal{F}_g)) \leq \|f - g\|_\infty.$$

Le résultat tient pour X triangulable et avec f et g continues et tame, i.e., qu'elles n'ont qu'un nombre fini de valeurs critiques, et définies sur le même espace topologique X . Frédéric CHAZAL, COHEN-STEINER, GLISSE et al. 2009 généralisent le résultat, en utilisant le concept de module de persistance. Les résultats de stabilité n'ont plus besoin des conditions de continuité et de triangulabilité, le *tameness* est relaxé, les fonctions peuvent être définies sur différents espaces topologiques. Ce formalisme permet donc de ré-établir le résultat précédent dans des conditions plus générales, ainsi que d'autres résultats de stabilité, par exemple pour d'autres distances comme la Gromov-Hausdorff (Frederic CHAZAL, Vin DE SILVA, GLISSE et al. 2013).

4.6.3. Consistance d'un diagramme de persistance

Nous avons ainsi vu le formalisme permettant de nous intéresser à la forme de nos données. On pose une structure combinatoire, un complexe simplicial, pour lequel on calcule l'homologie persistante. On résume cette information dans le diagramme de persistance, qui encode toute l'information sur les groupes d'homologie persistante, et qui présente de bons résultats de stabilité. Cela en fait donc un bon outil pour décrire topologiquement nos données. Néanmoins, le travail statistique à partir de cet objet n'est pas trivial.

En effet, si on considère que l'homologie persistante est une bonne façon d'étudier la forme des données, elle ne prend néanmoins pas en compte leur nature aléatoire. Pour autant, plusieurs résultats de consistances ont été prouvés (MICHEL 2015; Frédéric CHAZAL et MICHEL 2021).

En pratique, on considère nos échantillons (x_1, \dots, x_n) comme vivant dans un espace métrique (\mathcal{X}, ρ) et échantillonnées à partir d'une mesure de probabilité inconnue \mathbb{P} dont le support est un espace compact noté $\mathcal{M}_{\mathbb{P}}$. On fait en particulier l'hypothèse que le support a une forme géométrique avec l'hypothèse de la variété.

Pour en apprendre plus sur $\mathcal{M}_{\mathbb{P}}$, on voudrait construire une filtration de celui-ci, $\mathcal{F}(\mathcal{M}_{\mathbb{P}})$, pour laquelle on calculerait l'homologie. N'ayant pas accès à $\mathcal{M}_{\mathbb{P}}$, on construit une estimation de celui-ci à partir de nos échantillons, $\hat{\mathcal{M}}_n = (x_1, \dots, x_n)$. Aussi, notre filtration $\mathcal{F}(\hat{\mathcal{M}}_n)$ est définie sur $\hat{\mathcal{M}}_n$, et donc l'homologie persistante est celle de $\hat{\mathcal{M}}_n$. Frédéric CHAZAL, GLISSE, LABRUÈRE et al. 2014 montrent que $\hat{\mathcal{M}}_n$ a un taux optimal de convergence vers $\mathcal{M}_{\mathbb{P}}$ par rapport à la distance Hausdorff.

Théorème 4.6.2. *Soit (M, ρ) un espace métrique et soit $a, b > 0$. On a*

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E} \left[d_b(\text{Dgm}(\mathcal{F}(\mathcal{M}_{\mathbb{P}})), \text{Dgm}(\mathcal{F}(\hat{\mathcal{M}}_n))) \right] \leq C \left(\frac{\log n}{n} \right)^{1/b}.$$

C ne dépend que de a et b, pas de M. Si on fait de plus l'hypothèse qu'il existe un point non-isolé x dans M et on considère une séquence $(x_n) \in (M \setminus \{x\})^{\mathbb{N}}$ tel que $\rho(x, x_n) \leq (an)^{-1/b}$ alors, pour tout estimateur $\hat{\text{Dgm}}_n$ de $\text{Dgm}(\mathcal{F}(\mathbb{X}_{\mu}))$,

$$\liminf_{n \rightarrow \infty} \rho(x, x_n)^{-1} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E} \left[d_b(\text{Dgm}(\mathcal{F}(\mathcal{M}_{\mathbb{P}})), \hat{\text{Dgm}}_n) \right] \geq C',$$

où C' est une constante absolue.

Ainsi, sous l'hypothèse (a, b) -standard, i.e., pour $a, b > 0$, pour $x \in \mathcal{M}_{\mathbb{P}}$ et $r > 0$,

$$\mathbb{P}(B(x, r)) \geq \min(ar^b, 1),$$

le taux de convergence de $\text{Dgm}(\mathcal{F}(\hat{\mathcal{M}}_n))$ à $\text{Dgm}(\mathcal{F}(\mathcal{M}_{\mathbb{P}}))$ est borné supérieurement pour la métrique bottleneck par $C \left(\frac{\log n}{n} \right)^{1/b}$. $\text{Dgm}(\mathcal{F}(\hat{\mathcal{M}}_n))$ est un estimateur optimal au sens minimax sur l'espace métrique (M, ρ) satisfaisant l'hypothèse (a, b) de $\text{Dgm}(\mathcal{F}(\mathcal{M}_{\mathbb{P}}))$.

4. Analyse Topologique du Signal – 4.6. Diagramme de persistance

On a ainsi une assise théorique assurant la convergence du diagramme de persistance d'un échantillon aléatoire d'un espace métrique vers le diagramme de persistance encodant les caractéristiques topologiques du support de la mesure à partir de laquelle les données ont été échantillonnées. D'autres résultats ont été démontrés, notamment pour l'analyse statistique de l'homologie persistante d'un diagramme, comme B. T. FASY, LECCI, RINALDO et al. 2014, qui permettent le calcul de régions de confiance basées sur la distance bottleneck pour aider à distinguer le bruit des caractéristiques topologiques. Frédéric CHAZAL, B. FASY, LECCI et al. 2018 proposent une alternative en utilisant du bootstrap.

Plus que cela, des résultats permettent de dériver des statistiques à partir d'un échantillon de diagrammes de persistance. Si l'espace des diagrammes de persistance est un espace métrique, dans lequel on peut calculer la distance entre deux objets, ce n'est pas un espace de Hilbert. La définition d'un diagramme de persistance moyen n'est par exemple pas unique. En cela, faire un travail statistique à partir d'un échantillon de ces objets peut se révéler complexe. Différentes stratégies existent afin de vectoriser les diagrammes, en sortir l'information qu'ils contiennent pour l'utiliser dans un travail d'apprentissage statistique. BUBENIK 2015 ; BUBENIK 2020 propose une stratégie pour représenter un diagramme de persistance dans un espace de Hilbert via le paysage persistant.

Définition 4.6.3 (Paysage persistant). *Soit un diagramme de persistance Dgm constitué de paires de points $\{(b_i, d_i)\}_{i \in I}$, les paires de naissance et de mort résumées dans le diagramme de persistance. Les diagrammes ont un nombre fini de points (b, d) avec $-\infty < b < d < \infty$. On définit*

$$f_{(b,d)}(t) = \max(0, \min(b + t, b - t)),$$

ce qui permet de définir le paysage persistant comme

$$\lambda(k, t) = k \max\{f_{(b_i, d_i)}(t)\}_{i \in I}$$

où $k \max$ note le k plus grand élément.

Le paysage persistant est une collection de fonctions continues linéaires par morceaux $\lambda_1, \lambda_2 \dots : \mathbb{R} \rightarrow \mathbb{R}$ où λ_k est la $k^{\text{ième}}$ fonction de paysage persistant. BUBENIK 2015 prouve que le paysage persistant présente lui aussi des propriétés de stabilité. Surtout, cette représentation étant dans un espace de Hilbert, il devient alors possible d'en calculer des statistiques.

Pour un échantillon de paysage persistant tiré d'une distribution aléatoire de l'espace des paysages, on peut calculer le paysage moyen. $\mathbb{E}[\lambda_X]$ a des informations topologiques stables sur la mesure sous-jacente \mathbb{P} , de laquelle les données sont générées, et le paysage persistant moyen empirique en est un estimateur non-biaisé (Frédéric CHAZAL, B. FASY, LECCI et al. 2015). On a de plus des résultats de normalité asymptotique et une convergence uniforme du multiplicateur *bootstrap* permettant le calcul de régions de confiance (Frédéric CHAZAL, B. T. FASY, LECCI et al. 2014). Ces

4. Analyse Topologique du Signal – 4.6. Diagramme de persistance

résultats tiennent pour d'autres représentations fonctionnelles des diagrammes de persistance, comme la silhouette (Frédéric CHAZAL, B. T. FASY, LECCI et al. 2014).

On peut donc faire un travail statistique sur un échantillon de données à partir de descripteurs topologiques extraits des diagrammes de persistance, via les paysages persistants ou les silhouettes par exemple. D'autres stratégies existent. Certains calculent des statistiques pour résumer le diagramme : COHEN-STEINER, H. EDELSBRUNNER, HARER et MILEYKO 2010 calculent la norme p du diagramme de persistance, ATIENZA, GONZALEZ-DIAZ et RUCCO 2019 l'entropie persistante, qu'ils prouvent stable et invariante (ATIENZA, GONZALEZ-DÍAZ et SORIANO-TRIGUEROS 2020), CARRIÈRE, OUDOT et OVSJANIKOV 2015 un vecteur stable servant de signature topologique à partir de la distance des points de la diagonale, FIREAIZEN, RON et BOBROWSKI 2022 un ensemble de statistiques. MAROULAS, MIKE et OBALLE 2019; MAROULAS, NASRIN et OBALLE 2020; MAROULAS, MICUCCI et NASRIN 2022 travaillent directement dans l'espace des diagrammes de persistance en utilisant la théorie des processus ponctuels. Ils suivent en cela une intuition que l'on trouve déjà en ouverture chez CARLSSON 2009, en y ajoutant une couche de modélisation bayésienne. En plus des approches fonctionnelles déjà citées comme le paysage persistant de BUBENIK 2015 ou la silhouette, qui est un paysage persistant modifié, de Frédéric CHAZAL, B. T. FASY, LECCI et al. 2014, d'autres méthodes existent pour vectoriser le diagramme de persistance : H. ADAMS, EMERSON, KIRBY et al. 2017 utilisent l'image persistante pour représenter le diagramme. D'autres objets existent, comme les vignobles, permettant de suivre l'évolution topologique d'un système dans le temps (BERGOMI et BARATÈ 2020; SALCH, REGALSKI, ABDALLAH et al. 2021). Enfin, d'autres procédures passent plutôt par l'intégration de la topologie dans le modèle. Une partie continue à s'appuyer sur les diagrammes de persistance : J.-Y. LIU, JENG et Y.-H. YANG 2016 utilisent des paysages persistants en entrée d'un CNN, K. KIM, ZAHEER, Frederic CHAZAL et al. 2020 proposent PLLay, une couche topologique basée sur les paysages persistants, quand DE SURREL, HENSEL, CARRIÈRE et al. 2022 calculent le diagramme de persistance directement dans le réseau. D'autres intègrent plutôt un paramètre permettant de prendre en compte la topologie des données dans la perte de leur modèle, comme MOOR, HORN, RIECK et al. 2020 avec l'autoencodeur topologique ou TROFIMOV, CHERNIAVSKII, TULCHINSKII et al. 2023, qui proposent un autre autoencodeur permettant d'apprendre une représentation des données respectant sa topologie, basée sur l'utilisation d'une perte topologique, de BARANNIKOV, TROFIMOV, BALABIN et al. 2022. Des revues listent les différentes stratégies pour exploiter l'information contenue dans les diagrammes de persistance (BARNES, POLANCO et PEREA 2021; HENSEL, MOOR et RIECK 2021), sans être conclusive sur la supériorité d'une méthode. Il semble plutôt que la stratégie à adopter dépende du problème et des données. Un des résultats de l'étude 5 consiste à comparer les différentes stratégies pour une tâche de classification de signal sonore.

On vient de présenter le formalisme et la théorie qui sous-tendent la TDA. On s'appuie sur ces outils et résultats dans le Chapitre 5 suivant. On ré-utilise une partie de ces outils également dans la dernière contribution, présentée dans le Chapitre 7.

5. Topological data analysis of human vowels : Persistent homologies across representation spaces

Sommaire

5.1. Introduction	104
5.2. The problem	104
5.2.1. Motivations for TDA	104
5.2.2. Representation space	105
5.2.3. Same signal, different representations, different topological characteristics	107
5.3. Topological Data Analysis : An overview	108
5.3.1. Persistent homology	109
5.3.2. Exploitation of the information from the diagram space	110
5.4. Experiments	111
5.4.1. Presentation of the data	111
5.4.2. Comparison on three supervised classification tasks	112
5.4.3. Computation of the homological information	112
5.4.4. Extraction of the information from the persistence diagram and comparison of the persistent variables	112
5.4.4.1. Persistent variables	113
5.4.4.2. Functional summary of the persistence diagrams	115
5.4.5. Step-wise selection of the variables	116
5.5. Results	117
5.5.1. Supervised task	117
5.5.1.1. TDA is useful for signal classification	117
5.5.1.2. Topological information alone is not enough to discriminate the signal	117
5.5.1.3. Complementarity of representations	119
5.5.2. Best topological variables	121
5.5.3. Unsupervised analysis, learning the manifold	121
5.6. Discussion	123
5.6.1. On the improvements on the prediction of labels	123

5. *Topological data analysis of human vowels : Persistent homologies across representation spaces –*

5.6.2. Different objects, different topologies	123
5.6.2.1. The difference between representation spaces	123
5.6.2.2. A different perspective	124
5.6.3. On the more present persistent variables	124
5.7. Conclusion	125

Ce Chapitre est la deuxième contribution de la thèse. Il a donné lieu à l'écriture d'un article, avec le co-autorat de Jean-Marc Freyermuth, Pierre Pudlo, Samuel Tronçon et Arnaud Rey. Il est sur le point d'être prépublié sur ArXiv et soumis à la revue *IEEE Transactions on Signal Processing*. Il a été présenté aux 54^{èmes} Journées de Statistique de la SFdS à Bruxelles. La base de données que nous avons récoltée pour ce travail est librement accessible sur <https://zenodo.org/record/7961904>.

Abstract

Topological Data Analysis (TDA) has been successfully used for various tasks in signal/image processing, from visualization to supervised/unsupervised classification. Often, topological characteristics are obtained from persistent homology theory. The standard TDA pipeline starts from the raw signal data or a representation of it. Then, it consists in building a multiscale topological structure on the top of the data using a pre-specified filtration, and finally to compute the topological signature to be further exploited. The commonly used topological signature is a persistent diagram (or transformations of it). Current research discusses the consequences of the many ways to exploit topological signatures, much less often the choice of the filtration, but to the best of our knowledge, the choice of the representation of a signal has not been the subject of any study yet. This paper attempts to provide some answers on the latter problem. To this end, we collected real audio data and built a comparative study to assess the quality of the discriminant information of the topological signatures extracted from three different representation spaces. Each audio signal is represented as i) an embedding of observed data in a higher dimensional space using Taken's representation, ii) a spectrogram viewed as a surface in a 3D ambient space, iii) the set of spectrogram's zeroes. From vowel audio recordings, we use topological signature for three prediction tasks : speaker sex, vowel type, and individual. We show that topologically-augmented random forest improves the Out-of-Bag Error (OOB) over solely based Mel-Frequency Cepstral Coefficients (MFCC) for the last two tasks. Our results also suggest that the topological information extracted from different signal representations is complementary, and that spectrogram's zeros offers the best improvement for sex prediction.

Keywords

TDA, topologically-augmented machine learning, persistent homology, representation space, signal classification, human vowel.

5.1. Introduction

Topological Data Analysis (TDA) is a fast-growing research area that relies on deep mathematical foundations (CARLSSON 2009; WASSERMAN 2018; Frédéric CHAZAL et MICHEL 2021). It offers novel and potentially fruitful angles of analysis of digital audio signals. This innovative approach to data science is based on extracting information from the *shape of data*.

TDA has already been applied to various signal processing tasks (BARBAROSSA et SARDELLITTI 2020a; BARBAROSSA et SARDELLITTI 2020b). It starts from the assumption that the data have a shape (FERRI 2018) and it computes its persistent homologies, which provide a compact representation of its topological features. These are stable to perturbations of input data and independent of dimensions and coordinate systems. However, this shape strongly depends on the way a signal is represented (i.e., on the representation space). The purpose of this work is to study how the computation of persistent homologies depends on the chosen representation spaces, considering the specific task of vowels categorization in human language. We study the impact of the representation space on the extracted topological information and determine if accessing higher-dimensional persistent homologies allows getting more discriminant information. We also discuss what is the best way to summarize the information contained in a persistence diagram for our specific classification tasks.

This article is organized as follows. First, we introduce the problem and its rationale. Second, we describe in a nutshell the theory and the processing pipeline of TDA. Third, we present the strategy for investigating the problem, the nature of the acquired dataset and our classification aims. Fourth, we report the main results that are extensively discussed in the last section.

5.2. The problem

5.2.1. Motivations for TDA

Topology is the branch of mathematics that deals with the qualitative geometric information of a space (CARLSSON 2009). The tools provided by algebraic topology allow us to capture the shape of the data (A. J. ZOMORODIAN 2005). The topological approach frees itself from the question of metrics and coordinates by studying the properties of a space through its connectivity. It has an interesting explanatory power thanks to its great potential for visualization, and the topological features have a discriminant power which makes TDA a particularly interesting candidate for the classification of natural signals. In this section, we give an overview of application of TDA in signal processing. Additional details on the theoretical foundations of TDA are given in Section 5.3.

A central feature in TDA is the computation of persistent homologies (OTTER, PORTER, TILLMANN et al. 2017). Several pipelines have been proposed for the computation and use of such topological descriptors in data analysis. It typically consists in

5. Topological data analysis of human vowels : Persistent homologies across representation spaces – 5.2. The problem

calculating persistent homologies on the input data (e.g. an audio signal) or a representation of it (e.g., its spectrogram), vectorizing persistence diagrams, and using these characteristics in a model (Frédéric CHAZAL et MICHEL 2021 ; BOISSONNAT, Frédéric CHAZAL et MICHEL 2022). Persistence diagrams and some of their representations have been shown to be stable against noise (COHEN-STEINER, H. EDELSBRUNNER et HARER 2007 ; Frédéric CHAZAL et MICHEL 2021 ; Frédéric CHAZAL, B. T. FASY, LECCI et al. 2014), meaning that small perturbations of the input data results in small changes in the persistent diagram. This stability makes the topological approach an excellent candidate for the description of natural signals. Altogether, the interesting perspectives offered by TDA to face Big Data challenges combined with the fast development of tools for the efficient computation of topological descriptors has led to a proliferation of studies demonstrating the added value of TDA in a variety of contexts.

For example, BARBAROSSA et SARDELLITTI 2020a ; BARBAROSSA et SARDELLITTI 2020b demonstrate the usefulness of TDA for signal processing and for the study of signals on graphs. Topological tools are also useful in analyzing the shape of time series (SEVERSKY, S. DAVIS et BERGER 2016), for object detection in images (PATRANGENARU, BUBENIK, PAIGE et al. 2019) or in sound detection (FIREAIZEN, RON et BOBROWSKI 2022). Besides detection tasks, topological descriptors are also useful for the classification of sound signals (J.-Y. LIU, JENG et Y.-H. YANG 2016) and musical signals (BERGOMI et BARATÈ 2020). It impacts multiple scientific domains alike biology, medicine, ecology (PEREIRA et DE MELLO 2015), neurosciences, e.g., for fMRI (SALCH, REGALSKI, ABDALLAH et al. 2021) or EEG data (NASRIN, OBALLE, BOOTHE et al. 2019 ; Xiaoqi XU, DROUGARD et R. N. ROY 2021) for which TDA allows the construction of invariant signal descriptors. Topological features are complementary to more classical descriptors, and they allow capturing global and high-dimensional information useful for signal analysis. However, the impact of the representation space of the signal on the extracted topological information is, to our knowledge, still an unexplored question.

5.2.2. Representation space

Our study focuses on physical signals, in particular, sound signals. A raw data signal can be represented as different *data objects*, in different, here-called, representation spaces. Figure 5.1 shows the same raw signal in three different ways : as a spectrogram, viewed as a surface in 3D Euclidean ambient space, as the spectrogram's zeros, or as a point cloud using Taken's embedding.

5. Topological data analysis of human vowels : Persistent homologies across representation spaces – 5.2. The problem

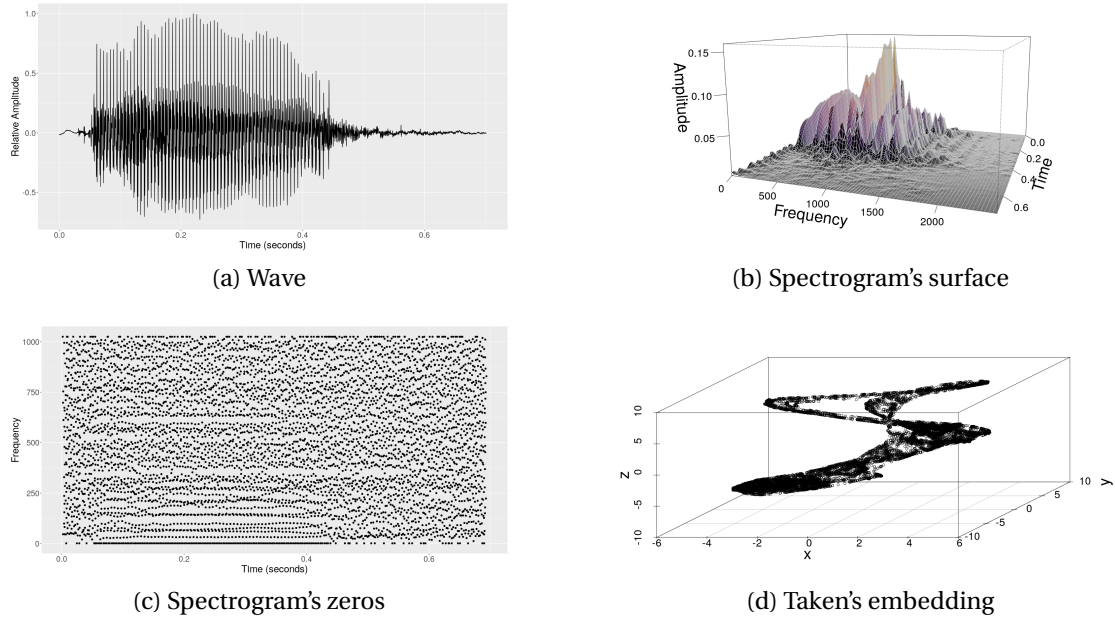


FIGURE 5.1. – Different representation of the same signal. 5.1a is the initial wave of the sound; 5.1b is the surface of its spectrogram; 5.1c is the zeros of its spectrogram in the time-frequency plane; 5.1d is its Taken's embedding.

FLANDRIN 2018 refers to time-frequency analysis as the language of signal processing. Among the plethora of existing representation methods, the spectrogram is certainly the most widely used. Let us consider a sound signal $x \in L_1(\mathbb{R}) \cap L_2(\mathbb{R})$, its spectrogram $S_x^{(h)}(t, \omega)$ is defined as :

$$S_x^{(h)}(t, \omega) = |F_x^{(h)}(t, \omega)|^2, \quad (5.1)$$

where $(t, \omega) \in \mathbb{R}^2$ are time and frequency variables, $h(t) \in \mathbb{R}$ is a window function and $F_x^{(h)}(t, \omega)$ is the Short-Time Fourier Transform (STFT) :

$$F_x^{(h)}(t, \omega) = \int_{-\infty}^{+\infty} x(s)h(s-t) \exp\{-i\omega(s-t/2)\} ds. \quad (5.2)$$

We can therefore represent the spectrogram of a one dimensional sound signal x , with time window function $h(\cdot)$, as a surface $\mathcal{S}_x^{(h)}$ in a 3D ambient Euclidean space where the dimensions are time, frequency and amplitude, $\mathcal{S}_x^{(h)} := \left\{ \left(t, \omega, S_x^{(h)}(t, \omega) \right) \mid (t, \omega) \in \mathbb{R}^2 \right\} \subset \mathbb{R}^3$ (see Figure 5.1b). This representation has been proven effective in revealing geometrical structures, enhancing classification performances (LEVY, NAITSAT et ZEEVI 2022). Another possibility is to rather consider the spectrogram's zeros as introduced in FLANDRIN 2015. Choosing the window function to be Gaussian, i.e, $h(t) = \pi^{-1/4} \exp\{-t^2/4\}$, we readily show that the STFT can be rewritten as follows :

$$F_x^{(h)}(t, \omega) = \exp\{-|z|^2/4\} \mathcal{F}_x(z), \quad (5.3)$$

5. Topological data analysis of human vowels : Persistent homologies across representation spaces – 5.2. The problem

where, $z = \omega + it$ and $\mathcal{F}_x(z)$ is the Bargmann transform of x . It is easy to see that it is an entire function of order 2, which admits a Weierstrass-Hadamard form :

$$\mathcal{F}_x(z) \propto \prod_{n=1}^{\infty} \left(1 - \frac{z}{z_n}\right) \exp \left\{ \frac{z}{z_n} + \frac{1}{2} \left(\frac{z}{z_n}\right)^2 \right\}, \quad (5.4)$$

where $\mathcal{Z}_x := \{z_n = \omega_n + it_n\}_n$ is the set of zeroes of $\mathcal{F}_x(z)$. The spectrogram is therefore completely characterized by its zeroes, in other words, it can be represented by a point cloud in the time-frequency plane. FLANDRIN 2018 pioneered the idea of using topological characteristics for describing a spectrogram by visualizing the distribution of the edges of its zero-based Delaunay-triangulation. In the sequel, we build on this idea for extracting topological characteristics using the tools described hereafter in Section 5.3.

We introduce another representation of sound signals that is not related to time-frequency analysis. Considering a sound signal as a discrete-time digital audio recording (or time series) $\{x_1, \dots, x_T\}$ and assuming it comes from a dynamic system, we can borrow the tools of dynamical system analysis to find out an informative representation. The Taken's theorem states that it is possible to obtain a representation of this time series that is topologically equivalent to the attractor of the system via a delay embedding, which contains useful information about the system (TAKENS 1981). We thus transform the digital audio recording into a point cloud in a higher dimensional space, $\mathcal{P}_{D,x} = \{p_1, \dots, p_m\} \subset \mathbb{R}^D$, each element $p_i \in \mathcal{P}_{D,x}$ is a vector of dimension D , constructed by taking a time delay τ :

$$p_i = (x_i, x_{i+\tau}, x_{i+2\tau}, \dots, x_{i+(D-1)\tau})'. \quad (5.5)$$

We therefore embed a one dimensional digital audio recording into a higher dimensional space to obtain a point cloud. It is necessary to estimate the two hyperparameters : D , the dimension of the space, and τ , the time delay. D is estimated using the Cao's algorithm (L. CAO 1997). τ is selected using the Average Mutual Information (AMI). The coordinates of the phase-space embedding must be independent enough (to avoid aggregation around the diagonal in the embedding). τ is then chosen to be the smallest value such that $AMI(\tau) < \frac{1}{e}$. An example of this point cloud representation is shown in Figure 5.1d.

5.2.3. Same signal, different representations, different topological characteristics

We described three ways of representing the same audio signal. It naturally raises some questions, such as, is there a representation that carries more topological (discriminant) information than the other? To get an intuition about this question, Figure 5.2 shows three persistence diagrams computed on the representations of the same signal illustrated on Figure 5.1 (see section 5.4.3 for more information about how this topological information is computed). Although the signal is the same, the extracted

5. Topological data analysis of human vowels : Persistent homologies across representation spaces – 5.3. Topological Data Analysis : An overview

topological characteristics are very different because of the choice of the representation. Since the representation spaces can be of different dimensions, the homological features describing the topology of the object of these spaces will also be of different dimensions. This raises another, more specific questions, is the access to higher dimensions providing relevant information? We address these issues in a quantitative way, through a case study, a classification task of human vowels.

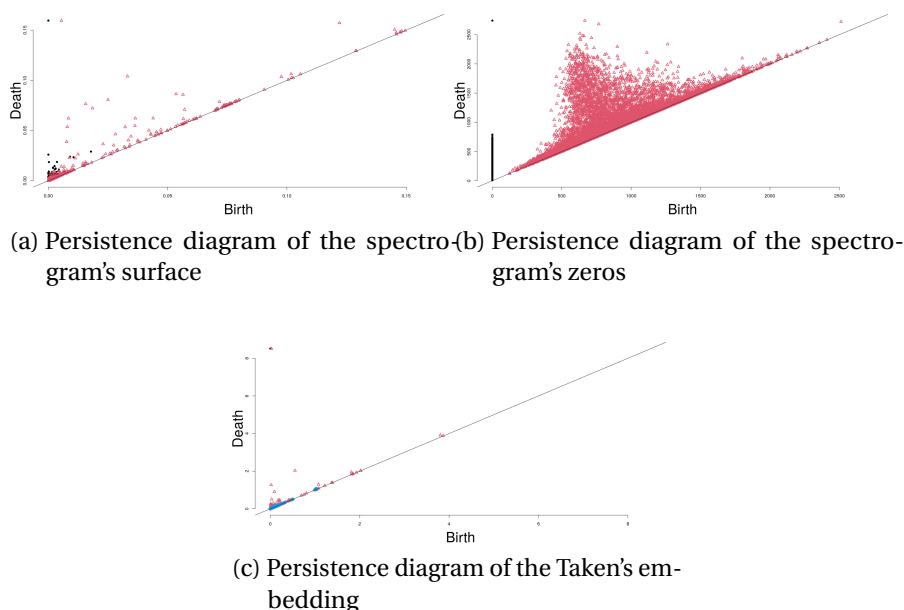


FIGURE 5.2. – Three persistent diagrams computed for the three different representation of the same signal of Figure 5.1. 5.2a is the persistence diagram of the spectrogram's surface using sublevel sets; 5.2b is the persistence diagram of the zeros of the spectrogram using an Alpha complex; 5.2c is the persistence diagram of the Taken's embedding using an Alpha complex. The points in the diagram represent the computed persistent homologies. The different colors represent the different dimensions : black for $p = 0$, red for $p = 1$ and blue for $p = 2$.

5.3. Topological Data Analysis : An overview

This section presents the theory and data analysis pipeline of TDA in a nutshell. For more details, we refer the reader to three important textbooks that offer a fairly broad presentation of TDA and its theoretical foundations (i.e., A. ZOMORODIAN et CARLSSON 2005; H. EDELSBRUNNER et HARER 2009; DEY et Y. WANG 2022) as well as the excellent introduction for data scientists Frédéric CHAZAL et MICHEL 2021.

5.3.1. Persistent homology

In a prosaic summary, TDA is the enumeration of holes of different dimensions in the *shape of data*. This intuitive notion of holes is formalized by the mathematical concept of a homology group. Homology groups allow us to treat the holes of a topological space mathematically, by studying the connectivity of the space. Let us denote as $H_p(\mathbb{M})$ the p^{th} homology group of the topological space \mathbb{M} . Every homology group has the topological properties of its dimension. For $p = 0$, $H_0(\mathbb{M})$ takes the connected elements of \mathbb{M} , for $p = 1$, $H_1(\mathbb{M})$ takes the tunnels, for $p = 2$, the voids. The rank of $H_p(\mathbb{M})$ defines the p^{th} Betti numbers, denoted as $\beta_p = \text{rank}(H_p)$ hereafter. Betti numbers basically count the number of different topological features. Homology groups and Betti numbers are topological invariants which characterizes the shape of \mathbb{M} .

Instead of the homology groups of \mathbb{M} , we are interested in its persistent homology groups. We compute them through a filtration F of \mathbb{M} . A filtration F is a parametrized nested family of subspaces $F = (M_r)_{r \in T}$, where $T \subseteq \mathbb{R}$, such that for any $r, r' \in T$, if $r \leq r'$, $M_r \subseteq M_{r'}$, and $\mathbb{M} = \cup_{r \in T} M_r$. The parameter $r \in T$ is the scale parameter. Let $f: \mathbb{M} \rightarrow \mathbb{R}$ and $M_r = f^{-1}(-\infty, r]$ be the sublevel set for value r , the family $\{M_r\}_{r \in T}$ is the sublevel set filtration. A filtration allows multiscale topological description; for each value of r we have the associated homology groups. The persistence of homologies makes it possible to track the lifetime of homologies, to determine their birth and their death according to the scale parameter r .

We denote as \mathbb{X} our data objects that can be either \mathcal{I}_x or $\mathcal{P}_{x,D}$ (point clouds), or with the two-dimensional surface $\mathcal{S}_x^{(h)}$. A point cloud typically does not carry interesting topological information, nevertheless, it is possible to retrieve some via the construction of a simplicial filtration on the top of it. We can define a simplicial filtration by constructing a simplicial complex on the data. The filtration is then a nested sequence of simplicial complexes, for each of which we compute the (simplicial) homologies. In this paper, we use the Alpha complexes, a family of subcomplexes of the Delaunay complex, since it is equivalent but smaller than other complexes such as the Cech complexes. For any $x \in \mathbb{X}$, with $\mathbb{X} \subset \mathbb{R}^d$, and $r \in \mathbb{R}^+$, we define $B_x(r) = x + r\mathbb{B}^d$ the closed ball centered at x and of radius r . The union of these balls is the set of points at a distance at most r from at least one of the points of \mathbb{X} . We define the Alpha complex

$$\text{Alpha}(r) = \{\sigma \subseteq \mathbb{X} \mid \cap_{x \in \sigma} B_x(r) \neq \emptyset\}, \quad (5.6)$$

where $R_x(r)$ is the intersection of each Euclidean ball with its corresponding Voronoi cell. In the context of filtration, we construct a nested sequence of Alpha complexes on \mathbb{X} , taking an increasing value of r .

We recover the homological information of \mathbb{X} computed for different scale values r . The birth and death of the topological features (the elements that make up each homology group) are recorded and summarized in the persistence diagram, for each value of r . Thus, each persistent topological feature resulting from the filtration is expressed by a pair (b_i, d_j) , the moment of its birth and its death, i.e., two values of r .

The most persistent homologies are those for which the difference $d_j - b_j$ is maximal (B. T. FASY, LECCI, RINALDO et al. 2014). We define a persistence diagram as a multiset of points in $\mathcal{D} := D \times \{1, \dots, P\}$, where

$$D := \{(b, d) \in \mathbb{R}^2 \mid d \geq b \geq 0\}. \quad (5.7)$$

Each triplet $(b, d, p) \in \mathcal{D}$ represents a p -dimensional homological feature that appears when $r = b$ and disappears when $r = d$ (MAROULAS, NASRIN et OBALLE 2020).

In our experiment, we use filtration adapted to the representation spaces. More specifically, the alpha-complex filtration for Taken’s embeddings and spectrogram zeroes and the sublevel set filtration for the spectrogram surface.

5.3.2. Exploitation of the information from the diagram space

Persistence diagrams provide multiscale homological descriptions of the data. However, the space of persistence diagrams has very complicated geometry and topology which makes it difficult to exploit directly (DIVOL et LACOMBE 2021). Many strategies were developed in order to extract the relevant information and enable machine learning applications. We refer the readers to BARNES, POLANCO et PEREA 2021 ; HENSEL, MOOR et RIECK 2021 that offer comprehensive review of existing methods.

Common approaches consist in computing functional representations alike persistent surfaces (H. ADAMS, EMERSON, KIRBY et al. 2017), persistence landscapes (BUBENIK 2015) or persistent silhouettes (Frédéric CHAZAL, B. T. FASY, LECCI et al. 2014), and then, to discretized them to be used as vector-based input for machine learning algorithms. These representations have been proven to enjoy stability properties (BUBENIK 2020) meaning that they are slightly modified by small variations in the input data. Their nice properties and ease to compute make them widely employed in applications. For example, J.-Y. LIU, JENG et Y.-H. YANG 2016 use persistence landscapes as input to a CNN for musical signal classification, and PLLay (K. KIM, ZAHEER, Frederic CHAZAL et al. 2020) is a topological layer based on persistence landscapes.

Instead of functional representations, persistence diagrams can be summarized by scalar descriptors. There is a myriad of such, among which, the p -norm of the persistence diagram (COHEN-STEINER, H. EDELSBRUNNER, HARER et MILEYKO 2010) or the persistent entropy (ATIENZA, GONZALEZ-DIAZ et RUCCO 2019). Both are proven to be stable to perturbations of the input data (ATIENZA, GONZALEZ-DÍAZ et SORIANO-TRIGUEROS 2020). Aggregating those descriptors together appears to be an interesting strategy. CARRIÈRE, OUDOT et OVSJANIKOV 2015 construct a stable vector of the persistence diagram from the distances between points and between each point and the diagonal, while FIREAIZEN, RON et BOBROWSKI 2022 use a whole set of descriptors on the persistence diagrams. In the sequel, we will consider sets of descriptors as well as some standard functional summaries.

5.4. Experiments

In this section, we investigate how the choice of representation space of a digital audio signal affects topological information. This information is quantified in terms of Out Of Bag (OOB) error in a supervised classification task inspired by the topical question in acoustic of vowel classification (KORKMAZ, BOYACI et TUNCER 2019; GEORGIU 2023). It finds applications in emotion classification (DEB et DANDAPAT 2019), in evaluating developmental trouble (VAVRINA, ZETOCHA et TUCKOVA 2012) or in order to distinguish between healthy patients and patients with neurological disorder (VASHKEVICH et RUSHKEVICH 2021).

We collected our own dataset consisting of French-speaking adults pronouncing vowels indoors. For each of these recordings, we extract homological information from three possible representation spaces, as well as classical frequency descriptors from the associated spectrogram. We identify the recordings with the pronounced vowel, the individual pronouncing it, and their sex. Thus, we can quantify whether topological descriptors provide additional information to the more conventional frequency descriptors; whether there is a difference in the topological information depending on the representation space; whether there is a better way of vectorizing the persistence diagrams. In order to visualize the extracted information in a more qualitative way, we also learn the underlying manifold from the set of all extracted topological features of each representation and visualize how the signals are distributed over each manifold.

In what follows, we describe the recorded data and three associated classification tasks on which the test is carried out. Then, we present the three representation spaces and associated filtration we use for each signal. Finally, we present different procedures for vectorization of persistent diagrams and the methodology we use to compare them.

5.4.1. Presentation of the data

The data are digital audio recordings of French-speaking adults. These recordings were made in a controlled environment to keep low and stable signal-to-noise ratio and to avoid junk sounds. We used a Zoom H6 recording device with a stereophonic microphone XYH-6. The sampling rate of the recordings is 44100 Hz, 16 bits. For the present analyses, we subsampled the signals to 16kHz and convert them to monophonic. We recorded 20 individuals, 15 women and 5 men. Each individual pronounced 8 vowels (ä , ã , ə , i , o , õ , u , y) in 7 conditions : natural, low voice, high voice, short, long, on an ascending and descending scale. There were 10 utterances for each condition. Therefore, each vowel has been recorded 1400 times and each individual 560 times. In total, 8400 recordings were made by females and 2800 by males. The audio data set is freely accessible (BONAFOS, FREYERMUTH, PUDLO et al. 2023) and can be downloaded on <https://zenodo.org/record/7961904>.

5.4.2. Comparison on three supervised classification tasks

We consider the three following tasks : prediction of the vowel, prediction of the individual who pronounced the vowel, prediction of the sex of the individual who pronounced the vowel. For each task (and each representation space), we use a random forest as a classification model with the same number of 500 trees. We report the OOB error, which allows us to estimate the error by cross-validation at a lower cost, for models using either topological variables only, or frequency variables only, or both topological and frequency variables (topologically-augmented).

5.4.3. Computation of the homological information

For each record, we compute the spectrogram with a Gaussian window of 11.6 ms and an overlap of 90%. On the spectrogram's surface representation, we apply sublevel set filtration to compute the persistent homologies while on the spectrogram's zeroes, we use an alpha-complex filtration (see section 5.3).

For the representation using Taken's embeddings, we first estimate for each record the two parameters required to build the embedding, τ according the AMI and D according to the Cao's algorithm. Then, we calculate the embeddings of each record according to these two values. Finally, we harmonize the dimension of embedding spaces over recordings by reducing the dimension to 3 via UMAP (MCINNES, HEALY et MELVILLE 2020). The records are represented by point clouds in \mathbb{R}^3 . On this space, we compute the persistent homologies using an alpha-complex filtration.

At the end, each audio signal is associated to three persistent diagrams carrying potentially different topological information.

We also compute the Mel-Frequency Cepstral Coefficients (MFCC) (CHACHADA et KUO 2014; SUEUR 2018). Those are more classical frequency descriptors of the signal, specifically designed for human speech analysis and generally used in machine learning for classification tasks involving human speech. They will serve as a baseline to compare our topological descriptors. Finally, in a topologically-augmented machine learning fashion, we will merge the MFCC with the topological descriptors.

5.4.4. Extraction of the information from the persistence diagram and comparison of the persistent variables

As discussed in part 5.3.2, there are many ways to summarize the information carried by persistence diagrams into variables that facilitate further statistical/machine learning usage. We do not claim to be exhaustive in our comparison. We follow various proposals from the literature to form a set of variables computed on persistence diagrams, here so-called *persistent variables*. In addition to persistent variables, we also compute functional summaries. To be fair in comparison, we use the same classification model in each case.

5.4.4.1. Persistent variables

Let \mathcal{D} a persistence diagram, the multiset defined in equation (5.7). We compute each variable for $p = \{0, 1\}$, for the Taken's embeddings we also compute for $p = 2$.

From ATIENZA, GONZALEZ-DIAZ et RUCCO 2019, we compute the persistent entropy E_p . Let $L_p = \{\ell_i = d_i - b_i | 1 \leq i \leq n\}$ the set containing the lifetime of each homological features of dimension p of the diagram D . We define the persistent entropy as

$$E_p(F) = - \sum_{i=1}^n p_i \log(p_i), \quad (5.8)$$

where $p_i = \frac{\ell_i}{S_L}$ and $S_L = \sum_{i=1}^n \ell_i$. We compute it for each dimension.

We follow FIREAIZEN, RON et BOBROWSKI 2022 to compute several descriptors of the persistence diagram. Using L_p , we compute some statistics from the vector :

- i) the mean μ_p .
- ii) the variance σ_p^2 .
- iii) the 5 top longest lifetimes $L_{p,i}$ for $i = 1, \dots, 5$.
- iv) two normalized longest lifetime,

$$\frac{L_{p,1}}{|L_p|}$$

.

- v) and

$$\frac{L_{p,1}}{\mu_p}.$$

- vi) the number of α -long lifetime $N_{p,\alpha}$. We choose α to distinguish between topological noise and information from the diagram. We take $\alpha = 0.05$.

$$N_{p,\alpha} = \#(L_{p,i} > \alpha).$$

- vii) the ratio of means

$$\frac{\mu_{p_1}}{\mu_{p_2}}.$$

p_1 and p_2 being two dimensions of homological features, with $p_1 < p_2$.

- viii) the ratio of α -long cycles

$$\frac{N_{p_1,\alpha}}{N_{p_2,\alpha}}.$$

- ix) the products of top longest lifetimes

$$L_{p_1,i} \cdot L_{p_2,i},$$

for $i = 1, \dots, 6$.

5. Topological data analysis of human vowels : Persistent homologies across representation spaces – 5.4. Experiments

x) the products

$$L_{p_1,i} \cdot (L_{p_2,i} - L_{p_2,i+1})$$

for $i = 1, \dots, 6$.

xi) the Periodicity Score

$$PS = 1 - \frac{L_{1,2}}{L_{1,1}}.$$

xii) the Quasi-Periodicity Score

$$QPS = L_{1,2} \cdot L_{2,1}.$$

xiii) the Frequency Shift Score

$$FSS = \frac{L_{2,1} \cdot L_{2,2}}{L_{1,1}}.$$

We compute four more variables from PEREIRA et DE MELLO 2015 :

— the persistent Betti number. For any pair of indices $0 \leq k \leq l \leq n$ and any dimension p , the p^{th} persistent Betti number is

$$\beta_p^{k,l} = \sum_{i \leq k} \sum_{j > l} \kappa_p^{i,j}, \quad (5.9)$$

where $\kappa_p^{i,j}$ is the number of p -dimensional homology that are born at X_i and die at X_j , X_i and X_j being two subsets of the filtration. The persistent Betti number is the number of holes for each dimension, and is called so in PEREIRA et DE MELLO 2015. We follow H. EDELSBRUNNER et HARER 2009 and call it the persistent Betti number;

— the maximum hole lifetime in each dimension

$$\max_p = \max_{\ell_i \in L_p} (\ell_i); \quad (5.10)$$

— the number of relevant holes

$$\text{n_rel}_p = \sum_{\ell_i \in L_p} f(\ell_i, \max_p, \text{ratio}), \quad (5.11)$$

where $f(\ell_i, \max_p, \text{ratio})$ equals 1 if $\ell_i \geq \max_p \times \text{ratio}$, 0 otherwise. It is the number of points in the persistent diagram that are relatively distant from the diagonal. We chose to count the holes with a lifetime at least greater than a quarter of the longest (*i.e.*, $\text{ratio} = 0.25$);

— the sum of all lifetime

$$\text{sum}_p = \sum_{\ell_i \in L_p} (\ell_i). \quad (5.12)$$

Following COHEN-STEINER, H. EDELSBRUNNER, HARER et MILEYKO 2010, we com-

pute the p -norm of the persistence diagram,

$$\|D\|_p = \left[\sum_{u \in D} \text{pers}(u)^p \right]^{\frac{1}{p}}, \quad (5.13)$$

where u is a point of the diagram and $\text{pers}(u)$ the absolute value of the difference between the coordinates. In practice, we compute it for $p = 2$.

The set of persistent variables is used alone in the different classification tasks, as well as in combination with the MFCCs. This data fusion method, which we expect to be the most efficient according to the results from the literature, can be found under the name "Topology augmented" in Table 5.1, which summarizes the results for the different tasks.

5.4.4.2. Functional summary of the persistence diagrams

In addition to the set of persistent variables, we consider functional summary of the persistence diagrams. After being discretized, they can use it as input of a classification model. We present and test two of these methods.

First, we compute the silhouette of a persistence diagram (Frédéric CHAZAL, B. T. FASY, LECCI et al. 2014). It follows the persistence landscapes (BUBENIK 2015; BUBENIK 2020). The silhouette summarizes the persistence diagram in a single function. We choose this representation to illustrate the mapping of the persistence diagram into a Hilbert space. We define it as

$$\phi^{(\gamma)}(t) = \frac{\sum_{j=1}^m |d_j - b_j|^\gamma \lambda_j(t)}{\sum_{j=1}^m |d_j - b_j|^\gamma}. \quad (5.14)$$

The silhouette takes a parameter γ . This determines whether all points are treated equally (γ small) or whether the most persistent pairs of points are given more weight (γ large). We set $\gamma = 1$, because some results show the importance of what is sometimes considered as "topological noise" (e.g., (PATRANGENARU, BUBENIK, PAIGE et al. 2019)). We also have to fix the number of sample on which we build the silhouette. We set $n_{\text{sample}} = 2^9 = 512$.

Next, we compute the persistence images of the diagrams (H. ADAMS, EMERSON, KIRBY et al. 2017). The idea is to map the persistence diagram to an integrable surface, so-called the persistence surface

$$\rho_D(u, v) = \sum_{(x,y)=(b,d-b) \in D} w(x, y) g_{(x,y)}(u, v),$$

where $g_{(x,y)}(u, v) : \mathbb{R}^2 \rightarrow \mathbb{R}$ is bivariate Gaussian distribution centered at each point $(x, y) = (b, d - b) \in D$ and $w : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a continuous and piecewise differentiable weighting function. $(u, v) \in \mathcal{D}$ is a compact domain, e.g., the domain of definition of the points $(x, y) = (b, d - b) \in D$. Then, we divide the domain in a collection of non-overlapping subdomains, the pixels P_i , with $\mathcal{D} = \bigcup P_i$. We integrate the persistence

surface over the fixed grid to define the persistence image, taking the average of ρ in each pixel,

$$I_{P_i}(\rho_D) = \int \int_{P_i} \rho_D(u, v) dudv. \quad (5.15)$$

This outputs an image representing the persistence diagram, with a density distribution that is more or less important depending on the distribution of homologies on the persistence diagram. Again, we can use these vectors as input to a classification model. A sampling parameter must be set, which is fixed at 10.

5.4.5. Step-wise selection of the variables

As aforementioned in Section 5.3.2, the question of how to extract information from the persistence diagram for statistical purposes remains widely open and is not to be addressed in this paper. We instead focus on the study of the differences in persistent homologies between representation spaces of a given signal, and we bring complementary information to the field by comparing the most frequently occurring topological variables.

In order to carry out this comparison, we follow a step-wise strategy. For each classification task, we start with the complete set of persistent variables. We estimate the model with this set of variables, and then we remove the least important variable. We retrain the model with this new set of variables. This process continues until we remove all the variables in the training set. In the end, we retrain the model with the smallest OOB error. We report in Table 5.1 the results of the best model, and we list the variables present in the set of variables used to train this model. We count each time a variable is in the set of the best model, for each task and for each representation.

This information can be found in the Table 5.2. We can then compare whether there is a difference according to the input representation, whether certain topological variables carry more interesting information than others, or at least whether they are more often found in the best model for the classification tasks in question. Since our procedure includes MFCC, we potentially remove some MFCC descriptors while keeping some topological variables in models producing the best results.

There are three classification tasks (prediction of the vowel, of the sex, and of the individual), three initial signal representations (spectrogram's surface, spectrogram's zeros, taken's embeddings). For each, we follow our step-wise strategy to find the best model, eliminating iteratively the less useful variable, with two conditions : with and without the MFCC. We save the 18 best models, and we count the remaining persistent variables in each of them. Lastly, we follow the same procedure by including the topological variables of all representations in the same model, for the three tasks. Thus, a variable can be counted in a maximum of 24 models.

We put in bold in Table 5.2 the variables that appear at least in 50% of the best models. In the signal representation columns, each variable can appear at best 12 times. Here again, we put in bold those that appear at least in 50% of the best models. This way, it will be possible to compare whether certain variables appear more than others globally, whatever the representation and whatever the task. We can also check

if there are differences depending on the representation space. For the sake of clarity, we resume the top 5 longest lifetimes $L_{p,i}$ and their product $L_{p_1,i} \cdot L_{p_2,i+1}$ by a row in Table 5.2 for each dimension. Also, the maximum count for these variables is not 24, but $24 \times 5 = 120$.

5.5. Results

5.5.1. Supervised task

The results for all supervised classification tasks can be found in Table 5.1. All models are random forests with the same number of trees. For models using as covariates either persistent variables or both persistent variables and MFCCs, the reported results are those of the best model (i.e., the model following the stepwise procedure, enjoying the lowest OOB error).

5.5.1.1. TDA is useful for signal classification

Topological information improves the results of two over three classification tasks. For vowel classification, the MFCCs alone obtain an error of 8.71%. The addition of topological information improves the results, whatever the chosen representation space. The error reduces to 8.43%, 8.03%, 8.49% using the persistent variables obtained from spectrogram's surfaces, spectrogram's zeros and Taken's embeddings, respectively. The best improvement is obtained by taking all persistent variables from all representation spaces, resulting in an error of 7.98%.

The individuals' classification, also benefits from topological information. Indeed, MFCCs alone are outperformed when adding topological information, whatever the representation space. The best results are obtained when the persistent variables are extracted from the spectrogram's zeros, lowering the error from 11.54% to 9.24%. On the contrary to vowel classification, taking all persistent variables from all representation spaces deteriorates the results.

Finally, on the sex classification task, topology augmented approach fails to improve the results over the MFCCs alone, which exhibits the lowest error at 4.54%.

5.5.1.2. Topological information alone is not enough to discriminate the signal

Topological information alone, whatever the chosen representation space, is outperformed by classical frequency descriptors on each task. For the vowel classification task, while the MFCCs alone achieve an error of 8.71%, the best result using only topological variables is of about 41%. It is obtained with persistent variables from all representation.

For the sex classification task, using MFCCs alone give the best result, with an error of 4.54%. Using topological information, the best models always consider the persistent

5. Topological data analysis of human vowels : Persistent homologies across representation spaces – 5.5. Results

TABLEAU 5.1. – Comparison of Out Of Bag error (OOB) for different signal representations. Bold numbers emphasize the best signal representation for each task

Signal Representation	Vectorization ^a	Error (%)		
		Vowel	Sex	Individual
MFCC		8.71	4.54	11.54
Spectrogram's Surface	Silhouettes $p = 0$	72.97	20.37	73.72
	Silhouettes $p = 1$	46.91	19.34	63.34
	Silhouettes $p = 0, 1$	45.04	16.44	53.86
	Persistent Image $p = 0$	79.31	28.46	88.27
	Persistent Image $p = 1$	79.26	28.73	87.97
	Persistent Image $p = 0, 1$	76.99	26.1	83.43
	Persistent Variables	52.03	15.72	50.07
	Topology augmented ^b	8.43	4.71	10.42
Spectrogram's Zeros	Silhouettes $p = 0$	79.89	23.69	78.21
	Silhouettes $p = 1$	73.33	17.61	75.69
	Silhouettes $p = 0, 1$	70.11	16.14	67.5
	Persistent Image $p = 0$	82.48	25.08	87.77
	Persistent Image $p = 1$	70.5	16.79	71.15
	Persistent Image $p = 0, 1$	70.7	17.19	70.68
	Persistent Variables	69.85	15.19	62.15
	Topology augmented ^b	8.03	5.55	9.24
Taken's embeddings	Silhouettes $p = 0$	84.39	27.71	89.15
	Silhouettes $p = 1$	81.51	25.32	84.44
	Silhouettes $p = 2$	79.71	25.02	81.19
	Silhouettes $p = 0, 1$	78.03	24.59	80.63
	Silhouettes $p = 0, 2$	78.79	24.56	77.86
	Silhouettes $p = 1, 2$	78.56	24.68	77.22
	Silhouettes $p = 0, 1, 2$	76.41	23.98	75.53
	Persistent Image $p = 0$	85.11	29.63	90.24
	Persistent Image $p = 1$	81.89	27.06	84.87
	Persistent Image $p = 2$	83.5	27.15	87.42
	Persistent Image $p = 0, 1$	80.76	26.29	82.42
	Persistent Image $p = 0, 2$	81.54	25.94	82.93
	Persistent Image $p = 1, 2$	80.26	26.21	82.24
	Persistent Image $p = 0, 1, 2$	79.53	25.76	79.86
	Persistent Variables	69.85	20.27	60.89
	Topology augmented ^b	8.49	4.79	10.77
All together	Persistent Variables	41.56	10.89	32.13
	Topology augmented ^b	7.98	6.01	10.3

^a Strategy to extract information from the persistence diagram

^b Signal representation = MFCC + persistent variables. We follow the step-wise strategy presented in section 5.4.5 on this set and present the result of the best model.

5. *Topological data analysis of human vowels : Persistent homologies across representation spaces – 5.5. Results*

variables, with an error ranging from 15.19% to 20.27%. Aggregating the persistent variables of each representation improves the results, with an error of 10.89%.

For the individual classification task, it is once again the set of persistence variables of all three representations, that obtains the best results among models considering only topological information. It performs poorly with an error of 32.13% while MFCCs alone reach an error of 11.54%.

5.5.1.3. **Complementarity of representations**

Topology augmented approaches improve the results over the MFCCs for both vowel and individual classification. Comparing the topology augmented approaches for the three representations on the vowel classification task, it is the variables extracted from the persistence diagrams of the spectrogram's zeros that achieve the minimum error. For the individual classification task, the variables extracted from the persistence diagrams of the zeros of the spectrograms also provide the best score for all the approaches compared for this task.

While the topology augmented approach with the addition of the persistent variables of the spectrogram's zeros provides the best improvement for both tasks, there does not appear to be one better representation of the signal or one with a more informative topology than the others. The differences are held and, for the sex classification task, the topology augmented models with the persistence diagrams of the spectrogram's surfaces and Taken's embeddings outperform the spectrogram's zeros. For the vowel classification task, the best model is the one taking the full set of persistent variables in addition to the MFCCs, those computed on all representations.

The resulting 'data objects' have different topology, which would be potentially complementary. Indeed, we note that, for each task, the model learned with the set of all persistent variables, computed on the three representations, and without the MFCCs, performs better than each model learned with the set of persistent variables computed on each representation. Thus, for vowel classification, while the persistent variables of the spectrogram's surfaces have an error of 52.03%, those of the spectrogram's zeros 69.85% and those of the Taken's embeddings 69.85%, the model trained on the set of these persistent variables reaches an error of 41.56%. Moreover, by adding the MFCCs, it is this set that gives the best results for this task. For sex classification, the persistent variables computed on the spectrogram's surfaces have an error of 15.72%, those on the spectrogram zeros 15.19% and those on the Taken's embeddings 20.27%. All together, the model still obtains a clear improvement, with an error of 10.89%. For classification of the individuals, the persistent variables computed on the surfaces of the spectrogram have an error of 50.07%, those of the zeros of the spectrogram 62.15%, those of the Taken's embeddings 60.89%. The model learned on the whole of these variables obtains an error of 32.13%. Thus, we improve each time the results by merging with all the persistent variables.

5. Topological data analysis of human vowels : Persistent homologies across representation spaces – 5.5. Results

TABLEAU 5.2. – Most frequently kept persistent variables in each best model. The variables present in more of 50% of the best models are emphasized in bold

Variable	Frequency in the best model (%) ^a	Signal Representation		
		Spectrogram's Surface	Spectrogram's Zeros	Taken's embeddings
β_0 (5.9)	62.5	33.33	41.67	50
β_1	33.33	50	0	16.67
β_2	25	\emptyset	\emptyset	50
\max_0 (5.10)	29.17	58.33	0	0
\max_1	0	0	0	0
\max_2	0	\emptyset	\emptyset	0
n_rel_0 (5.11)	0	0	0	0
n_rel_1	4.17	0	1	0
n_rel_2	0	\emptyset	\emptyset	0
sum_0 (5.12)	41.67	66.67	0	2
sum_1	50	5	2	5
sum_2	8.33	\emptyset	\emptyset	16.67
E_0 (5.8)	29.17	50	0	8.33
E_1	33.33	50	0	16.67
E_2	4.17	\emptyset	\emptyset	8.33
$\ D_0\ _p$ (5.13)	37.5	50	16.67	8.33
$\ D_1\ _p$	50	25	50	25
$\ D_2\ _p$	0	\emptyset	\emptyset	0
μ_0 (i)	87.5	83.33	75	16.67
μ_1	79.17	66.67	50	41.67
μ_2	4.17	\emptyset	\emptyset	8.33
σ_0^2 (ii)	45.83	33.33	41.67	16.67
σ_1^2	66.67	25	100	8.33
σ_2^2	0	\emptyset	\emptyset	0
$\frac{L_{0,1}}{ L_0 }$ (iv)	25	33.33	0	16.67
$\frac{L_{1,1}}{ L_1 }$	33.33	58.33	0	8.33
$\frac{L_{2,1}}{ L_2 }$	0	\emptyset	\emptyset	0
$\frac{L_{0,1}}{\mu_0}$ (v)	54.17	41.67	41.67	25
$\frac{L_{1,1}}{\mu_1}$	37.5	50	8.33	16.67
$\frac{L_{2,1}}{\mu_2}$	0	\emptyset	\emptyset	0
$N_{0,\alpha}$ (vi)	29.17	0	58.33	0
$N_{1,\alpha}$	25	0	8.33	41.67
$N_{2,\alpha}$	0	\emptyset	\emptyset	0
$\frac{\mu_0}{\mu_1}$ (vii)	87.5	100	33.33	41.67
$\frac{\mu_1}{\mu_2}$	12.5	\emptyset	\emptyset	25
$\frac{\mu_2}{\mu_0}$	4.17	\emptyset	\emptyset	8.33
$\frac{N_{0,\alpha}}{N_{1,\alpha}}$ (viii)	20.84	0	33.33	8.33
$\frac{N_{0,\alpha}}{N_{2,\alpha}}$	0	\emptyset	\emptyset	0
$\frac{N_{1,\alpha}}{N_{2,\alpha}}$	4.17	\emptyset	\emptyset	8.33
PS (xi)	12.5	25	0	0
QPS (xii)	0	\emptyset	\emptyset	0
FSS (xiii)	0	\emptyset	\emptyset	0
$L_{0,i}^b$ (iii)	21.67	36.67	5	1.67
$L_{1,i}^b$	50.83	43.33	51.67	6.67
$L_{2,i}^b$	5	\emptyset	\emptyset	10
$L_{0,i} \cdot L_{1,i}^b$ (ix)	27.5	31.67	15	8.33
$L_{0,i} \cdot L_{2,i}^b$	0	\emptyset	\emptyset	0
$L_{1,i} \cdot L_{2,i}^b$	2.5	\emptyset	\emptyset	5
$L_{0,i} \cdot (L_{1,i} - L_{0,i+1})^b$ (x)	5	5	3.33	1.67
$L_{0,i} \cdot (L_{2,i} - L_{0,i+1})^b$	0.3	\emptyset	\emptyset	1.67
$L_{1,i} \cdot (L_{2,i} - L_{1,i+1})^b$	0	\emptyset	\emptyset	0

^a We follow the step-wise strategy to choose the best model for each classification task. we count in how many models each variable appears.

^b We count if one of them is kept, $\forall i \in \{1, \dots, 6\}$

5.5.2. Best topological variables

For these classification tasks, persistent silhouettes or persistent images perform worse than the persistent variables in all scenarios. For this reason, we build the topologically augmented approach only with the persistent variables. They are therefore not taken into account in the stepwise procedure to identify the best topological variable (see Section 5.4.5). All the results are presented in Table 5.2.

In general, there are 8 persistent variables that we encounter at least in 50% of the best models. These are the persistent Betti number in dimension $p = 0$, the sum of L_p in dimension $p = 1$, the 2-norm of the persistence diagram for $p = 1$, the average of L_p for $p = 0$ and $p = 1$ and its variance for $p = 1$, the normalized longest lifetime for $p = 0$, with the normalization computed with the mean of L_p of this dimension, the ratio of the mean of L_0 and L_1 .

The number of variables retained in 50% of the best models varies significantly from one representation space to another. We identified 11, 5, 2 persistent variables retained for the spectrogram's surface, for the spectrogram's zeros and for the Taken's embedding, respectively.

There are two persistent variables that seem to stand out the most : the persistent Betti number and the norm of the persistence diagram. The other persistent variables which appear regularly are linked and summarized in the vector L_p . We find statistics on this vector that summarize the information of the persistence diagrams (mean, variance, normalized maximum), whatever the initial representation of the signal.

5.5.3. Unsupervised analysis, learning the manifold

In order to describe better the behavior of the homological features associated to the different representation spaces, we provide a visual and qualitative help using the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) algorithm. Figure 5.3 represents the projection of the recordings onto the 2d space learned with the UMAP. It distinguishes MFCC, persistent variables accordingly to the representation space, and the three tasks considered therein.

Figure 5.3 reveals that the manifold learned on the MFCCs creates clear clusters for vowels. This is less the case for sex, where observations from both classes are quite close in this space. Same comment applies for individuals. Although some observations are centered, the distribution of individuals in this space can be quite spread out. It is even harder to see a pattern for the manifold learned on persistent variables, whatever the representation space. In fact, the manifolds learned from the persistent variables are much more connected, and we do not distinguish clear clusters in these spaces.

Interestingly, the overall shapes of manifolds learned on Taken's embeddings and on spectrogram's surface look quite similar, and rotated by 180° . While connectivity is still present, manifolds learned on topological variables extracted from zero's spectrogram looks very different. This might be explained by the number of coordinates in the representation space within the persistent homology. Taken's embedding and the

5. Topological data analysis of human vowels : Persistent homologies across representation spaces – 5.5. Results

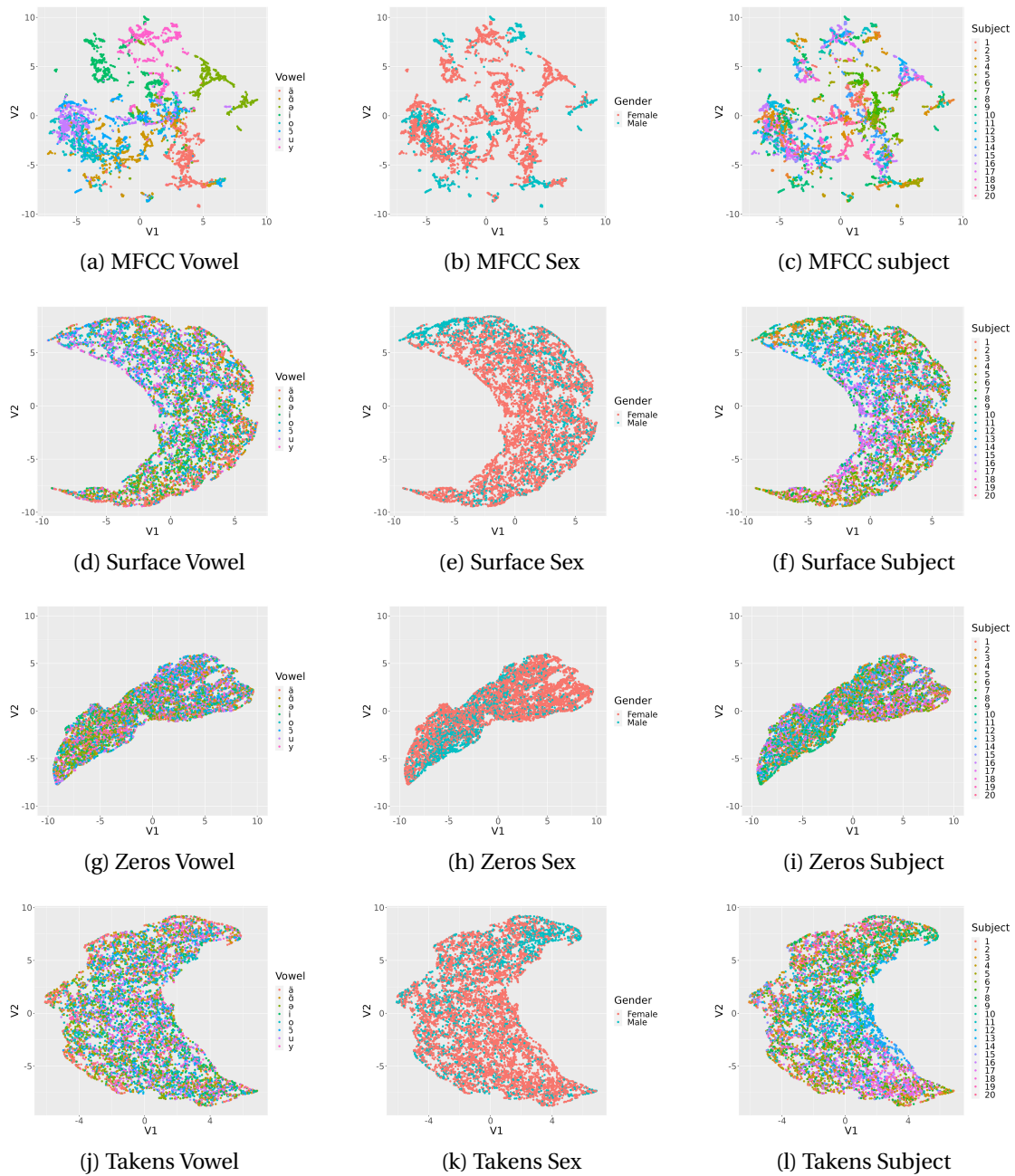


FIGURE 5.3. – Projection of the records on the 2d space learned with UMAP, depending on the input of the algorithm : MFCC, topological variables from the spectrogram’s surface, the spectrogram’s zeros or the Taken’s embedding.

spectrogram’s surface lie in a 3-dimensional space, whereas the spectrogram’s zeros are in the plane.

5.6. Discussion

The main results that emerge from the present study are fourfold. Firstly, topological information improved some classification results. Secondly, the choice of the signal representation space has an impact on the topological information that is extracted. Thirdly, the best way to use a persistence diagram for a classification task depends on the task. Finally, topological information extracted from the spectrogram's zeroes appeared to be particularly complementary to the MFCC and provided the best results out of all tested models for individual prediction.

5.6.1. On the improvements on the prediction of labels

On the one hand, no matter how the information is extracted from the persistence diagrams, topological information never outperformed the MFCCs. On the other hand, adding topological variables to frequency variables improves the results in two of the three considered classification tasks. We thus confirmed results already described in the literature (CARRIÈRE, OUDOT et OVSJANIKOV 2015; Xiaoqi XU, DROUGARD et R. N. ROY 2021; Frédéric CHAZAL et MICHEL 2021). Topological descriptors of signals bring complementary information to MFCCs, which are specifically designed for human speech analysis.

These results contribute to fill an existing gap in the analysis of time-varying data, as noted in some recent reviews (HENSEL, MOOR et RIECK 2021). They are encouraging enough to continue investigating the performance of topologically-augmented machine learning approaches even for natural signal suffering from much lower signal-to-noise ratio as one might consider robust filtration alike Distance-To-Measure introduced in Frédéric CHAZAL, B. FASY, LECCI et al. 2018.

Our results show no improvement for the sex classification task. Remark that it is the simplest task as it is binary classification, while the other have 8 and 20 classes, for vowels and individuals prediction, respectively. Another possible reason is the pitch difference between men and women. Topological characteristics are certainly invariant to pitch modulations, and MFCC's particularly well suited.

When the task becomes more complicated, topological information proved to be useful. The most notable improvement is for the prediction of the speaker. Persistent homologies seem to carry this additional information needed to improve the classification performance.

5.6.2. Different objects, different topologies

5.6.2.1. The difference between representation spaces

As already noticed, it is difficult to identify one representation of the signal as being better than another. Nevertheless, there are two points to bear in mind from the classification results according to the initial representation space :

- For the two tasks where topological information improves the results, i.e., individual and vowel classification, the persistent homologies extracted from the time-frequency plane of the spectrogram zeros improve the OOB error the most. Taken’s embeddings are the representation with the least improvement for both tasks. Thus, access to higher dimensional homologies does not seem to provide discriminative information about the signal.
- The aggregation of topological information is also advantageous. It allows the best improvement in the prediction of vowels and, when it does not allow the best improvement in the prediction of individuals, the improvement is better than for the other two representations. Moreover, for the models trained only on the persistent variables, without the MFCCs, the best results are always obtained when all the representations are combined. The homological information of the representation spaces therefore seems complementary.

This shows that a particular representation space, reveal different salient topological features. Very often, Taken’s embeddings are considered in topological signal analysis. In this paper, we introduced two ways to derive topological information from spectrogram that have the advantage of being more interpretable.

5.6.2.2. A different perspective

The topological signatures we computed on the different signal representation spaces are complementary to more classical descriptors such as MFCCs. This has been highlighted in supervised classification tasks (CARRIÈRE, OUDOT et OVSJANIKOV 2015; Xiaoqi XU, DROUGARD et R. N. ROY 2021 ; Frédéric CHAZAL et MICHEL 2021).

The manifolds we learn from the persistent variables, presented in Figure 5.3, are strongly connected. The topological approach is metric-free and based on the connectivity (CARLSSON 2009). A topological latent space does not clearly cluster the data.

We find a similar observation in the topological autoencoder (MOOR, HORN, RIECK et al. 2020). This property makes entangled structures appear that are impossible in a clear spatial separation. This would be useful to illustrate a hierarchy in the data, with a parent-child structure.

The value of the topological approach depends on the task (WASSERMAN 2018) and is particularly suited to tasks requiring analysis of nested categories. If one expects such a structure in the data, and we wish to highlight it, it is interesting to examine its topology, in order to look at the problem from a blind spot to more classical analysis tools.

5.6.3. On the more present persistent variables

Finally, regarding the vectorization of persistence diagrams, it seems difficult to characterize topological signatures, at least among those tested. We confirm a result already present in other reviews studying this question (BARNES, POLANCO et PEREA 2021), which find that the most interesting topological signatures seem to depend

on the studied data set. Here we found that mapping the persistence diagram onto a functional space, such as Frédéric CHAZAL, B. T. FASY, LECCI et al. 2014; H. ADAMS, EMERSON, KIRBY et al. 2017, was less efficient than taking a set of scalars that summarize the information contained in the diagram. Nevertheless, these approaches have the advantage of having well established stability properties and have been used elsewhere (BUBENIK 2020; J.-Y. LIU, JENG et Y.-H. YANG 2016; K. KIM, ZAHEER, Frederic CHAZAL et al. 2020). This does not disqualify them for other tasks, and they might even be effective on this task, if handled more carefully (for example, by tuning finely the different parameters they depend on, or by computing well-chosen summary data).

This question also raises the issue of what is considered to be topological noise. Topological noise generally designates homological features with very short lifespans. It can be seen on persistent diagrams as the set of points along the main diagonal. We might ask ourselves, is topological noise really noise? Indeed, it has already been noticed that topological noise can be useful to characterize a signal (PATRANGENARU, BUBENIK, PAIGE et al. 2019). The persistence diagrams obtained after alpha-complex filtration on the zeros of the spectrograms are visually the diagrams with the most persistent homologies, many of them close to the diagonal, as can be seen in Figure 5.2. Yet, it is this representation that gives the best improvements. Perhaps these homologies are not just noise and contain information. It is important to bear this in mind when choosing a way to use the information contained in persistence diagrams, as not all methods treat elements close to the diagonal in the same way.

Finally, we did not test all the existing approaches to this question. In fact, we adopted a strategy consisting of extracting information from the persistence diagrams, either by mapping persistence diagrams onto function space (persistent landscape lives in Banach space) or by calculating vectors to describe them.

5.7. Conclusion

We discussed the potential added value of the topological approach to sound signal processing by studying the differences according to the representation space of the signal. We tested it on three classification tasks, predicting the sex of the speaker, the pronounced vowel and the identity of the speaker. For two of these tasks, vowel and identity prediction, the topological features improve the results compared to the baseline. Although it is difficult to distinguish one representation space as being more informative than another, it seems that the topological descriptors computed on each of them are complementary. Our results suggest the use of less common representations than Taken's embeddings, such as spectrogram's surfaces or spectrogram's zeros. For individual classification, the zeros give the best results with the topologically-augmented approach. Moreover, parameter selection and interpretation of a spectrogram is simpler than Taken's embeddings. We analyzed different ways of vectorizing information from persistence diagrams. The best results were obtained using a set of persistent variables regardless of the representation space. This shows that the topological approach offers a complementary and interesting angle of analysis.

5. Topological data analysis of human vowels : Persistent homologies across representation spaces – 5.7. Conclusion

It is not sufficient on its own to discriminate the signal, but it does provide additional information about its hierarchical structure. It would be interesting to analyze more theoretically the complementarity of the representation spaces of the signal, and to see the possibility of working directly in the space of persistence diagrams.

In the end, we believe that TDA has a lot to offer to sound signal processing, in particular for individual recognition tasks.

6. Modélisation bayésienne

Sommaire

6.1. Remise en contexte et résumé de l'étude	127
6.2. Définition du processus de Dirichlet	128
6.2.1. Définition formelle	128
6.2.2. Stick-breaking	130
6.2.3. Processus du Restaurant Chinois	131
6.3. Modèle de mélange avec processus de Dirichlet	134
6.3.1. Le modèle de mélange classique	135
6.3.2. Cas non-paramétrique	136
6.4. Estimation du modèle	138
6.4.1. Les propositions algorithmiques	138
6.4.2. Retour sur le modèle	140
6.4.3. Détermination de la partition	142

6.1. Remise en contexte et résumé de l'étude

Nous avons présenté dans le Chapitre 3 notre méthode pour extraire les segments de vocalisations dans les enregistrements continus. Nous avons ensuite mesuré la plus-value à intégrer une information topologique dans la représentation de ce signal étant donné notre problème dans le Chapitre 5 précédent, et comparé les différences entre les descripteurs topologiques du signal selon l'objet représentant le signal sur lequel on les avait calculés. Nous avons tous les outils pour pouvoir proposer une classification non-supervisée de ces vocalisations, la troisième contribution de cette thèse.

Pour cela, on estime un modèle de mélange gaussien avec processus de Dirichlet sur les vocalisations qu'un enfant a produit durant une année. On représente le signal par des descripteurs cepstraux et des variables topologiques. On étudie ensuite la partition induite par le modèle, notamment en comparant les exemples les plus profonds de chaque classe, ceux-ci étant supposément les plus représentatifs.

Cette classification se base donc sur l'utilisation d'un modèle bayésien non-paramétrique, et plus particulièrement sur un modèle de mélange avec processus de Dirichlet. Comme un modèle de mélange classique, un modèle de mélange non-paramétrique permet de faire une partition d'un ensemble de données en K classes mais, alors que

le nombre de classes K doit être spécifié a priori dans l'approche paramétrique classique, comme un choix de modélisation, celui-ci est estimé sur l'ensemble de données dans l'approche non-paramétrique. Un modèle de mélange non-paramétrique est un modèle de mélange pour lequel K tend vers l'infini. La complexité du modèle s'adapte ensuite aux données, on utilise finalement un nombre K fini d'éléments de mélange étant donné l'ensemble de données fini, tout en gardant la possibilité d'intégrer un savoir expert grâce à l'a priori propre à la méthode bayésienne.

C'est d'ailleurs par l'a priori que l'on passe du modèle bayésien paramétrique au modèle non-paramétrique. On équipe l'espace des paramètres d'une loi a priori permettant de le représenter comme un espace infini dénombrable. En l'occurrence, on utilise un processus de Dirichlet. On présente dans ce Chapitre les éléments nécessaires sur lesquels repose le modèle utilisé dans le Chapitre 7 suivant.

On commence par présenter ce qu'est un processus de Dirichlet, les différentes façons de le représenter. Puis, on revoit ce qu'est un modèle de mélange et son extension non-paramétrique. Enfin, on revient sur les différentes façons d'estimer un modèle de mélange avec processus de Dirichlet et on présente plus précisément le modèle de l'étude.

Ce Chapitre se base principalement sur E. B. (B. SUDDERTH 2006 ; HJORT, HOLMES, P. MÜLLER et al. 2010 ; GHOSAL et VAN DER VAART 2017 ; MURPHY 2023.

6.2. Définition du processus de Dirichlet

Un modèle non-paramétrique est un modèle pour lequel on considère la dimension de l'espace des paramètres comme infinie. Le passage d'un modèle paramétrique à un modèle non-paramétrique se fait via la spécification de la loi a priori. Pour cela, on considère que les paramètres sont distribués selon un processus de Dirichlet.

On peut représenter un modèle de dimension infinie de différentes façons (ORBANZ et TEH 2010). On s'intéresse dans cette partie à trois façons de représenter le processus de Dirichlet : en regardant la distribution du processus stochastique comme une collection de distributions de dimension finie ; par une représentation explicite qui permet de tirer des distributions du processus ; par une représentation implicite qui permet de tirer des exemples de la distribution. Ces différentes façons seront utiles pour pouvoir utiliser en pratique le modèle.

6.2.1. Définition formelle

On passe avec le bayésien non-paramétrique à des espaces infinis. On doit poser une loi a priori le permettant. Cela se fait via un processus stochastique, une distribution de probabilité sur un ensemble de variables aléatoires. Quand on souhaite faire de la classification non-supervisée, le processus adapté est le processus de Dirichlet.

Un processus de Dirichlet est une distribution sur les distributions (comme le processus gaussien était une distribution sur les fonctions). Il est paramétrisé par un paramètre de concentration α et une distribution de base G_0 , une distribution

6. Modélisation bayésienne – 6.2. Définition du processus de Dirichlet

sur un espace Θ . Une variable que l'on tire dans un processus de Dirichlet est une distribution sur Θ . On note $G \sim DP(\alpha, G_0)$ une distribution aléatoire G tirée dans un processus de Dirichlet. Pour cela, on considère G comme une mesure aléatoire sur un ensemble Θ , *i.e.* une application qui assigne des valeurs à des sous-ensembles $T \subseteq \Theta$ en satisfaisant les règles usuelles des probabilités et la σ -additivité (image comprise entre 0 et 1, image de l'univers égale à 1, union disjointe d'un nombre dénombrable de sous-ensembles égale à la somme des images).

Dans la suite de ce chapitre, on considère Θ comme un espace Polonais, *i.e.* un espace topologique qui est un espace métrique séparable et complet par rapport à une métrique qui définit sa topologie.

Définition 6.2.1 (Processus de Dirichlet). *Soit (T_1, \dots, T_K) une partition finie et mesurable de l'espace Θ , *i.e.*, (T_1, \dots, T_K) sont des ensembles disjoints dont l'union est Θ .*

Pour une mesure aléatoire G , soit $(G(T_1), \dots, G(T_K))$ le vecteur de probabilités des éléments de la partition.

G possède une distribution de processus de Dirichlet, que l'on note $G \sim DP(\alpha, G_0)$, si, pour toute partition finie et mesurable de Θ ,

$$(G(T_1), \dots, G(T_K)) \sim Dir(\alpha G_0(T_1), \dots, \alpha G_0(T_K)),$$

loi de Dirichlet de dimension K .

Le processus $DP(\alpha, G_0)$ est ainsi une loi a priori sur les mesures de probabilité G qui, pour toute partition finie, a la distribution jointe du vecteur de probabilités associé distribué selon une loi de Dirichlet. Comme le processus gaussien, le processus de Dirichlet est défini implicitement par un ensemble de distributions de dimension finie, ici par la distribution de G projetée sur toute partition finie.

Quand on équipe les mesures de probabilité G avec pour loi a priori un processus de Dirichlet, on spécifie deux paramètres : le paramètre de concentration α , un nombre réel positif, et G_0 , généralement une mesure aléatoire standard sur Θ . Le paramètre de concentration α est la masse totale de la mesure de base. On a donc $\alpha = G_0(\Theta)$. Il mesure à quel point le processus de Dirichlet G est concentré autour de G_0 . Il joue un rôle de variance inverse. Plus α est grand, plus on concentre la masse de probabilité de $G(T)$ autour de $G_0(T)$.

Les deux permettent de définir les moments de G .

Proposition 6.2.1 (Moments). *Pour $G \sim DP(\alpha, G_0)$ et tout ensembles mesurables T_1 et T_2 ,*

$$\begin{aligned} \mathbb{E}[G(T_1)] &= G_0(T_1), \\ \mathbb{V}[G(T_1)] &= \frac{G_0(T_1)(1 - G_0(T_1))}{1 + \alpha}, \\ \text{Cov}[G(T_1), G(T_2)] &= \frac{G_0(T_1 \cap T_2) - G_0(T_1)G_0(T_2)}{1 + \alpha}. \end{aligned}$$

6. Modélisation bayésienne – 6.2. Définition du processus de Dirichlet

Enfin, la propriété qui rend le processus de Dirichlet si utile en pratique est qu'il est une loi a priori conjuguée : pour une loi a priori de Dirichlet sur G et des observations de G , la distribution de la loi a posteriori sur G est encore de Dirichlet. La loi a posteriori met à jour les paramètres : le paramètre de concentration est $\alpha + N$ et la mesure de base est une combinaison convexe de la mesure originale et de la distribution empirique des observations.

Proposition 6.2.2. *Soit $G \sim DP(\alpha, G_0)$ une mesure aléatoire distribuée selon un processus de Dirichlet. Pour N observations indépendantes $\bar{\theta}_i \sim G$, la mesure a posteriori suit aussi un processus de Dirichlet*

$$p(G|\bar{\theta}_1, \dots, \bar{\theta}_N, \alpha, G_0) = DP\left(\alpha + N, \frac{1}{\alpha + N}(\alpha G_0 + \sum_{i=1}^N \delta_{\bar{\theta}_i})\right).$$

La forme conjuguée du processus est une conséquence de la définition 6.2.1, qui définit la distribution jointe des mesures de probabilité des éléments d'une partition finie de l'espace comme une distribution de Dirichlet, qui est une loi a priori conjuguée. Comme on a un vecteur qui suit une distribution de Dirichlet, on peut utiliser les propriétés de conjugaison de la distribution pour construire la loi a posteriori de G . Pour une partition de l'espace Θ , on a, sous la loi a priori de processus de Dirichlet, un vecteur qui suit une distribution de Dirichlet, et on peut alors utiliser la conjugaison pour la partition en question.

La réalisation d'un processus de Dirichlet est discret avec probabilité 1, même quand la mesure de base G_0 est absolument continue. En revanche, son support est très large. Tant que la mesure de base est supportée sur tout l'ensemble Θ , le support du processus est l'ensemble de toutes les mesures de probabilité sur Θ . Le fait que le processus de Dirichlet n'échantillonne que des distributions discrètes le rend très utile pour des applications de classification.

6.2.2. Stick-breaking

La définition précédente du processus de Dirichlet est formelle et n'indique pas comment tirer la mesure aléatoire G ou comment tirer des échantillons de G . La construction *stick-breaking* permet de répondre à la première question, en proposant une représentation explicite et constructive du processus de Dirichlet.

Les mesures de probabilité G tirées d'un processus de Dirichlet sont des mesures discrètes avec probabilité un, de la forme

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}. \quad (6.1)$$

La mesure aléatoire G est constitué d'une infinité d'atomes, le $k^{\text{ième}}$ atome est localisé en θ_k et a une pondération π_k . Pour un processus de Dirichlet, la localisation θ_k des atomes est tirée indépendamment, suivant la mesure de base G_0 . Le paramètre de concentration va ici contrôler la distribution des poids π_k . Comme G est une

6. Modélisation bayésienne – 6.2. Définition du processus de Dirichlet

mesure aléatoire, une séquence de poids (π_1, π_2, \dots) somme à 1. La simulation d'une de ces séquences se fait via le processus de cassage de bâton, *stick-breaking*.

On commence par un bâton de taille 1, représentant la masse de probabilité totale. On casse ce bâton séquentiellement en suivant une loi $Beta(1, \alpha)$, chaque morceau ainsi cassé formant π_k . On réapplique un tirage dans la loi $Beta(1, \alpha)$ et on casse la fraction restante du bâton. Pour $k = 1, 2, \dots$,

$$\begin{aligned} \beta_k &\sim Beta(1, \alpha), \\ \theta_k &\sim G_0, \\ \pi_k &= \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell) = \beta_k (1 - \sum_{\ell=1}^{k-1} \pi_\ell). \end{aligned} \tag{6.2}$$

Les points π_k sont associés aux points θ_k en cassant un bâton de longueur unité aléatoirement en une infinité de fragments. On commence par casser le bâton à β_1 et on assigne la masse β_1 à un point aléatoire $\theta_1 \sim G_0$. La masse restante $(1 - \beta_1)$ est divisée à β_2 . On assigne la masse $\beta_2(1 - \beta_1)$ à un point aléatoire $\theta_2 \sim G_0$. Le processus continue une infinité de fois pour assigner toute la masse à un nombre dénombrable de points. Le processus qui en résulte est un processus de Dirichlet $DP(\alpha, G_0)$. On a donc $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \sim DP(\alpha, G_0)$. On peut trouver une preuve chez GHOSAL et VAN DER VAART 2017. On note la distribution sur les poids

$$\pi \sim GEM(\alpha)$$

d'après Griffiths, Engen et McCloskey. La figure 6.1 illustre la construction *stick-breaking*.

Cette représentation du processus de Dirichlet nous permet d'en générer de manière approximative. On peut donc en pratique utiliser un processus de Dirichlet même si les expressions analytiques ne sont pas disponibles, en calculant les quantités a posteriori à partir de simulation de leur distribution a posteriori. En fixant un nombre d'étapes fini, on peut simuler une mesure aléatoire et traiter le problème comme un problème paramétrique, pour lequel les méthodes existent.

À noter, d'autres mesures aléatoires pourraient être construites à partir de la représentation *stick-breaking*, en remplaçant $Beta(1, \alpha)$ par d'autres possibilités. Dans le cas de l'introduction de covariables, on pourrait introduire une dépendance entre les mesures aléatoires en permettant par exemple une dépendance entre les points θ_k ou les poids π_k .

6.2.3. Processus du Restaurant Chinois

La représentation *stick-breaking* nous permet donc de tirer une mesure aléatoire du processus. Il reste à répondre à la deuxième question, comment tirer des échantillons de G . Le fait que le processus de Dirichlet tire des mesures de probabilité discrètes et qu'il soit conjugué va être utile pour construire un modèle d'urne de Pólya pour la distribution prédictive du processus. Celui-ci est à la base du processus du restaurant

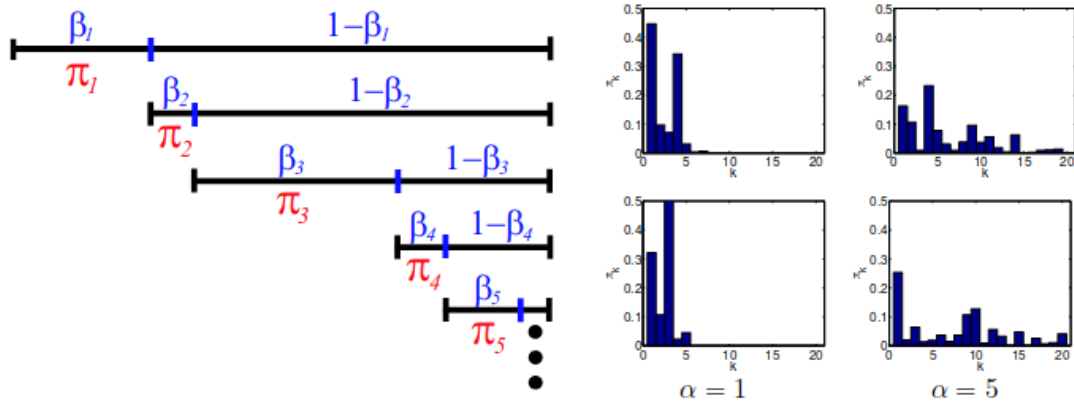


FIGURE 6.1. – Illustration de la construction *stick-breaking* de l'ensemble infini de poids $\pi \sim GEM(\alpha)$ correspondant à la mesure $G \sim DP(\alpha, G_0)$. À gauche, un bâton de longueur un est cassé à un point aléatoire β_1 tiré dans $Beta(1, \alpha)$. La longueur de la partie cassée produit π_1 . On casse récursivement le bout de bâton restant pour produire π_2, π_3, \dots . À droite, les $K = 20$ premiers poids générés par quatre constructions *stick-breaking* aléatoires, deux pour $\alpha = 1$, deux pour $\alpha = 5$. Image tirée de E. B. (B. SUDDERTH 2006.

chinois. Cette représentation permet de tirer des échantillons suivant un G lui-même tiré suivant $DP(\alpha, G_0)$. Il joue donc un rôle de première importance pour les méthodes impliquant un processus de Dirichlet.

Pour une observation $\bar{\theta}_1$ d'une mesure aléatoire d'un processus de Dirichlet G , la probabilité que $\bar{\theta}_1$ soit à l'intérieur d'un ensemble $T \subseteq \Theta$ est donné par $\mathbb{E}[G(T)] = G_0(T)$, selon la définition 6.2.1. La première observation est donc distribuée comme la mesure de base du processus

$$p(\bar{\theta}_1 = \theta | \alpha, G_0) = G_0(\theta).$$

Si maintenant on s'intéresse à N observations, comme $DP(\alpha, G_0)$ produit des mesures de probabilité discrètes, il y a une probabilité strictement positive que plusieurs observations $\bar{\theta}_i \sim G$ prennent des valeurs identiques. Pour N observations $\{\bar{\theta}_i\}_{i=1}^N$, on suppose qu'elles prennent $K \leq N$ valeurs distinctes $\{\theta_i\}_{i=1}^K$. Étant donné la définition 6.2.1 de l'espérance d'un processus de Dirichlet et la définition 6.2.2 de la distribution de G , on peut écrire l'espérance a posteriori d'un sous-ensemble $T \subseteq \Theta$

$$\mathbb{E}[G(T) | \bar{\theta}_1, \dots, \bar{\theta}_N, \alpha, G_0] = \frac{1}{\alpha + N} \left(\alpha G_0(T) + \sum_{k=1}^K N_k \delta_{\theta_k}(T) \right),$$

avec, pour $k = 1, \dots, K$

$$N_k = \sum_{i=1}^N \mathbf{1}\{\bar{\theta}_i = \theta_k\}.$$

6. Modélisation bayésienne – 6.2. Définition du processus de Dirichlet

N_k est donc le nombre d'observations précédentes égales à θ_k et K est une variable aléatoire.

À partir de cette expression, on a un schéma d'échantillonnage par urne de Pólya qui nous permet de produire des échantillons d'une mesure aléatoire d'un processus de Dirichlet. On peut en effet caractériser la distribution prédictive de la prochaine observation $\bar{\theta}_{N+1} \sim G$

$$p(\bar{\theta}_{N+1} = \theta | \bar{\theta}_1, \dots, \bar{\theta}_N, \alpha, G_0) = \frac{1}{\alpha + N} \left(\alpha g(\theta) + \sum_{k=1}^K N_k \delta_{\theta_k} \right), \quad (6.3)$$

où g est la densité de la mesure de base G_0 . Un processus de Dirichlet a ainsi une distribution prédictive qui peut être évaluée et interprétée comme un modèle d'urne de Pólya.

En assignant à chaque observation $\bar{\theta}_i$ une valeur θ_k , le processus de Dirichlet fait de manière implicite une partition des données. Si on ajoute maintenant des variables discrètes (z_1, \dots, z_N) qui indiquent le sous-ensemble auquel est associée l'observation i , tel que $\bar{\theta}_i = \theta_{z_i}$, on peut dissocier la structure de la partition, contrôlée par α , des paramètres des classes, contrôlés par G_0 . Si on remplace dans l'équation (6.3) les $\bar{\theta}$ par les z ,

$$p(z_{N+1} = z | z_1, \dots, z_N, \alpha) = \frac{1}{\alpha + N} \left(\sum_{k=1}^K N_k \mathbf{1}\{z = k\} + \alpha \mathbf{1}\{z = k^+\} \right), \quad (6.4)$$

où k^+ indique une nouvelle classe vide. On a ici le processus du restaurant chinois, une représentation équivalente du processus de Dirichlet, que l'on représente à la Figure 6.2, appelé ainsi d'après l'analogie suivante.

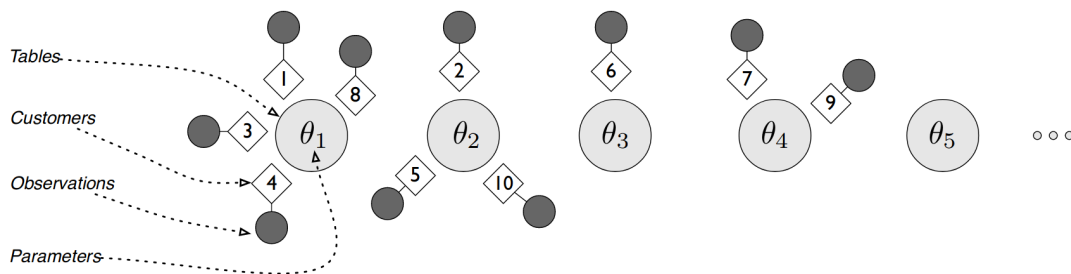


FIGURE 6.2. – Illustration du processus du restaurant chinois. Les diamants représentent les clients, associés à l'observation correspondante, les cercles sombres. Les grands cercles représentent les tables, qui sont donc les classes, ainsi que les paramètres θ associés. Ces derniers ne font pas partie du processus du restaurant chinois en tant que tel mais plutôt de modèle du mélange. Image tirée de GERSHMAN et D. M. BLEI 2012.

Soit un restaurant avec un nombre infini de tables. Les clients correspondent aux observations, les tables aux classes. Chaque table a un dessert particulier, ce dessert

correspond au paramètre θ de la classe. Le premier client arrive et s'assoit à la première table. Quand le deuxième client entre dans le restaurant, il a le choix de s'asseoir à une table déjà ouverte, ou bien de s'asseoir à une nouvelle table inoccupée. Les clients sont pro-sociaux et choisissent de rejoindre une table déjà occupée avec une probabilité proportionnelle aux nombres de clients déjà assis à la table (il s'assied à la table k avec une probabilité proportionnelle à N_k). Sinon, il s'assied à une nouvelle table avec une probabilité α et commande un nouveau dessert en tirant dans la mesure de base G_0 . Formellement, si z_i est l'assignement de l'observation i à une classe, on a

$$p(z_i = k | z_{1:i-1}) \propto \begin{cases} \frac{N_k}{i-1+\alpha} & \text{si } k \leq K_+, \text{ i.e., } k \text{ est une classe existante} \\ \frac{\alpha}{i-1+\alpha} & \text{sinon, i.e., } k \text{ est une nouvelle classe} \end{cases} \quad (6.5)$$

où N_k est le nombre d'observations dans la classe k , K_+ est le nombre de classes pour lesquelles $N_k > 0$.

La séquence $Z = (z_1, \dots, z_N)$ de l'assignement aux classes est une partition des entiers de 1 à N et le processus du restaurant chinois est une distribution de cette partition. Potentiellement tous les entiers sont une classe, mais tous n'ont pas un élément assigné. Pour cela, on note K_+ le nombre de classes pour lesquelles $N_k > 0$, les autres "existent" mais n'ont pas eu d'observations assignées.

On retrouve α comme paramètre de concentration. Plus α est grand, plus des classes sont créées, avec moins d'observations par classes.

La probabilité de créer de nouvelles tables diminue avec le nombre d'observations mais reste non-nul. Le nombre de tables occupées K croît logarithmiquement avec le nombre d'observations, $\mathbb{E}[K] = \alpha \log N$. On a un phénomène des riches qui s'enrichissent avec les tables occupées qui ont plus de chance de se voir attribuer de nouveaux clients.

Enfin, le processus du restaurant chinois a une propriété d'invariance importante malgré sa séquentialité. C'est une distribution sur une partition, mais une distribution échangeable. La distribution jointe est invariante à l'ordre auquel les observations sont assignées aux classes. Les probabilités de partition sont indépendantes des indices des observations. La numérotation des indices des clusters est arbitraire.

6.3. Modèle de mélange avec processus de Dirichlet

On vient de voir trois façons de représenter un processus de Dirichlet : en regardant la distribution du processus stochastique comme une collection de distribution de dimension finie ; par une représentation explicite qui permet de tirer des distributions du processus ; par une représentation implicite qui permet de tirer des exemples de la distribution. En le plaçant comme loi a priori d'un modèle de mélange, on peut le représenter d'une quatrième façon, comme la limite infinie de son pendant fini (ORBANZ et TEH 2010). On commence par présenter ici le modèle de mélange fini puis on s'intéresse au cas infini. Cela nous permettra de faire le lien dans la section

suivante avec le modèle que l'on utilise dans le Chapitre 7.

6.3.1. Le modèle de mélange classique

Considérons un modèle de mélange fini. Celui-ci prend la forme générale

$$p(\mathbf{x}|\boldsymbol{\pi}, \theta_1, \dots, \theta_K) = \sum_{k=1}^K \pi_k f(\mathbf{x}|\theta_k). \quad (6.6)$$

Chaque composant de mélange est une classe, représentée par une densité de probabilité $f(\mathbf{x}|\theta_k)$, pour laquelle on note la distribution $F(\theta_k)$. On a une distribution par classe, les paramètres de la distribution changeant pour chaque classe. Chaque point x_i est généré en sélectionnant une des K classes selon la distribution multinomiale π , puis en échantillonnant depuis la distribution de la classe choisie,

$$\begin{aligned} z_i &\sim \pi, \\ x_i &\sim F(\theta_{z_i}). \end{aligned} \quad (6.7)$$

La variable latente $z_i \in \{1, \dots, K\}$ indique la classe unique associée à chaque observation. On retrouve l'utilité des modèles de mélange pour faire de la classification, où l'attribution aux classes est utilisée pour faire une partition des données.

On choisit souvent $f(\mathbf{x}|\theta_k)$ dans la famille exponentielle. Dans le cadre de cette thèse, on choisira f gaussien. Ce choix est adapté à nos données, des vocalisations que l'on a projetées dans un espace de représentation adapté, et permettra de plus de faciliter les calculs selon les lois a priori que l'on choisit en faisant ressortir des formes conjuguées.

Comme on choisit des noyaux gaussiens pour notre mélange, les paramètres de chacun correspondent à la moyenne et à la variance de chaque classe, $\theta_k = (\mu_k, \Sigma_k)$. On va vouloir estimer sur nos données les valeurs de ces paramètres afin de caractériser chaque classe de vocalisation. On pose une loi a priori sur chacun de ces paramètres, que l'on note G_0 , qui prend un hyper-paramètre λ . On a donc

$$\theta_k \sim G_0(\lambda) \quad (6.8)$$

les paramètres de chaque classe distribués selon une loi G_0 . On peut par exemple choisir une loi a priori conjuguée et dire que θ_k suit une loi normale Wishart inverse pour $k = 1, \dots, K$.

On pose aussi une loi a priori sur les poids de mélange $\boldsymbol{\pi}$. Si l'on n'a pas d'information a priori sur les différentes classes, on choisit une loi de Dirichlet symétrique avec une précision α

$$\boldsymbol{\pi} \sim Dir\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right). \quad (6.9)$$

On a ainsi N points $\{x_i\}_{i=1}^N$, distribués dans K classes, paramétrisés par $\{\theta_k\}_{k=1}^K$.

6. Modélisation bayésienne – 6.3. Modèle de mélange avec processus de Dirichlet

On écrit la distribution jointe des données produit de la vraisemblance et l'a priori

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}) = \prod_{k=1}^K G_0(\theta_k) \prod_{i=1}^N f(x_i | \theta_{z_i}) p(z_i). \quad (6.10)$$

On a fait l'hypothèse que les observations sont indépendantes sachant l'assignement à la classe afin de faire le produit sur N . Étant donné un ensemble de données, on s'intéresse à sa partition, *i.e.*, à l'assignement de chaque observation à une classe. On le trouve en calculant la loi a posteriori par la règle de Bayes

$$p(\mathbf{z} | \mathbf{x}) = \frac{p(\mathbf{x} | \mathbf{z}) p(\mathbf{z})}{\sum_{\mathbf{z}} p(\mathbf{x} | \mathbf{z}) p(\mathbf{z})}. \quad (6.11)$$

On marginalise sur θ pour avoir la vraisemblance

$$p(\mathbf{x} | \mathbf{z}) = \int_{\theta} \left[\prod_{i=1}^N f(\mathbf{x} | \theta_{z_i}) \prod_{k=1}^K G_0(\theta_k) \right] d\theta. \quad (6.12)$$

En ayant pris G_0 une loi Gaussienne Wishart Inverse, la loi priori des paramètres des classes se conjugue à la vraisemblance des données, ce qui permet de calculer l'intégrale analytiquement. On approxime la loi a posteriori par des méthodes MCMC. Dans ce cas paramétrique, on a spécifié lors de la modélisation une valeur précise à K . On connaît en avance le nombre de classes et on veut une partition de notre ensemble de données sur ce nombre de classes. Mais la tâche que l'on se pose est justement de déterminer cette valeur sur les données. Nous voulons faire un travail de classification non-supervisée qui nous permet de dire, pour un ensemble de données, comment partitionner au mieux cet ensemble en différentes classes.

L'approche non-paramétrique permet de ne pas avoir à spécifier K et à l'estimer sur les données. Pour cela, on considère que K tend vers l'infini. Il existe un nombre infini de classes, mais juste un nombre fini d'entre elles est utilisé pour générer les données observées. La loi a posteriori est alors une distribution sur l'assignement des observations aux classes, les paramètres de chaque classe et en plus de cela, le nombre de classes. De plus, la distribution prédictive a posteriori permet d'assigner de futures observations à des classes non vues jusqu'à présent.

6.3.2. Cas non-paramétrique

Dans le modèle de mélange fini, on a défini une fonction de densité sur les données \mathbf{x}

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k f(\mathbf{x} | \theta_k),$$

où π_k est la proportion de mélange et θ_k sont les paramètres associés au composant k .

6. Modélisation bayésienne – 6.3. Modèle de mélange avec processus de Dirichlet

Cette fonction de densité peut être réécrite comme une intégrale

$$p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})G(\boldsymbol{\theta})d\boldsymbol{\theta},$$

où $G = \sum_{k=1}^K \pi_k \delta_{\theta_k}$ est une distribution de mélange discrète qui encapsule tous les paramètres du modèle. δ_{θ_k} est une distribution de Dirac, un atome, centrée en θ .

À partir de là, on peut réécrire la distribution de mélange en faisant tendre K vers l'infini. Un mélange bayésien non-paramétrique utilise un nombre infiniment dénombrable d'atomes dans sa distribution de mélange,

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}.$$

On a ainsi un modèle de mélange avec un nombre infini de composants. Quand on l'utilise sur un ensemble d'entraînement fini, on utilise seulement un nombre fini de composants. En effet, on associe chaque observation à exactement un seul composant et chaque composant peut être associé à plusieurs observations. Comme les observations sont finies, les composants le sont. On peut ainsi faire de l'inférence en reprenant le nombre d'éléments à utiliser et les paramètres associés à chaque composant.

Concrètement, comme on l'a dit précédemment, le moyen de ne pas avoir à spécifier le nombre de classes K lors de la modélisation réside dans la spécification de la loi a priori de l'assignation aux classes. Dans le cas fini, K est un entier et on pose une distribution de Dirichlet sur π . Dans le cas non-paramétrique, on considère que K tend vers l'infini. On pose une distribution sur des partitions infinies des nombres entiers. Pour cela, on utilise le processus de Dirichlet,

$$\begin{aligned} x_i &\sim F(\bar{\theta}_i), \\ \bar{\theta}_i &\sim G, \\ G &\sim DP(\alpha, G_0). \end{aligned} \tag{6.13}$$

Les autres façons de représenter le processus de Dirichlet que l'on a présenté vont être utiles pour définir le modèle, et surtout pour décrire des algorithmes permettant d'échantillonner depuis le modèle. La construction *stick-breaking* implique, pour $k = 1, 2, \dots$

$$\begin{aligned} G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}, \\ \pi &\sim GEM(\alpha), \\ \theta_k &\sim G_0(\lambda). \end{aligned} \tag{6.14}$$

Enfin, le processus de restaurant chinois, en reprenant la variable z_i pour l'assignation de x_i à sa classe, permet de représenter le processus génératif des données

comme

$$\begin{aligned} z_i &\sim \pi \\ x_i &\sim F(\theta_{z_i}) \end{aligned} \tag{6.15}$$

Pour des assignements aux autres classes z_{-i} fixées, on a la distribution a posteriori de z_i

$$p(z_i | z_{-i}, \mathbf{x}, \alpha, \lambda) \propto p(z_i | z_{-i}, \alpha) p(x_i | \mathbf{x}_{-i}, z, \lambda) \tag{6.16}$$

où on peut utiliser l'équation (6.5) pour le premier élément, et le second élément est la prédiction prédictive a posteriori évaluée au point x_i étant donné la partition actuelle. Elle nécessite de calculer la vraisemblance d'attribution à chaque classe existante et à une nouvelle classe, les deux calculables. La représentation permet donc d'établir un échantillonneur de Gibbs pour un modèle de mélange non-paramétrique, que l'on résume à l'algorithme 3.

Ces représentations du modèle sont donc utiles en pratique. Grâce à la représentation *stick-breaking*, on va contrôler directement la complexité du modèle sur les données. Selon notre échantillon fini, on estimera le nombre de composantes de mélange nécessaire. On va pouvoir utiliser cette représentation pour construire les algorithmes qui déterminent automatiquement le nombre de classes d'un ensemble de données. La représentation du processus du restaurant chinois implique une distribution prédictive qui est aussi utilisée pour la construction d'algorithmes pour nos modèles infinis. L'assignement aux classes du processus induit une partition qui induit elle-même un nombre de classes.

L'algorithme que l'on utilise est une méthode marginale et s'appuie sur la représentation du restaurant chinois. Les méthodes conditionnelles reposent généralement sur la représentation *stick-breaking*. On fait une présentation des différentes méthodes qui existent avant de présenter plus en détail un modèle non-paramétrique dans la section 6.4.2 suivante

6.4. Estimation du modèle

La loi a posteriori d'un modèle bayésien se définit comme la distribution conditionnelle des paramètres sachant les données et les hyper-paramètres. Cette définition ne nécessite pas obligatoirement la formule de Bayes. Bien qu'on la retrouve dans tous les modèles paramétriques, on ne la retrouve pas dans les modèles non-paramétriques. On calcule la loi a posteriori autrement. La conjugaison joue alors un rôle important en pratique.

6.4.1. Les propositions algorithmiques

Étant donné un ensemble de données $\{x_i\}_{i=1}^N$ que l'on modélise par un modèle de mélange non-paramétrique, on veut calculer la distribution a posteriori $p(z_1, \dots, z_N | x_1, \dots, x_N, \alpha, G_0)$, qui nous permet d'avoir la partition des données en un nombre de classes K automa-

tiquement estimé. Il existe plusieurs algorithmes permettant d'estimer un modèle de mélange avec processus de Dirichlet.

En général, on utilise des méthodes MCMC, et plus particulièrement de type Gibbs. On divise les algorithmes de Gibbs en deux groupes : les méthodes marginales ou conditionnelles (PAPASPILIOPOULOS et G. O. ROBERTS 2008). Les méthodes marginales marginalisent analytiquement la mesure aléatoire, alors que les conditionnelles exploitent des statistiques résumant la mesure aléatoire.

Les méthodes marginales ont été les premières développées (ESCOBAR 1994) et sont les plus utilisées encore aujourd'hui (R. M. NEAL 2000). Elles reposent sur des schémas d'urne de Pólya. Le processus de Dirichlet induit une structure de partition sur les données. Mais un échantillon est fini : pour un échantillon de n éléments, on peut avoir au plus n classes. Si le mélange peut avoir théoriquement un nombre infini de classes, *in fine*, seul un nombre fini d'entre elles sera associé aux données. Les variables associées à ces classes non représentées n'ont pas besoin de l'être. On peut donc marginaliser sur ces variables et obtenir une représentation implicite que l'on rend explicite quand il est nécessaire d'échantillonner depuis l'a priori. On intègre la loi a posteriori sur toutes les dimensions sauf un nombre fini d'entre elles, et on estime le modèle sur ce qui reste de dimension et les paramètres du modèle. On supprime ainsi les dimensions non utilisées et on a quelque chose d'utilisable. Les méthodes marginales sont intéressantes parce que simples et le nombre d'éléments à tirer à chaque itération est déterministe et borné. Par contre, l'incertitude a posteriori peut-être compliqué à calculer parce que les méthodes marginales ne produisent pas des réalisations de la distribution a posteriori, mais de son espérance conditionnelle, où l'espérance est prise par rapport à la mesure aléatoire.

Les méthodes conditionnelles ont été développées plus tardivement. Elles s'appuient sur la représentation *stick-breaking*, qui permet de produire des réalisations de la mesure aléatoire. On laisse cette fois-ci la distribution de dimension infinie dans le modèle et on échantillonne un nombre suffisant mais fini de variables aléatoires à chaque itération de la chaîne de Markov de la distribution stationnaire. Si les méthodes conditionnelles sont plus flexibles (*e.g.*, intégration de mesures de probabilité *stick-breaking* plus générale, permettant de modéliser une dépendance dans les données), leur implémentation peut poser des challenges méthodologiques auxquels différentes propositions peuvent être apportées (KALLI, GRIFFIN et WALKER 2011; ARBEL, DE BLASI et PRÜNSTER 2019; CANALE, CORRADIN et NIPOTI 2021).

À noter, dans les deux cas, méthodes marginales ou méthodes conditionnelles, on a besoin de connaître analytiquement le comportement a posteriori de la mesure aléatoire. Dans le cas des méthodes marginales, cela passe par les distributions prédictives, pour les méthodes conditionnelles, par une représentation a posteriori de G . Ainsi, dans les deux cas, même si des solutions peuvent exister, la conjugaison est importante.

Enfin, des méthodes variationnelles existent aussi (D. M. BLEI et Michael I. JORDAN 2004; D. M. BLEI et Michael I. JORDAN 2006).

6.4.2. Retour sur le modèle

Pour terminer cette partie, on présente le modèle que l'on utilise dans le Chapitre 7 suivant et l'algorithme que l'on utilise pour l'estimer.

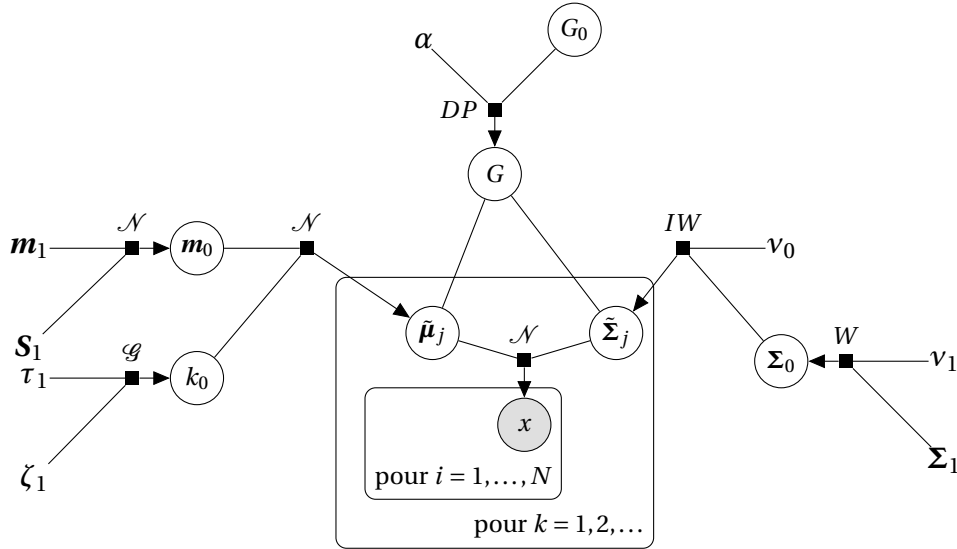


FIGURE 6.3. – Représentation graphique du modèle utilisé pour la classification non-supervisée des voyelles dans le chapitre 7. Modèle de mélange gaussien avec un processus de Dirichlet comme loi a priori sur la mesure aléatoire. La mesure de base est une loi normale Wishart-Inverse. La loi normale prend elle-même deux paramètres, un paramètre de moyenne et un paramètre de variance, sur lesquels on pose une distribution a priori, respectivement une loi normale et une loi Gamma. La loi de Wishart-Inverse prend, elle aussi, deux paramètres. On pose une loi a priori sur la matrice d'échelle, une loi de Wishart. Reste en hyper-paramètres tous les paramètres des hyper-priors, le degré de liberté de la loi de Wishart-Inverse et le paramètre de concentration α du processus de Dirichlet.

On modélise les vocalisations par un mélange de distributions. On utilise un modèle bayésien non-paramétrique pour déterminer le nombre de classes au sein de la distribution des vocalisations sans avoir à les spécifier a priori. Pour un ensemble de vocalisations $X = \{x_i\}_{i=1}^N$ avec $x_i \in \mathbb{R}^p$, on définit un modèle de mélange non-paramétrique par la distribution aléatoire

$$p(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}; \theta) dG(\theta), \quad (6.17)$$

où θ est l'espace des paramètres, f est une densité définie sur $\Theta \times \mathbb{R}^p$ et G est une mesure aléatoire discrète sur Θ . On fait l'hypothèse que G est distribué selon un

processus de Dirichlet. On peut réécrire le modèle sous une forme hiérarchique,

$$\begin{aligned} x_i | \theta_i &\sim f(x_i; \theta_i) & i = 1, \dots, N \\ \theta_i | G &\sim G \\ G &\sim DP(\alpha, G_0). \end{aligned} \quad (6.18)$$

Pour notre modèle, f est une Gaussienne multivariée de dimension p . Chaque composant de mélange est donc paramétrisé par la moyenne et la matrice de variance-covariance de la classe, $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. La densité de mélange est

$$p(\mathbf{x}) = \sum_{j=1}^{\infty} \pi_j f(\mathbf{x}; \tilde{\boldsymbol{\mu}}_j, \tilde{\boldsymbol{\Sigma}}_j). \quad (6.19)$$

Bien que des solutions existent si ce n'est pas le cas, on fait en sorte de choisir des lois a priori conjuguées afin de faciliter l'estimation du modèle. Ainsi, comme on a des noyaux Gaussiens multivariés, on doit choisir une mesure aléatoire de base G_0 sur $\Theta = \mathbb{R}^p \times S_+^p$, où S_+^p est l'espace des matrices $p \times p$ semi-définies positives. La loi a priori conjuguée sur G_0 est une loi normale Wishart Inverse,

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}_j &\sim IW(\nu_0, \boldsymbol{\Sigma}_0) \\ \tilde{\boldsymbol{\mu}}_j | \tilde{\boldsymbol{\Sigma}}_j &\sim \mathcal{N}\left(\mathbf{m}_0, \frac{\boldsymbol{\Sigma}_0}{k_0}\right). \end{aligned} \quad (6.20)$$

On ajoute des hypers a priori sur trois paramètres de nos lois a priori \mathbf{m}_0, k_0 et $\boldsymbol{\Sigma}_0$. On pose $\mathbf{m}_0 \sim \mathcal{N}_p(\mathbf{m}_1, \mathbf{S}_1)$, $k_0 \sim Ga(\tau_1, \zeta_1)$ et $\boldsymbol{\Sigma}_0 \sim W(\nu_1, \boldsymbol{\Sigma}_1)$. ν_0 est le degré de liberté de la loi de Wishart Inverse et se détermine selon p . Pour l'ensemble des six paramètres, $(\mathbf{m}_1, \mathbf{S}_1), (\tau_1, \zeta_1), (\nu_1, \boldsymbol{\Sigma}_1)$ on pose les hyper-paramètres à partir des données. \mathbf{m}_1 reprend les moyennes de chaque dimension de X , \mathbf{S}_1 les variances et covariances, $\tau_1 = \zeta_1 = 1$, $\nu_1 = p + 2$ et $\boldsymbol{\Sigma}_1$ est égale à la covariance de l'échantillon divisée par 2.

On retrouve une représentation hiérarchique du modèle à la Figure 6.3. Étant donné la spécification du modèle, celui-ci est conjugué. On peut l'estimer facilement via l'algorithme 3 de R. M. NEAL 2000.

La représentation du restaurant chinois est ici utile pour construire l'algorithme d'échantillonnage. Pour rappel, le modèle (6.18) est équivalent à un modèle de mélange classique pour lequel on fait tendre K à l'infini. La variable z_i indique la classe latente de l'observation i et θ_k les paramètres qui déterminent la distribution de la classe k . Typiquement, l'échantillonneur de Gibbs que l'on utilise est une extension de la version du modèle paramétrique. L'assignement aux classes $\{z_i\}_{i=1}^N$ produit par l'algorithme nous permet d'avoir une estimation du nombre de classes K présentes dans l'échantillon $\{\mathbf{x}_i\}_{i=1}^N$. Il suffit d'ajouter à l'algorithme la possibilité d'assigner l'observation i à une classe qui n'existe pas encore, ce que l'on a vu dans la partie 6.2.3. La probabilité d'assignement de l'observation i à la classe k est proportionnelle à (6.5). Le fait d'avoir choisi par ailleurs une forme conjuguée entre la vraisemblance et la

Algorithm 3 Échantillonneur de Gibbs pour modèle de mélanges bayésiens non-paramétriques

Require: Vecteur d'assignation aux classes $z = \{z_i\}_{i=1}^N$.

Require: Paramètres $\{\theta_k\}_{k=1}^K$ des K classes courantes.

for $i = 1, \dots, N$ **do**

Faire une permutation des données

for $k = 1 \dots, K$ **do**

Calculer la prédictive a posteriori de l'assignation à chaque classe

$$p(z_i = k | z_{-i}, \mathbf{x}_i) = \frac{N_{k,-i}}{N-1+\alpha} p_k(\mathbf{x}_i).$$

end for

Calculer la prédictive a posteriori de l'assignation à une nouvelle classe

$$p(z_i = k^+ | z_{-i}, \mathbf{x}_i) = \frac{\alpha}{N-1+\alpha} p(\mathbf{x}_i).$$

Tirer une nouvelle assignation z_i depuis $z_i \sim p(z_i | z_{-i}, \mathbf{x}_i)$.

Si $z_i = k^+$, créer une nouvelle classe et incrémenter K . Si des classes sont vides, les enlever et les soustraire à K .

end for

loi a priori nous permet de calculer analytiquement la vraisemblance prédictive de chaque classe. Le calcul des paramètres associés à chaque classe peut être fait. L'algorithme 3 permet l'estimation d'un modèle alors même que celui-ci a un nombre infini de paramètres, en étendant l'algorithme du cas fini. À noter, dans le cas où les lois a priori sont conjuguées, les vraisemblances sont calculables analytiquement. Des solutions existent si ce n'est pas le cas, chez R. M. NEAL 2000 avec son algorithme 8.

6.4.3. Détermination de la partition

Enfin, une fois l'estimation faite, un modèle bayésien non-paramétrique ne propose pas une solution unique pour partition. En effet, le modèle propose une distribution a posteriori sur l'ensemble des partitions possibles. Cela nous permet d'avoir une quantification de l'incertitude, mais cela demande aussi une façon de résumer la loi a posteriori afin d'avoir une valeur claire sur le nombre de classes que l'on trouve.

Si le fait d'avoir une distribution via la loi a posteriori est une richesse que l'on ne retrouve pas avec une estimation ponctuelle, il est néanmoins utile d'avoir une valeur qui nous permette de résumer l'information du modèle. On utilise alors souvent la loi a posteriori pour calculer un estimateur ponctuel, par exemple le maximum a posteriori. Ayant la distribution, il est aisé d'y joindre un intervalle de confiance, ou de crédibilité dans le vocabulaire bayésien, afin d'avoir une idée de l'incertitude de cette estimation.

Dans notre cas, l'estimateur qui nous intéresse est celui de K , le nombre de classes.

Combien de groupes de mélange avons-nous finalement dans notre distribution, quelle est la meilleure façon de partitionner notre espace de vocalisations? On suit pour cela la méthode proposée par WADE et GHAHRAMANI 2018.

Comme nous avons calculé la loi a posteriori en utilisant des techniques MCMC, nous avons beaucoup de partitions, qui sont autant d'échantillons de la loi a posteriori. Toutes les partitions ne sont pas identiques. L'idée est d'étendre ce qui se fait traditionnellement en statistique bayésienne, prendre une statistique de la distribution a posteriori comme la moyenne par exemple, au cas d'un a posteriori d'une partition.

Pour cela, on se base sur la théorie de la décision. On définit une fonction de perte sur la partition. Le point optimal est celui qui optimise l'espérance a posteriori de la fonction de perte, appelée risque a posteriori. On définit donc une fonction de perte $L(c, \hat{c})$ qui mesure la perte à estimer la vraie partition c par \hat{c} . Comme c est inconnue, le risque a posteriori que l'on minimise est la moyenne sur les partitions possibles, pondérées par la loi a posteriori :

$$c^* = \underset{\hat{c}}{\operatorname{argmin}} E[L(c, \hat{c}) | y_{1:N}] = \underset{\hat{c}}{\operatorname{argmin}} \sum_c L(c, \hat{c}) p(c | y_{1:N}). \quad (6.21)$$

On a la méthode pour déterminer le nombre de classes parmi toutes les propositions que l'on a dans notre loi a posteriori. Il suffit juste de définir la fonction de perte L que l'on utilise. C'est la variation d'information

$$VI(c, \hat{c}) = H(c) + H(\hat{c}) - 2I(c, \hat{c}), \quad (6.22)$$

où H est l'entropie et I est l'information mutuelle.

La fonction de perte (6.22) a plusieurs bonnes propriétés, discutées plus en détail par (WADE et GHAHRAMANI 2018). L'une d'elle est d'être une métrique de l'espace des partitions. Afin d'avoir une quantification de l'incertitude liée à notre estimation, on construit une région de crédibilité autour de l'estimateur, avec la boule centrée en c^* de rayon ϵ^*

$$B_{\epsilon^*}(c^*) = \{c : d(c^*, c) \leq \epsilon^*\}, \quad (6.23)$$

où ϵ^* est le plus petit ϵ tel que $P(B_{\epsilon}(c^*) | x) \geq 1 - \alpha$. On a ainsi la plus petite boule autour de c^* telle que la probabilité a posteriori est au moins $1 - \alpha$. Comme la variation d'information est une métrique de l'espace des partitions, on peut l'utiliser dans l'équation (6.23).

On obtiendra ainsi sur notre problème de partitionnement des vocalisations des bébés un nombre de classes de vocalisations, ainsi qu'une estimation de l'incertitude.

On vient de présenter les modèles de mélange bayésiens non-paramétriques, et plus particulièrement, les modèles avec processus de Dirichlet. On a présenté ces processus et différentes façons de les représenter, afin d'avoir une idée plus claire de ces modèles et de leur utilisation. Nous avons terminé par une présentation plus en détail du modèle que l'on utilise dans l'étude suivante du chapitre 7, qui nous permet de faire le travail de classification non-supervisée des vocalisations de bébé. Plusieurs solutions déjà implémentées existent sur R : DPpackage, PReMiuM, BNPdensity,

6. Modélisation bayésienne – 6.4. Estimation du modèle

`dirichletprocess`, `BNPMIXcluster`, `msBP`. On utilise `BNPmix` (CORRADIN, CANALE et NIPOTI [2021](#)).

7. Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development

Sommaire

7.1. Introduction	146
7.2. Data	147
7.3. Modelling	148
7.3.1. Topologically augmented signal representation	148
7.3.2. Nonparametric Bayesian modelling for clustering	152
7.3.3. Depth of the clusters for representative vocalizations	154
7.4. Data analysis	154
7.4.1. Partition	154
7.4.2. Comparison of clusters	156
7.5. Conclusion and discussion	160
7.6. Supplementary materials	161

Abstract

We cluster the vocalizations a human child has produced between birth and her first birthday. We use a topologically augmented representation of the vocalizations, based on two persistence diagrams for each vocalization : that of its Taken's embeddings and that of the surface of its spectrogram. A synthetic persistent variable is computed for each diagram, and added to the MFCCs. A non-parametric Bayesian mixture model is fitted on this representation. We consider a Dirichlet process prior for modelling the number of component K . We find 8 clusters of vocalization. A typical vocalization is constructed for each cluster by computing the deepest vocalization in each group, that we compare.

Keywords

7. *Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.1. Introduction*

clustering, Bayesian non-parametric, Dirichlet process, mixture model, topologically-augmented machine learning, TDA, babbling, language development

7.1. Introduction

Over the first year of life, the human baby's vocal tract undergoes major changes (BARBIER, BOË, CAPTIER et al. 2015). These changes determine her acoustic space, the set of sounds she will be able to produce. By the end of her first year, the child will be able to produce her first word (Patricia K. KUHL 2004). In the meantime, she follows a developmental path and explores her acoustic space. After cooing sounds, she starts producing quasi-vowel at around 3 months (D. Kimbrough OLLER 2000). These vowels are combined with consonants, leading to babbling at around six months of age (MORGAN et WREN 2018). The syllables are repeated, it is the phase of canonical babbling. With time and practice, the babbling becomes even richer and variegated. The quantity and proportion of babbling increases steadily over time (Margaret CYCHOSZ, CRISTIA, Erika BERGELSON et al. 2021), the complexity of vocalizations progresses, and by the end of the first year we arrive at the first word (BROOKS et KEMPE 2012). All this evolution of verbal production is conditioned by ambient sounds (BOYSSON-BARDIES, SAGART et C. DURAND 1984; S. A. S. LEE, B. DAVIS et MACNEILAGE 2010). The acoustic space of the child's vocal productions approximates that of the ambient language with which she is confronted. She calibrates her productions to this target language (TER HAAR, FERNANDEZ, GRATIER et al. 2021).

Monitoring these pre-language vocal productions is extremely important. Not only does it provide a better understanding of the different phases of language development, but above all, it is predictive of various disorders (D. K. OLLER, R. E. EILERS, A. R. NEAL et al. 1998; D. Kimbrough OLLER, Rebecca E EILERS, A. Rebecca NEAL et al. 1999; BARTL-POKORNY, POKORNY, GARRIDO et al. 2022).

This exploration of acoustic space and calibration to a target language by the child can be found in vocalizations that still predate babbling, such as crying (MAMPE, FRIEDERICI, CHRISTOPHE et al. 2009). Some of these pre-language vocalizations, such as squeaks or grunts, have an important property in human language : functional flexibility (D. Kimbrough OLLER, BUDER, RAMSDALL et al. 2013; JHANG 2017). They may also be predictive of disorders (CABON, MET-MONTOT, POREE et al. 2021; WERMKE, ROBB et SCHLUTER 2021). Moreover, the acoustic structures of these vocalizations such as the cries evolve during development (LOCKHART-BOURON, ANIKIN, PISANSKI et al. 2023). A better understanding of these early phases would open up the possibility of even earlier action. In addition, a more detailed view of babbling would enable us to better grasp the links with possible developmental disorders (PAUL, FUERST, RAMSAY et al. 2011).

The use of new storage and recording tools enables the construction of massive new databases. Coupled with new statistical analysis techniques, we can learn more about these early phases of language development (D. K. OLLER, P. NIYOGI, GRAY et al. 2010; MILLING, POKORNY, BARTL-POKORNY et al. 2022).

7. Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.2. Data

Current analysis of early vocal production focuses on formant analysis. Topological Signal Analysis (TDA) is an interesting candidate for augmenting the current representation of infantile vocalizations. TDA, which is proving its worth in more and more fields, including analysis of fMRI (SALCH, REGALSKI, ABDALLAH et al. 2021) or EEG data (NASRIN, OBALLE, BOOTHE et al. 2019; MAROULAS, MIKE et OBALLE 2019; Xiaoqi XU, DROUGARD et R. N. ROY 2021), has stability theoretic properties that make it useful for studying natural signals (COHEN-STEINER, H. EDELSBRUNNER et HARER 2007; Frederic CHAZAL, Vin DE SILVA, GLISSE et al. 2013). Topological descriptors of data, in complementarity with more classical descriptors, present good results on classification questions (BARNES, POLANCO et PEREA 2021; HENSEL, MOOR et RIECK 2021, Chapter 5). Taking topological information into account could provide useful additional information for refining the description of these vocal productions.

In this work, we propose an unsupervised classification of human baby vocalizations produced between 0 and 12 months. For this purpose, we recorded a child for one year, at home, at a rate of three days per month. Our aim is to propose a categorization that may be finer than the existing one, without specifying a priori the number of existing classes, but estimating them on the data. To do this, we use a non-parametric Bayesian model, a Dirichlet process mixture model. This clustering will be based on a topologically augmented representation of the signal, in order to integrate additional information about the topology of the vocalizations.

In the following, we present the database in section 7.2. We present the computation of the augmented topological representation and the clustering model in section 7.3. We present the classification results in section 7.4, before discussing them and concluding in section 7.5.

7.2. Data

The data set consists of a child’s vocalizations, produced between birth and the child’s first birthday. Each vocalization consists of a stereo-channel audio signal, sampled at 44100 Hertz in PCM format. It is a segment extracted from a longer audio file. We convert stereo to mono, taking the average of both channel, and rescale the pulse modulation signal such that it lies between -1 and 1 .

The vocalizations come from long-form audio recordings made at regular intervals of a child at home by her parents over a one-year period. The parents were equipped with a portable microphone three days a month. During these three days, they recorded as much as possible the child. The result was long audio recordings containing vocalization segments. We followed BONAFOS, PUDLO, FREYERMUTH et al. 2023 to extract vocalization segments from these continuous recordings.

This produced a dataset of 1924 vocalizations. We discarded vocalizations lasting more than 10 seconds, leaving a set of 1851 vocalizations, averaging 2.51 seconds in length. Table 7.1 shows the distribution of vocalizations detected over the first year.

Note that we don’t have vocalizations for every month. In fact, four months are missing : the first, fourth, fifth and tenth. It’s not that the vocalizations of these months

7. Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.3. Modelling

TABLEAU 7.1. – Number of vocalizations per month in the massive audio recordings, as well as the mean and the standard deviation of the duration of the vocalizations produced per month.

	Month							
	2	3	6	7	8	9	11	12
Effective	667	139	132	154	159	285	217	98
Mean duration (secs)	2.09	2.71	3.12	2.81	2.74	2.75	2.55	2.62
Standard deviation (secs)	1.57	2.00	2.12	1.94	2.15	2.03	1.90	2.09

have been left out, but the recordings couldn't be made for them. Necessarily, this lack of sampling means that we lose some of the information.

7.3. Modelling

We have a dataset of baby vocalizations. We want to cluster this set in order to distinguish K different classes of vocalizations. We decide to perform this clustering using a non-parametric Bayesian model, a Dirichlet process mixture model.

We learn the model on a representation of the vocalizations, taking into account the signal's topological characteristics.

7.3.1. Topologically augmented signal representation

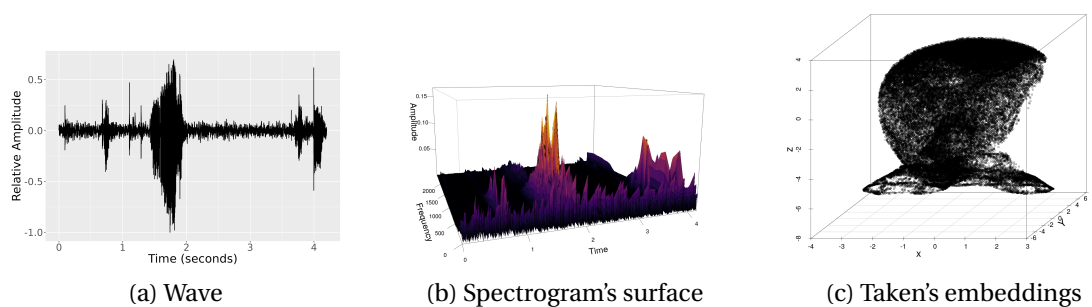


FIGURE 7.1. – Representation of the signal, by the surface of its spectrogram or by its Taken's embeddings. On each of these representations, we compute the persistent homology that we resume in the persistence diagram. We plot here the spectrogram surface and the Taken's embeddings for the deepest vocalization of each of the cluster we found ($N = 8$) with our Dirichlet process mixtures model.

7. Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.3. Modelling

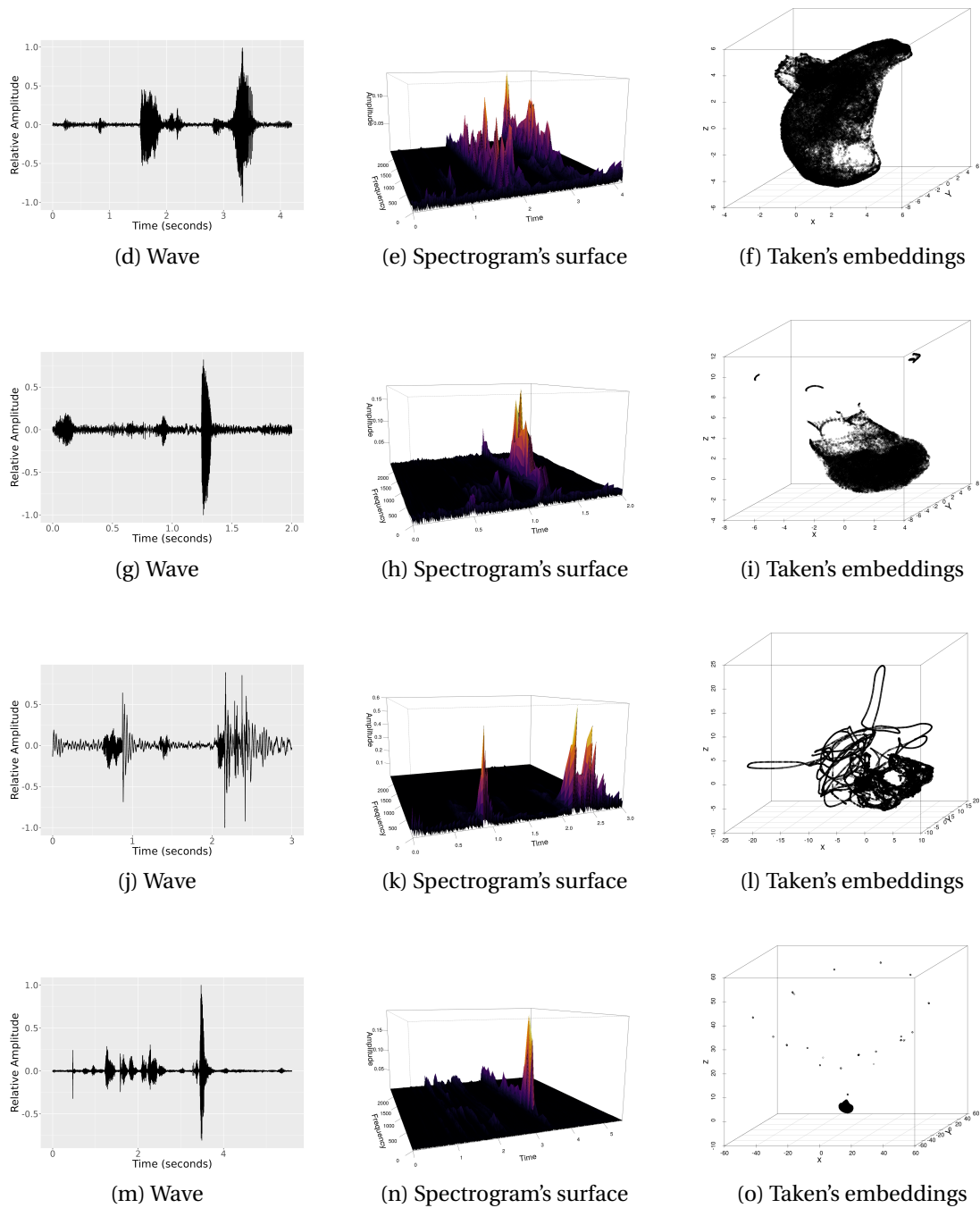


FIGURE 7.1. – Representation of the signal, by the surface of its spectrogram or by its Taken's embeddings. On each of these representations, we compute the persistent homology that we resume in the persistence diagram. We plot here the spectrogram surface and the Taken's embeddings for the deepest vocalization of each of the cluster we found ($N = 8$) with our Dirichlet process mixtures model (continuation).

7. Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.3. Modelling

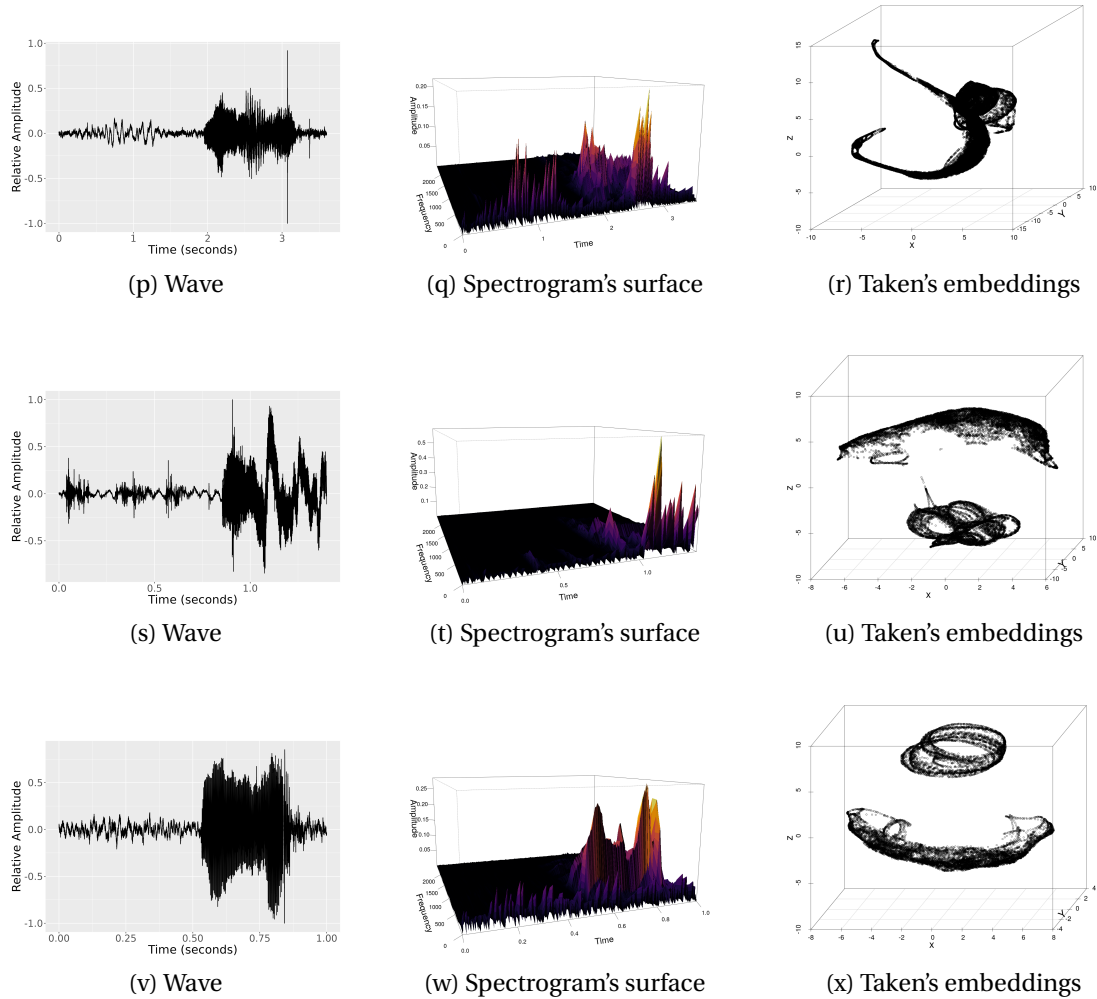


FIGURE 7.1. – Representation of the signal, by the surface of its spectrogram or by its Taken's embeddings. On each of these representations, we compute the persistent homology that we resume in the persistence diagram. We plot here the spectrogram surface and the Taken's embeddings for the deepest vocalization of each of the cluster we found ($N = 8$) with our Dirichlet process mixtures model (continuation).

7. Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.3. Modelling

The representation of the signal is the first step of the statistical analysis. Depending on what the representation highlights, we could discover or not some pattern. We want to have a topologically augmented signal representation.

We consider the data as sampled from a manifold. The manifold hypothesis is a common working hypothesis in machine learning, based on physical constraints (GOODFELLOW, Y. BENGIO et COURVILLE 2016; FEFFERMAN, MITTER et NARAYANAN 2016; BERENFELD 2022). This hypothesis motivates the use of Topological Data Analysis (TDA), which captures properties of data manifold (TROFIMOV, CHERNIAVSKII, TULCHINSKII et al. 2023). We are analyzing vocalizations from birth to first birthday, a period during which the child develops motor control of the buco-phonatory apparatus. This motor control and its evolution should govern the data manifold. Then, we want to integrate topological features of the underlying manifold.

TDA assumes that data has a shape (A. J. ZOMORODIAN 2005; CARLSSON 2009). We build this shape through a filtration, a nested family of simplicial complexes (Frédéric CHAZAL et MICHEL 2021). We derive from the filtration the persistent homology, topological descriptors of the data. In our case, we can represent a vocalization in different spaces. Depending on the representation space, we adapt the filtration to compute the persistent homology.

For each record, we compute the spectrogram, with a Gaussian window of 11.6 ms and an overlap of 90%. We project the spectrogram $S(t, \omega) = |F(t, \omega)|^2$, where $F(t, \omega)$ is the Short Time Fourier Transform, onto the z -axis to define a surface in \mathbb{R}^3 . The dimensions are time t , frequency ω and amplitude S . We apply a sublevel set filtration to compute the persistent homology of the spectrogram, *i.e.*, for $f : S \mapsto \mathbb{R}$, we compute a nested sequence of topological spaces $S_r = f^{-1}(-\infty, r]$ for increasing value of r .

We also compute the Taken's embeddings (TAKENS 1981). We estimate the time delay parameter τ such that $AMI(\tau) < \frac{1}{e}$, where AMI is the Average Mutual Information. We estimate the dimension of the embeddings D with the Cao's algorithm (L. CAO 1997). We reduce the dimension D to 3 for all embeddings using UMAP (MCINNES, HEALY et MELVILLE 2020) so that all embeddings have the same dimension. Thus, we represent the vocalization as a point cloud $\mathcal{P}_D = \{p_1, \dots, p_D\} \subset \mathbb{R}^D$ with

$$p_i = (x_i, x_{i+\tau}, x_{i+2\tau}, \dots, x_{i+(D-1)\tau}).$$

We apply an Alpha filtration to compute the persistent homology of the embeddings, *i.e.*, we compute a nested family of Alpha complex

$$Alpha(r) = \{\sigma \subseteq \mathcal{P} \mid \bigcap_{p \in \sigma} R_x(r) \neq \emptyset\},$$

where $R_x(r)$ is the intersection of each Euclidean ball with its corresponding Voronoi cell, for increasing value of r .

We present in Figure 7.1 the two representation spaces of the same signal, for which we compute persistent homology. For both objects, we have an increasing sequence of topological spaces. We compute the homology at all scale, *i.e.*, for all r of the sequence. We resume in a persistence diagram the persistence homology of the object, where a

point in a diagram has to coordinate the value of r of its birth and the value of r of its death. With the persistent homology, we have a multiscale topological description of the object (WASSERMAN 2018).

The space of the persistence diagrams is not a space from which statistical analysis is trivial. We extract information from the diagrams computing a set of variables from them : persistent entropy (ATIENZA, GONZALEZ-DIAZ et RUCCO 2019), the p -norm of the diagram (COHEN-STEINER, H. EDELSBRUNNER, HARER et MILEYKO 2010), the persistent Betti number (H. EDELSBRUNNER et HARER 2009), descriptors of the vector collecting the lifetime of the points of the diagram following FIREAIZEN, RON et BOBROWSKI 2022 and PEREIRA et DE MELLO 2015.

From this set of variables, computed for each diagram, we compute a synthetic persistent variable using PCA. We keep the first principal component of the PCA as synthetic persistent variable, explaining 27.79% of the variance of the set of the variables from the persistence diagram of the surface of the spectrogram, and 65.94% of the variance of the set of the variables from the persistence diagram of the Taken's embeddings.

Plus, we compute the Mel Frequency Cepstral Coefficients (MFCC), classical frequency descriptors of human speech analysis (CHACHADA et KUO 2014; SUEUR 2018). We compute twelve coefficients, with a window length of 25 milliseconds and an overlap of 40%. We take the average of the twelve coefficients to have the same number of MFCC for all the vocalizations.

We end up with a topologically augmented representation of the vocalizations, composed by fourteen dimensions, 12 MFCC and 2 synthetic persistent variables, one resuming the persistence diagram computed on the surface of the spectrogram, another resuming the persistence diagram computed on the Taken's embeddings.

7.3.2. Nonparametric Bayesian modelling for clustering

Let $\mathbf{x}_i = (x_1, \dots, x_t) \in X$ be a vocalization. We build a 14 dimensional representation of the vocalizations of the set by computing the variables presented in the previous section. We now want to determine, from this representation, how many clusters there are in the set.

To this aim, we use a Dirichlet process mixture model. We define a mixture model by the random distribution

$$p(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}; \theta) dG(\theta), \quad (7.1)$$

where Θ is the space of parameters, f is a kernel defined on $\Theta \times \mathbb{R}^p$ and G is a random discrete probability measure on Θ .

Because of the representation of \mathbf{x} , the dimension of each representation of a vocalization is $p = 14$. We consider for our model f as a p -dimensional Gaussian kernel. Consequently, $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\Theta = \mathbb{R}^p \times S_+^p$, where S_+^p is the space of semi definite positive $p \times p$ matrices.

7. Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.3. Modelling

We are interested in the number K of mixture components and the assignment of vocalizations to components. We can learn the complexity of the model on the data, and thus partition the dataset, by setting a Dirichlet process as prior on G . We write the model

$$\begin{aligned} \mathbf{x}_i | \theta_i &\sim f(\mathbf{x}_i; \theta_i) & i = 1, \dots, N \\ \theta_i | G &\sim G \\ G &\sim DP(\alpha, G_0). \end{aligned} \tag{7.2}$$

We do not specify the complexity of the model : it is going to grow and fit the data. It is equivalent to put an infinite dimensional prior (HJORT, HOLMES, P. MÜLLER et al. 2010; GHOSAL et VAN DER VAART 2017). We will select a finite number of dimension given the dataset. We can write the mixture density as

$$p(\mathbf{x}) = \sum_{j=1}^{\infty} \pi_j f(\mathbf{x}, \mu_j, \Sigma_j), \tag{7.3}$$

For the precision parameter of the Dirichlet process, we choose α such that $\mathbb{E}[K|n, \alpha] = \sum_{i=1}^n \frac{\alpha}{\alpha+i-1}$ (ANTONIAK 1974; DORAZIO 2009). Consequently, we put a prior on the number of clusters we are expecting in the set. Following the literature (Meg CYCHOSZ, SEIDL, Erika BERGELSON et al. 2019), we expect $K = 5$. We put as prior on the base measure G_0 a normal-inverse Wishart,

$$\begin{aligned} \Sigma_j &\sim IW(\nu_0, \Sigma_0) \\ \mu_j | \Sigma_j &\sim \mathcal{N}\left(\mathbf{m}_0, \frac{\Sigma_0}{k_0}\right). \end{aligned} \tag{7.4}$$

With this prior, we have conjugacy. We take for the degree of freedom of the inverse Wishart a non-informative prior $\nu_0 = p$. We add as hyperpriors

$$\begin{aligned} \mathbf{m}_0 &\sim \mathcal{N}(\mathbf{m}_1, \mathbf{S}_1) \\ k_0 &\sim \mathcal{Gamma}(\tau_1, \xi_1) \\ \Sigma_0 &\sim W(\nu_1, \Sigma_1), \end{aligned} \tag{7.5}$$

with the hyperparameters computed on the dataset : \mathbf{m}_1 is the mean of each dimension of X , \mathbf{S}_1 the variance-covariance matrix, $\tau_1 = \xi_1 = 1$, $\nu_1 = p + 2$ and $\Sigma_1 = \mathbf{S}_1 / 2$.

We use the collapsed Gibbs sampler of R. M. NEAL 2000, based on the Chinese Restaurant Process representation, to sample the indicator variable $z = \{z_i\}_{i=1}^N$, which assigns each vocalization to a latent cluster, by marginalizing mixture weights and parameters. This assignment gives us the clustering. We run the MCMC with 10,000 iterations and discard the first 4,000 ones as burn-in.

Being a Bayesian model, the posterior gives a distribution on the possible clustering of the dataset and not just a single point estimate. We select the best clustering following WADE et GHAHRAMANI 2018.

7.3.3. Depth of the clusters for representative vocalizations

Finally, once we have the partition, we compare the different clusters. To do so, we compute the depth of all points using Mahalanobis depth function (SERFLING et ZUO 2000), and we use the deepest point of each cluster as the representative of its cluster. The deepest point is like the multivariate median (MOSLER 2013). This most central point of the cluster (MOSLER et MOZHAROVSKIY 2022) should give us a good example of the cluster detected.

7.4. Data analysis

7.4.1. Partition

We find with our model 8 different clusters. We resume the size of each cluster, as well as some descriptive statistics about the duration of the vocalizations composing the cluster, in Table 7.2. We learn the manifold of the vocalizations on their topologically augmented representation with UMAP (MCINNES, HEALY et MELVILLE 2020) and we project the clustering in Figure 7.2.

TABLEAU 7.2. – Number of vocalizations per cluster, as well as the mean and the standard deviation of the duration of the vocalizations of the cluster.

	Cluster							
	1	2	3	4	5	6	7	8
Effective	360	252	229	13	370	45	511	66
Mean duration (secs)	2.01	3.23	2.58	6.92	1.72	3.76	2.87	2.06
Standard deviation (secs)	1.57	2.24	2.10	1.43	1.01	2.80	1.85	1.26

Initially, we found 9 clusters, but one of these clusters were only composed by 5 records, and none of them included baby vocalizations. This cluster served as a "garbage cluster" and grouped together false positives still here in the dataset. Interestingly, this class was separated of the manifold that we learn using the topologically augmented representation of the recordings. We represent the initial manifold, with this cluster composed by non-vocalizations, in the Figure S7.1, in the supplementary materials. Learning the latent space of the baby vocalizations on their topologically augmented representations through UMAP reveals a connected manifold, composed by the vocalizations of the baby, but if we project other sounds in this space, they are not near the manifold, as it is the case for this "garbage" cluster. Moreover, based on this representation, the Dirichlet process mixture model is able to group together recordings that are different from the rest of the dataset. It automatically detects and creates a "garbage" class.

7. Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.4. Data analysis

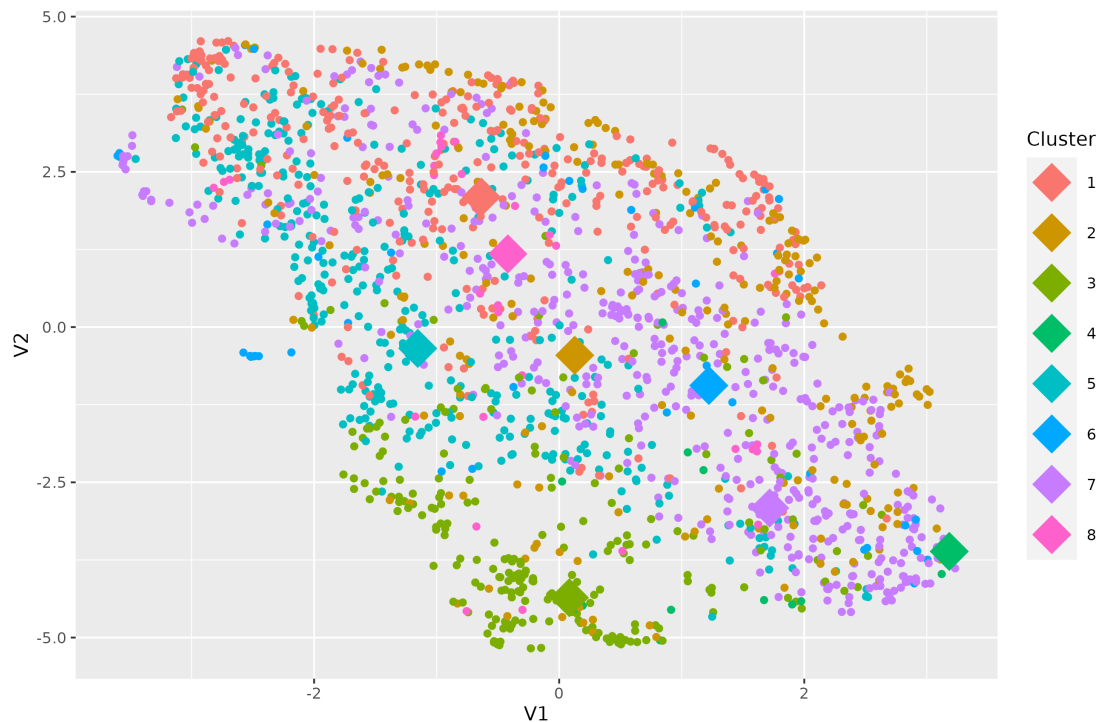


FIGURE 7.2. – Manifold of the vocalizations of the baby produced during one year, learned on the topologically augmented representation of the vocalizations, on which we project the clustering. We highlight the deepest vocalization of each cluster. Each point is a vocalization, each color a cluster. The diamonds correspond to the deepest vocalization of each class.

We represent the final partition, composed by the eight clusters we have detected, in Figure 7.2. We add on the figure the deepest point of each cluster. Once we project the clustering on the manifold, we can see the distribution of the clusters in this space. Cluster 1 and cluster 8 are quite close. We expect a lot of similarity between them. On the contrary, cluster 3 is the farthest, with a relatively concentrated distribution. It is the same for cluster 4, for which the distribution is also narrow, but this cluster is not as far from the others as cluster 3 and is smaller. We expect cluster 3 and cluster 4 to have more discriminant features. Cluster 7, on the other hand, which is also the largest, has a highly dispersed distribution.

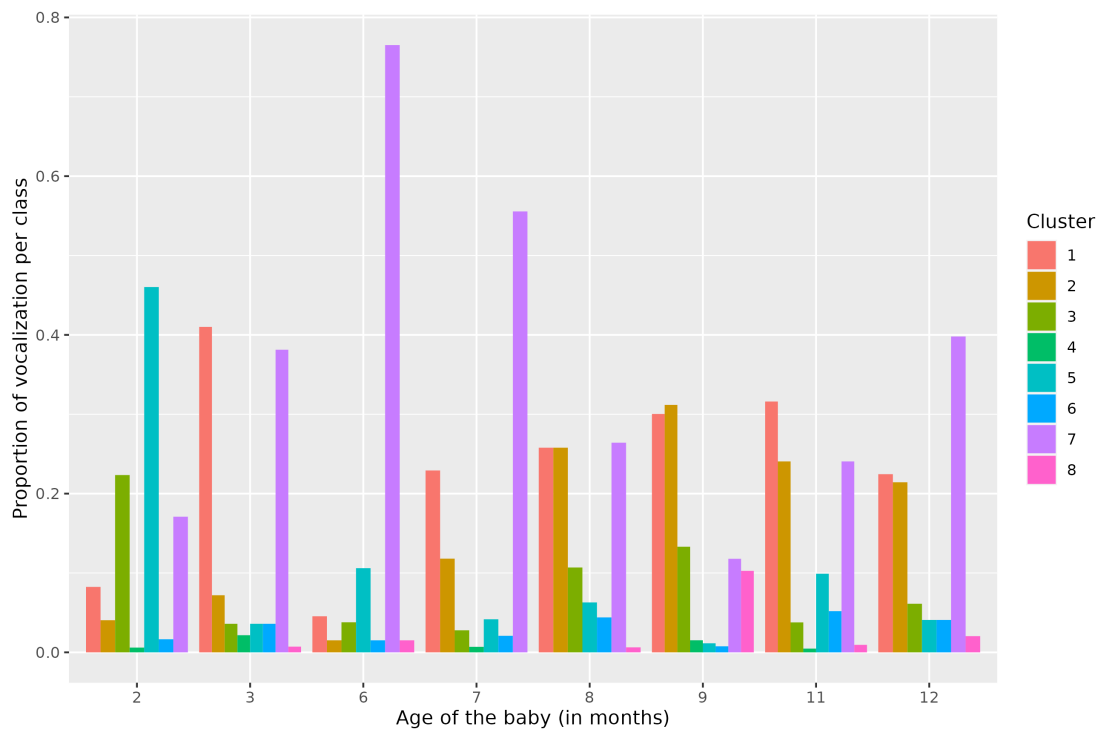


FIGURE 7.3. – Proportion of monthly production of vocalization per cluster. Parents did not record during three months, yet the gap.

We plot in Figure 7.3 the proportion of vocalization of each cluster at each month. We see that the distribution of the production per cluster is not the same along the year. The first month, the vocalizations are mainly vocalizations of cluster 5, followed by cluster 3. The proportion of the cluster 7, the biggest class, is the third. It is the most produced vocalization up to 8 months of age. At this month, cluster 1 and 2 are almost equally produced, and become most produced in the ninth month. At this month, a cluster appears in the distribution of vocalization, the cluster 8.

Whereas in Figure 7.3 we have the proportion of vocalization of each month per cluster (*i.e.*, it sums to one per month), we have the proportion of vocalization of each cluster per month (*i.e.*, it sums to one per cluster) in Table 7.3. We have complementary information about we have just said with Table 7.3. The vast majority of the production of the vocalizations of cluster 3 and 5 is done during the first months. We can hypothesize that these clusters are the first clusters of vocalizations, from which other vocalization classes will emerge. On the contrary, cluster 8 is mainly produced from the ninth month onwards. This suggests that this vocalization class appears later.

7.4.2. Comparison of clusters

We have computed, for each cluster, the deepest vocalization, that we use as the representative of its cluster. For the deepest vocalization of each cluster, we compute the spectrogram, that we plot in Figure 7.4. We have also plotted in Figure 7.1 the

7. *Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.4. Data analysis*

TABLEAU 7.3. – Proportion of production for each cluster during the year.

Cluster	Month							
	2	3	6	7	8	9	11	12
1	15.28	15.83	1.67	9.17	11.39	21.94	18.61	15.83
2	10.76	3.98	0.80	6.77	16.33	32.67	20.32	3.98
3	65.07	2.18	2.18	1.75	7.42	15.28	3.49	2.18
4	30.77	23.08	0	7.69	0	30.77	7.69	23.08
5	82.97	1.35	3.78	1.62	2.70	0.81	5.68	1.35
6	24.44	11.11	4.44	6.67	15.56	4.44	24.44	11.11
7	22.31	10.37	19.77	15.66	8.22	6.07	22.31	10.37
8	0	2.86	5.71	0	2.86	77.14	0	2.86

deepest vocalization of each cluster in three different manners : the waveform, the surface of the spectrogram (for which we have computed the persistence diagram, presented in Figure 7.5) and the Taken's embeddings (for which we have computed the Taken's embeddings, presented in Figure 7.6).

Cluster 1 and cluster 8 are close on the Figure 7.2, but cluster 1 is produced throughout the year, whereas cluster 8 really starts to be produced from the ninth month. In Figure 7.4, cluster 8 has the clearest formants, whereas cluster 1 shows a less clear distribution of energy peaks.

We can see from the Table 7.3 that cluster 2 is a rather late vocalization in the first year. Indeed, a third of cluster 3's production takes place at the ninth month, 20% at the eleventh, which means that more than half of cluster 3's production takes place after the ninth month. Moreover, its production increases, with 16% achieved in the eighth month. On the spectrogram of the deepest vocalizations in Figure 7.4, cluster 3 is also the vocalization with the most formant. The child is thus making increasing use of the resonators in his vocal tract, which are crucial for vowel production.

In contrast to cluster 2, which is a cluster produced mainly in the last 4 months of the first year, and which is produced more and more, cluster 5 is a cluster containing vocalizations produced at over 80% during the first months of life. We thus have a cluster that includes post-natal vocalizations, which are produced less and less as the child learn to produce the other vocalizations.

Moreover, we look at the topological differences of each cluster. We compute the persistent diagram of the surface of the spectrogram for each cluster, and we plot them in Figure 7.5. There are also here differences depending on the cluster. Cluster 8, for example, which groups later vocalizations, has homological features of dimension 1 that clearly deviate from the diagonal, suggesting that they are not topological noise. The surface of the spectrogram of the deepest vocalization of this class is thus characterized by different holes, corresponding to the frequency peaks that stand

7. Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.4. Data analysis

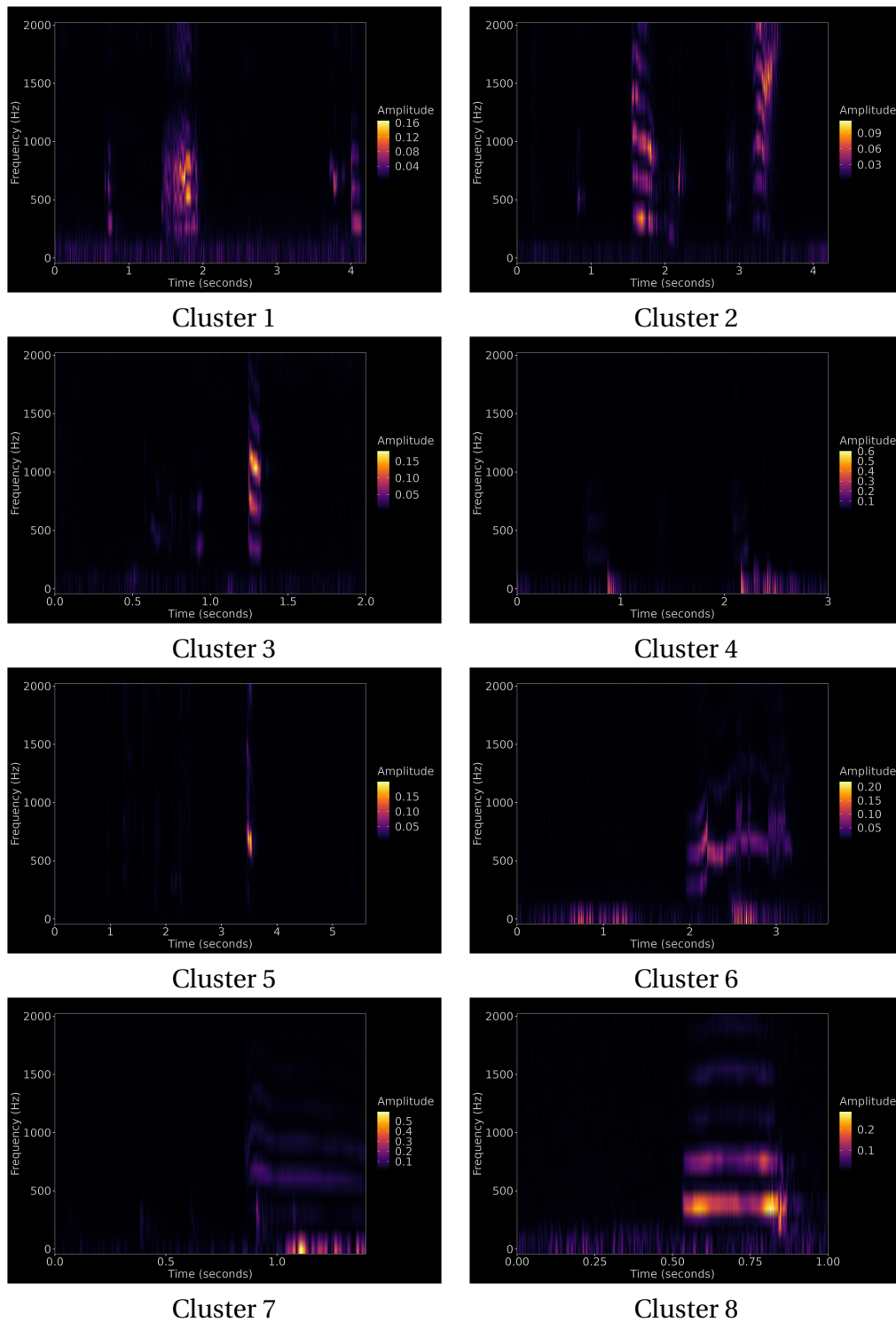


FIGURE 7.4. – Spectrograms of the deepest vocalization of each cluster.

out on the spectrogram. Cluster 1, said to be close to cluster 8 on the Figure 7.2 manifold, also exhibits topological features of dimension 1 away from the diagonal,

7. Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.4. Data analysis

but topological noise is more present, reflecting a less "clear" spectrogram energy distribution for the deepest vocalization in this cluster.

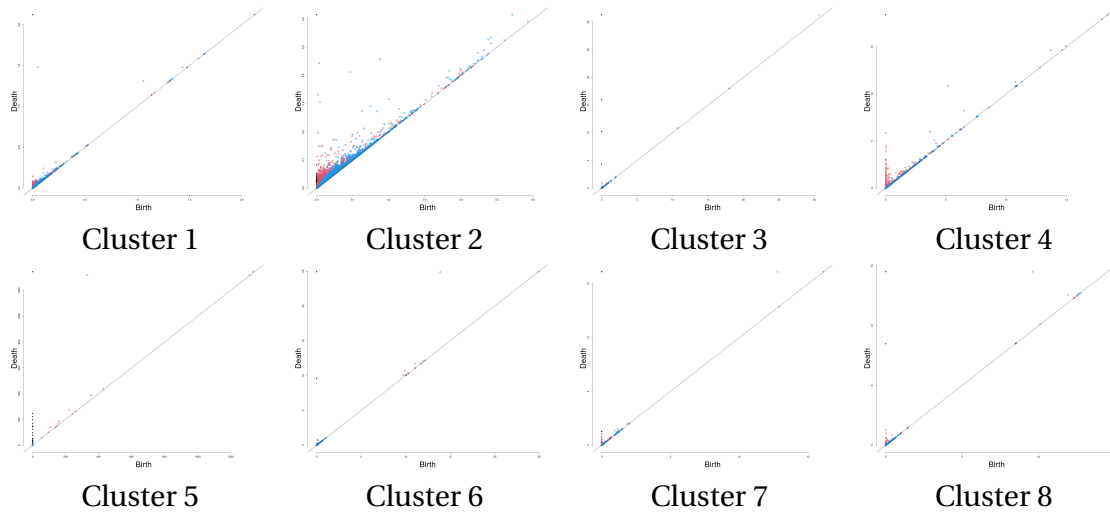


FIGURE 7.6. – Persistence diagrams of the Taken's embeddings of the deepest vocalization of each cluster. The black, red and blue points correspond respectively to homological features of dimension 0, 1 and 2.

Similarly, we compute the persistence diagram of the Taken's embeddings of the deepest vocalization of each cluster, and we plot them in Figure 7.6. Except for cluster 2, the persistence diagrams of the Taken's embeddings are less noisy than those of

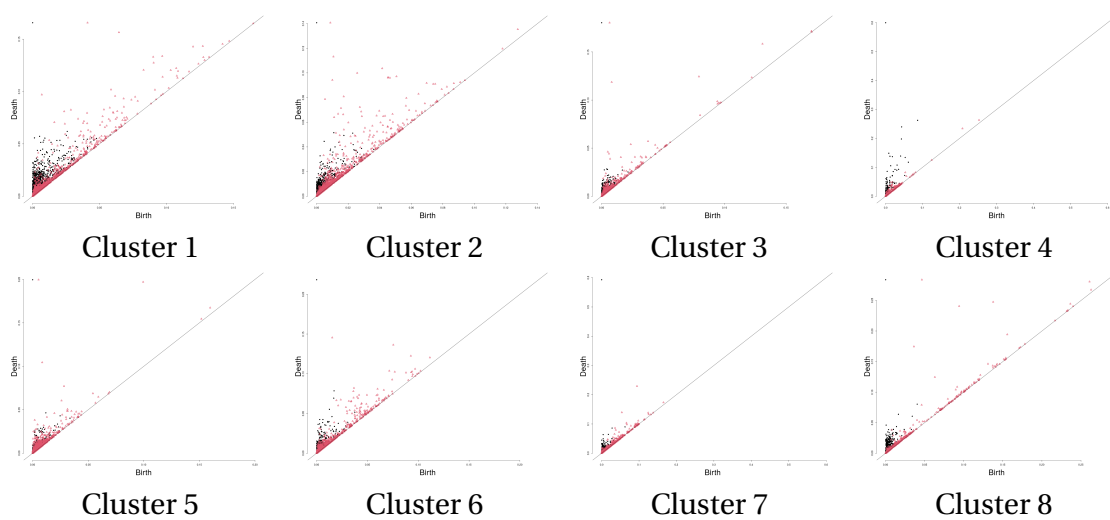


FIGURE 7.5. – Persistence diagrams of the surface of the spectrogram of the deepest vocalization of each cluster. The black, red and blue points correspond respectively to homological features of dimension 0, 1 and 2.

the surface of the spectrogram. There are also differences between clusters, with the Taken's embedding in cluster 4, for example, having a homology of dimension 2 quite far from the diagonal, which none of the others seem to have. Interpretation is more complicated for this signal representation, as the link with the physics of the child's vocal apparatus is not made as it is for the spectrogram.

7.5. Conclusion and discussion

In conclusion, we detected 8 different types of vocalization out of 1851 vocalizations produced by a child over one year. We adapted an unsupervised strategy and worked on a topologically augmented representation of the signal. A signal representation that takes account of the signal's topology provides welcome additional information that may be useful for interpretation.

We note that certain vocalizations only appear from a certain period onward, while being fairly close to each other. Cluster 1 and 8 for instance are close on the vocalizations manifold we compute, but cluster 1 is produced throughout the year and cluster 8 mainly from the ninth month. Moreover, we note a difference of distribution of energy of the spectrograms between the two clusters, energy of cluster 1 being more dispersed than that of cluster 8. From the point of view of language development, this suggests, in line with the idea of calibration (TER HAAR, FERNANDEZ, GRATIER et al. 2021), that the child discovers and learns to master its vocal apparatus during the first few months, to begin producing vocalizations whose sounds are increasingly close to the phonemes of its mother tongue. The literature shows that canonical babbling starts around the sixth month and becomes increasingly complex. It is increasingly produced, accounting for 15% of vocal productions between the ninth and tenth months of age (D. Kimbrough OLLER, BUDER, RAMSDELL et al. 2013; Margaret CYCHOSZ, CRISTIA, Erika BERGELSON et al. 2021). These two close clusters could be two different stages of a class of vocalization.

The present analysis has some limits. First, we take non-informative priors. It could be interesting to incorporate more expert knowledge into the modelling. In particular, the choice of α could be changed. It may have an impact on the final clustering, an increase mechanically leading to the creation of more clusters. We compute it as we expect 5 clusters, following the number of clusters of the public dataset Meg CYCHOSZ, SEIDL, Erika BERGELSON et al. 2019, but we can already distinguish more type of vocalizations during the first year of life of babies in the literature (D. Kimbrough OLLER, BUDER, RAMSDELL et al. 2013; JHANG 2017; BROOKS et KEMPE 2012). We should try to refine the clustering tuning this prior to study the impact on the number of clusters and their characteristics.

The topological representation here used should also be improved. We know from past studies (Chapter 5) that topological information, computed for different representations, is useful and complementary to more conventional descriptors for the classification of vocal signal. To avoid the curse of dimensionality, we build synthetic persistent variables, with which we lose a lot of information (specially for the

7. *Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.6. Supplementary materials*

persistent homology of the spectrogram). We should explore ways of constructing a lower-dimensional signal representation that incorporates topological information, *e.g.*, following MOOR, HORN, RIECK et al. 2020; TROFIMOV, CHERNIAVSKII, TULCHINSKII et al. 2023.

This work needs to be extended, and the analyses made in greater depth. First, dependency should be taken into account. We have dismissed this question here and considered all vocalizations to be exchangeable. Reconsidering the dependency and longitudinality of vocalizations is undoubtedly the first thing to take into account, by defining a non-parametric regression (QUINTANA, Peter MÜLLER, JARA et al. 2022). This would also make it possible to incorporate covariates into the analysis. If, as a second step, we were to add a hierarchy in order to integrate several children, we could compare the differences in vocal productions and their evolution over the first year of life according to the values of the covariates we have integrated into the analysis.

7.6. Supplementary materials

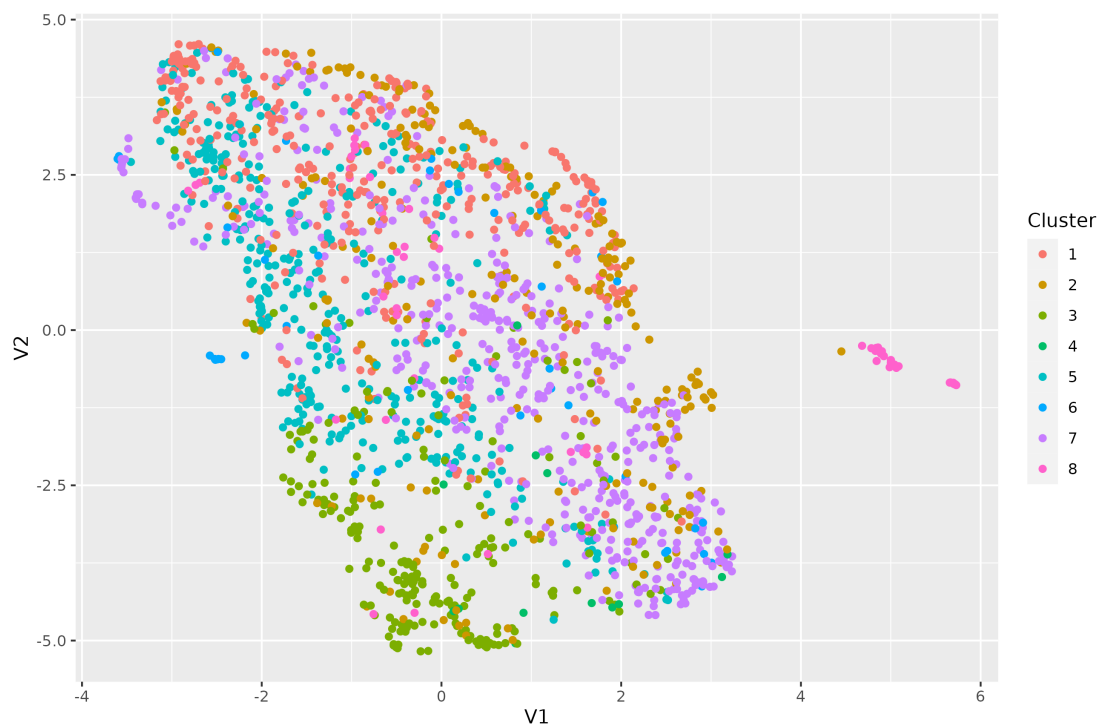


FIGURE S7.1. – Initial manifold of the vocalizations of the baby produced during one year, on which we project the clustering of the Dirichlet process mixtures model, before to remove the "garbage cluster". This class is separated from the other classes, composed by baby vocalizations.

Conclusion

Dans notre première contribution, nous avons proposé une méthodologie *end-to-end* pour détecter les segments de vocalisation de durées variables dans les enregistrements audio de longue durée et les étiqueter en fonction d'un ensemble de données d'entraînement pré-étiquetées. Elle permet le traitement rapide d'enregistrements massifs, réalisés dans des conditions d'enregistrement naturelles et bruyantes, qui peuvent varier d'un enregistrement à l'autre. Elle est basée sur un réseau de neurones profond qui effectue les deux tâches simultanément : la détection et la classification des vocalisations. Nous avons réussi à entraîner le réseau de neurones à partir d'un ensemble de données d'entraînement réduit (≈ 60 minutes). Il peut être utilisé pour traiter, segmenter et classifier automatiquement des signaux audio massifs enregistrés dans différentes conditions (orientation du microphone, position de la source par rapport au microphone) et dans des conditions non contrôlées (bruits de fond) pour un coût de calcul raisonnable. Nous répondons aux attentes fixées (signaux sonores massifs et bruyants enregistrés dans différentes conditions, mais petit ensemble de données d'entraînement), tout en incluant toutes les étapes de pré-traitement dans le réseau, notamment la question de la représentation des données audio. Il est généralisable à différentes espèces, différentes conditions et différents écosystèmes. Notre contribution est importante pour la littérature sur l'apprentissage automatique pour trois raisons. Tout d'abord, elle supprime certaines des limitations qui avaient été soulignées par la communauté bioacoustique (STOWELL 2022). Notre contribution débloque donc ces points et rend possible l'utilisation d'une approche par apprentissage profond pour les problèmes de bioacoustique avec peu de données étiquetées et des enregistrements continus massifs et bruyants. La mise à disposition du code avec un exemple, ainsi qu'une nouvelle base de données, devrait faciliter et encourager son utilisation. Deuxièmement, notre méthode propose d'entraîner un réseau de neurones à un coût de calcul relativement maîtrisé. À l'heure où la question de l'impact écologique des méthodes d'apprentissage se pose, nous contribuons à enrichir la littérature d'exemples de méthodes d'apprentissage conscientes de cette problématique et cherchant à l'intégrer. Troisièmement, notre contribution permet la création de nouvelles bases de données étiquetées qui peuvent être utilisées comme de nouvelles bases d'entraînement plus massives pour l'apprentissage de nouveaux modèles. La base de données créée dans le cadre de notre contribution est librement accessible. Nous avons testé notre *pipeline* sur deux problèmes différents, afin de démontrer sa capacité à s'adapter à différents contextes. Dans les deux cas, les données ne sont pas synthétiques, mais naturelles. Dans le premier cas, nous traitons un problème de bioacoustique : un groupe de babouins enregistré en continu pendant un mois dans son habitat, un centre de primatologie. Deux microphones ont été placés

7. *Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.6. Supplementary materials*

à proximité de l'enclos et ont enregistré le groupe en continu. Les singes sont en semi-liberté : ils peuvent se déplacer librement dans leur enclos. Par conséquent, les sources (les singes) sont en mouvement constant par rapport aux microphones. En outre, nous enregistrons tous les sons émis à l'extérieur. La deuxième étude concerne le développement du langage, l'objet de cette thèse : les enfants ont été enregistrés à intervalles réguliers, trois jours par mois, pendant un an. Les parents ont placé un microphone à proximité des enfants pour enregistrer leurs vocalisations. À chaque nouvel enregistrement, les microphones changeaient potentiellement d'orientation et de positionnement par rapport à la source. Là encore, les enregistrements sont massifs, avec beaucoup de sons parasites et le signal d'intérêt en faible quantité par rapport à la durée des enregistrements continus. Dans les deux cas, nous obtenons d'excellents résultats. À partir d'une base d'entraînement de 72 et 77 minutes respectivement, nous apprenons un réseau de neurones pour chaque étude, avec une précision de 94,58% et 99,76% respectivement. Chaque enregistrement audio long, d'une durée de 443 et 174 heures respectivement, est traité en moins de 10 heures sur un ordinateur portable équipé d'un seul GPU. Nous produisons une nouvelle base de données de productions vocales pour chaque espèce, d'une durée de 38,8 et 35,2 heures respectivement. La nouvelle base de données de vocalisations de singes étiquetés est librement accessible sur zenodo (<https://zenodo.org/record/8239697>). Pour des raisons juridiques, il n'est pas possible de rendre la base de données des vocalisations des enfants librement accessible. Le code est également accessible sur gitlab (<https://gitlab.com/papers4375727/detection-and-classification-of-vocal-productions>).

Dans notre deuxième contribution, nous apportons une preuve empirique de l'utilité de la TDA pour classifier des signaux de vocalisation humaine. On trouve dans la littérature récente des preuves de son intérêt pour traiter différents types de données, mais rien jusqu'à présent n'a été fait concernant le signal vocal humain. Nous montrons grâce à ce travail sa complémentarité avec des descripteurs du signal plus classique. De plus, nous traitons une question jusqu'à présent totalement nouvelle pour la littérature, à savoir l'impact de la représentation du signal. Nous nous sommes intéressés et avons comparé l'homologie persistante de différents objets représentant un même signal, et en quoi l'information homologique que l'on extrayait de ces objets était différente ou non, complémentaire ou non, et si un de ces objets était supérieur ou non d'un point de vu informatif. On a introduit pour cela des représentations que l'on retrouve peu dans la littérature du traitement topologique du signal, notamment les zéros du spectrogramme (FLANDRIN 2015), qui permet par ailleurs une plus grande amélioration des résultats que des représentations que l'on retrouve plus classiquement comme les plongements de Taken. Pour étudier cela, nous avons récolté une nouvelle base de données, librement accessible sur zenodo <https://zenodo.org/record/7961904>, de huit voyelles prononcées par 20 francophones, sous 7 conditions, 10 fois chacune, soit 11 200 enregistrements. Les conditions d'enregistrement sont contrôlées, contrairement aux enregistrements de bébé que nous avons extrait avec la première contribution, et on connaît plusieurs des étiquettes de

chacune de ces productions vocales : la voyelle prononcée, le sexe de l'émetteur et son identité. Cette information nous permet d'estimer clairement l'intérêt de l'approche topologique pour classifier des données vocales humaines, tout en faisant, encore une fois, un travail sur des données réelles et pas sur des jeux de données synthétiques. Pour les trois problèmes de classification que nous avons, l'ajout d'une information topologique permet d'améliorer les résultats pour deux d'entre eux : on passe d'une erreur (OOB) de 8.71% à 7.98% pour la prédiction des voyelles et de 11.54% à 9.24% pour la prédiction de l'individu. Les meilleurs résultats sont obtenus à chaque fois avec une approche topologiquement augmentée, *i.e.*, en construisant un vecteur de représentation du signal qui incorpore des variables topologiques ainsi que des descripteurs plus classiques, les MFCC. Nous avons mis en place une procédure de sélection des variables *step-wise* afin de voir quelles étaient les variables qui restaient à chaque fois dans le meilleur modèle, en supprimant itérativement la variable la moins utile. Chacun des meilleurs modèles prenait en compte au moins une variable topologique, confirmant que ce n'est pas une inflation des variables qui a permis d'améliorer les résultats, et que l'information portée par les variables topologiques peut se substituer à une partie de l'information fréquentielle portée par les MFCC. Cette comparaison illustre également que, pour classifier un signal vocal, il n'y a pas de signature topologique particulière qui se dégage comme meilleure que les autres. La meilleure stratégie est de prendre un ensemble de variables topologiques calculé sur le diagramme de persistance. Il n'y a pas non plus de représentation du signal qui serait supérieur aux autres. Si la surface et surtout les zéros du spectrogramme améliorent davantage les résultats que les plongements de Taken, il semble y avoir une complémentarité des représentations.

Dans notre troisième contribution, nous analysons une nouvelle base de données, un sous-ensemble des enregistrements que nous avons récolté, constitué de vocalisations d'un enfant. Ces vocalisations ont été produites alors que l'enfant était chez lui, tout au long de sa première année de vie, à différents moments de la journée, dans différentes conditions, quand il était seul ou entouré. On a ainsi une diversité importante et une richesse assez inédite. Nous avons choisi, pour représenter ces vocalisations, de passer par une méthode topologiquement augmentée. Ainsi, en plus des MFCC, nous avons calculé les filtrations pour deux représentations du signal, la surface du spectrogramme et les plongements de Taken. Nous avons extrait des diagrammes de persistance qui en résultaient un ensemble de variables topologiques, suivant les méthodes présentées et comparées dans le Chapitre 5. Afin d'éviter un problème de dimensionnalité lors de la modélisation, nous avons construit, pour chaque diagramme de persistance, une variable persistante synthétique par combinaison linéaire de l'ensemble des variables calculées. Afin de proposer une classification possiblement plus fine des vocalisations de l'enfant durant sa première année de vie, nous avons adopté une modélisation permettant de faire une classification non-supervisée. Nous avons utilisé un modèle de mélange avec processus de Dirichlet, celui-ci permettant d'estimer le nombre K de composants de mélange nécessaire pour l'échantillon. Nous avons détecté pour cet enfant huit catégories de vocalisations, avec

7. *Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.6. Supplementary materials*

des caractéristiques fréquentielles et temporelles différentes. Leur production est aussi marquée dans le décours développemental : certaines de ces classes sont produites très majoritairement à certains mois de la première année de l'enfant. Ainsi, une des classes, la classe 5, est produite à 82,97% pendant le deuxième mois, tout comme la classe 3, produite à 65,07% lors de ce deuxième mois. Nous avons ainsi une diversité relativement importante dès le deuxième mois (JHANG 2017). La classe 8 est produite à 77.14% au neuvième mois, moment où les babillages gagnent en complexité et sont de plus en plus produits (Margaret CYCHOSZ, CRISTIA, Erika BERGELSON et al. 2021 ; LANG, BARTL-POKORNY, POKORNY et al. 2019 ; D. Kimbrough OLLER 2000). Des extensions et une analyse approfondie serait nécessaire afin de tirer plus d'information de ces résultats.

Les résultats présentés ouvrent plusieurs perspectives. Tout d'abord, relativement à la représentation du signal et à l'intégration d'une information topologique. Nous avons exploré ici une piste se basant sur les diagrammes de persistance. Nous avons calculé un objet permettant de représenter une vocalisation, un spectrogramme, un plongement de Taken, pour lequel nous avons regardé la topologie. Nous nous sommes intéressés à l'homologie persistante de chaque objet en calculant son diagramme de persistance. L'espace des diagrammes de persistance n'étant pas un espace vectoriel, nous avons mis en place des stratégies pour vectoriser les diagrammes de persistance à partir de ce qu'il y avait dans la littérature. Néanmoins, d'autres stratégies, mentionnées au cours de ce travail, mais non explorées, pourraient se révéler intéressantes et fructueuses. Dans la suite de TROFIMOV, CHERNIAVSKII, TULCHINSKII et al. 2023 ou MOOR, HORN, RIECK et al. 2020, un autoencodeur avec une perte topologique (BARANNIKOV, TROFIMOV, BALABIN et al. 2022) permettrait de construire une représentation du signal de dimension réduite, tout en gardant une information sur la topologie de celui-ci. De plus, nous avons analysé la topologie des objets représentant le signal, mais il pourrait être instructif, dans la veine de ce qu'a fait CARLSSON, ISHKHANOV, Vin DE SILVA et al. 2008 pour les images, de s'intéresser à la topologie de l'espace des vocalisations.

La question de la modélisation du Chapitre 7 ouvre, elle aussi, de nombreuses perspectives. Tout d'abord, si le modèle a lui-même groupé des vocalisations produites à partir d'une certaine période dans une même classe, ce qui fait sens étant donné les caractéristiques des productions vocales de l'enfant, nous n'avons pas pris en compte la temporalité et la dépendance des productions vocales au cours de la première année. C'est une limite de la modélisation que l'on pourrait intégrer avec les processus de Dirichlet dépendants (CAMPBELL, M. LIU, KULIS et al. 2013). Ensuite, nous avons modélisé les vocalisations produites par un enfant unique. Or, nous avons enregistré plusieurs enfants. Afin de considérer les multiples "producteurs de son", on pourrait ajouter une couche de hiérarchie au modèle (TEH, Michael I JORDAN, BEAL et al. 2006 ; TEH et Michael I. JORDAN 2010). On aurait quelque chose de plus robuste, le fait de ne pas avoir échantillonné certains moments pour un enfant ne serait pas synonyme d'absence complet d'exemple pour le mois en question, et on garderait en même temps la possibilité d'avoir des propriétés particulières selon les enfants.

7. *Dirichlet process mixture model based on topologically augmented signal representation for vocalization clustering and language development – 7.6. Supplementary materials*

Avec l'introduction d'une hiérarchie, la taille de l'ensemble de données augmenterait considérablement, ce qui reposerait la question de l'algorithme d'estimation du modèle. Enfin, il serait souhaitable d'intégrer des covariables à notre modélisation (QUINTANA, W. O. JOHNSON, WAETJEN et al. 2016; QUINTANA, Peter MÜLLER, JARA et al. 2022). Cela nous permettrait de voir l'évolution des productions vocales au cours du temps, pour différents enfants, les liens entre les productions antérieures et celles qui suivent, ainsi que l'impact des variables environnementales pour lesquelles nous avons de l'information. On sait déjà qu'il y a par exemple une différence concernant les productions vocales entre les enfants nés à terme et les prématurés (SHINYA, KAWAI, NIWA et al. 2017; CABON, MET-MONTOT, POREE et al. 2021; SHAHRAMNIA, AHMADI, SAFFARIYAN et al. 2023). Étant donné la richesse de nos enregistrements, on pourrait étudier plus en profondeur le lien entre moment de naissance et évolution des vocalisations. On voudrait vérifier si d'autres variables ont aussi des conséquences sur les productions vocales, comme les conditions d'accouchement, ou quantifier l'impact de l'allaitement ou de la tétine, pour lesquelles nous avons l'information. On sait aussi, depuis HART et RISLEY 2003, l'impact de la classe sociale sur le vocabulaire des enfants. Les chercheurs de l'époque se sont rendu compte qu'une intervention à l'entrée de l'école était bien souvent déjà trop tardive pour réduire les différences de vocabulaire et leur courbe d'apprentissage entre des enfants de différents milieux. Si le fossé de 30 millions de mots a été sur-évalué parce que provenant d'extrapolations des données récoltées, les répliques suivantes ont remontré la différence qui existe selon l'origine sociale (GILKERSON, RICHARDS, WARREN et al. 2017; ROMEO, LEONARD, ROBINSON et al. 2018). Différents facteurs viennent contribuer à cette différence, notamment les interactions de l'enfant (LOGAN, JUSTICE, YUMUŞ et al. 2019), mais leur impact n'est pas tout à fait clair et mérite encore une analyse approfondie (CASILLAS, P. BROWN et LEVINSON 2020). Il serait ainsi intéressant d'analyser si les différences se jouent déjà dans les productions pré-langagières, si oui dans quelles proportions et avec quelle variance selon les individus ou les groupes. Si une information plus tardive du niveau langagier des enfants est disponible, une modélisation longitudinale hiérarchique prenant en compte les variables socio-économiques permettrait d'étudier quantitativement cette question. On peut faire un lien avec ce qui se fait dans la médecine personnalisée (PEDONE 2022; PEDONE, ARGIENTO et STINGO 2023). Grâce à cette modélisation, on pourra estimer, pour chaque enfant, étant donné les valeurs de différentes covariables, quels sont les risques qu'il y ait un retard développemental dès la première année de vie, et donc possiblement intervenir le plus tôt possible.

Bibliographie

- [Abd+14] Ossama ABDEL-HAMID, Abdel-rahman MOHAMED, Hui JIANG et al. « Convolutional Neural Networks for Speech Recognition ». In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.10 (oct. 2014), p. 1533-1545. ISSN : 2329-9290, 2329-9304. DOI : [10.1109/TASLP.2014.2339736](https://doi.org/10.1109/TASLP.2014.2339736). (Visité le 01/10/2020) (cf. p. 24).
- [ACL19] Sajjad ABDOLI, Patrick CARDINAL et Alessandro LAMEIRAS KOERICH. « End-to-End Environmental Sound Classification Using a 1D Convolutional Neural Network ». In : *Expert Systems with Applications* 136 (déc. 2019), p. 252-263. ISSN : 09574174. DOI : [10.1016/j.eswa.2019.06.040](https://doi.org/10.1016/j.eswa.2019.06.040). (Visité le 21/06/2022) (cf. p. 46).
- [Ada+17] Henry ADAMS, Tegan EMERSON, Michael KIRBY et al. « Persistence Images : A Stable Vector Representation of Persistent Homology ». In : *The Journal of Machine Learning Research* 18.1 (jan. 2017), p. 8830. ISSN : 1532-4435 (cf. p. 101, 110, 115, 125).
- [AJA16] Elaine ANGELINO, Matthew James JOHNSON et Ryan P. ADAMS. « Patterns of Scalable Bayesian Inference ». In : *Foundations and Trends® in Machine Learning* 9.2-3 (nov. 2016), p. 119-247. ISSN : 1935-8237, 1935-8245. DOI : [10.1561/22000000052](https://doi.org/10.1561/22000000052). (Visité le 18/09/2023) (cf. p. 35).
- [Ant74] Charles E. ANTONIAK. « Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems ». In : *The Annals of Statistics* 2.6 (1974), p. 1152-1174. ISSN : 00905364. (Visité le 27/09/2023) (cf. p. 153).
- [ADP19] Julyan ARBEL, Pierpaolo DE BLASI et Igor PRÜNSTER. « Stochastic Approximations to the Pitman–Yor Process ». In : *Bayesian Analysis* 14.4 (déc. 2019). ISSN : 1936-0975. DOI : [10.1214/18-BA1127](https://doi.org/10.1214/18-BA1127). (Visité le 11/09/2023) (cf. p. 34, 139).
- [Are+21a] Sonia ARENILLAS-ALCÓN, Jordi COSTA-FAIDELLA, Teresa RIBAS-PRATS et al. « Neural Encoding of Voice Pitch and Formant Structure at Birth as Revealed by Frequency-Following Responses ». In : *Scientific Reports* 11.1 (mars 2021), p. 6660. ISSN : 2045-2322. DOI : [10.1038/s41598-021-85799-x](https://doi.org/10.1038/s41598-021-85799-x). (Visité le 22/01/2023) (cf. p. 18).
- [Are+21b] Sonia ARENILLAS-ALCÓN, Teresa RIBAS-PRATS, Carles ESCERA et al. « Encoding of Fundamental Frequency and Fine Structure of Speech Sounds : A Comparison between Adults and Newborns ». In : *The Journal of the Acoustical Society of America* 150.4 (oct. 2021), A63-A63. ISSN : 0001-4966. DOI : [10.1121/10.0007632](https://doi.org/10.1121/10.0007632). (Visité le 22/01/2023) (cf. p. 18).

- [Arn+23] Vincent ARNAUD, François PELLEGRINO, Sumir KEENAN et al. « Improving the Workflow to Crack Small, Unbalanced, Noisy, but Genuine (SUNG) Datasets in Bioacoustics : The Case of Bonobo Calls ». In : *PLoS computational biology* 19.4 (avr. 2023), e1010325. ISSN : 1553-7358. DOI : [10.1371/journal.pcbi.1010325](https://doi.org/10.1371/journal.pcbi.1010325) (cf. p. 26).
- [ACB19] Devansh ARPIT, Victor CAMPOS et Yoshua BENGIO. *How to Initialize Your Network? Robust Initialization for WeightNorm & ResNets*. Oct. 2019. DOI : [10.48550/arXiv.1906.02341](https://doi.org/10.48550/arXiv.1906.02341). arXiv : 1906.02341 [cs, stat]. (Visité le 02/09/2023) (cf. p. 42).
- [ASN98] Richard N. ASLIN, Jenny R. SAFFRAN et Elissa L. NEWPORT. « Computation of Conditional Probability Statistics by 8-Month-Old Infants ». In : *Psychological Science* 9.4 (juill. 1998), p. 321-324. ISSN : 0956-7976, 1467-9280. DOI : [10.1111/1467-9280.00063](https://doi.org/10.1111/1467-9280.00063). (Visité le 04/06/2019) (cf. p. 18).
- [AGR19] Nieves ATIENZA, Rocio GONZALEZ-DIAZ et Matteo RUCCO. « Persistent Entropy for Separating Topological Features from Noise in Vietoris-Rips Complexes ». In : *Journal of Intelligent Information Systems* 52.3 (juin 2019), p. 637-655. ISSN : 0925-9902, 1573-7675. DOI : [10.1007/s10844-017-0473-4](https://doi.org/10.1007/s10844-017-0473-4). (Visité le 03/11/2022) (cf. p. 101, 110, 113, 152).
- [AGS20] Nieves ATIENZA, Rocio GONZALEZ-DÍAZ et Manuel SORIANO-TRIGUEROS. « On the Stability of Persistent Entropy and New Summary Functions for Topological Data Analysis ». In : *Pattern Recognition* 107 (nov. 2020), p. 107509. ISSN : 0031-3203. DOI : [10.1016/j.patcog.2020.107509](https://doi.org/10.1016/j.patcog.2020.107509). (Visité le 02/12/2022) (cf. p. 101, 110).
- [AVT16] Yusuf AYTAR, Carl VONDRICK et Antonio TORRALBA. « SoundNet : Learning Sound Representations from Unlabeled Video ». In : *arXiv :1610.09001 [cs]* (oct. 2016). arXiv : 1610.09001 [cs]. (Visité le 09/03/2021) (cf. p. 65).
- [Bar94] Serguei BARANNIKOV. « The Framed Morse Complex and Its Invariants ». In : *Advances in Soviet Mathematics. Singularities and Bifurcations* 21 (avr. 1994), p. 93-116. DOI : [10.1090/advsov/021/03](https://doi.org/10.1090/advsov/021/03). (Visité le 15/08/2023) (cf. p. 94).
- [Bar+22a] Serguei BARANNIKOV, Ilya TROFIMOV, Nikita BALABIN et al. « Representation Topology Divergence : A Method for Comparing Neural Network Representations. » In : *Proceedings of the 39th International Conference on Machine Learning*. PMLR, juin 2022, p. 1607-1626. (Visité le 15/08/2023) (cf. p. 30, 101, 165).
- [BS20a] Sergio BARBAROSSA et Stefania SARDELLITTI. « Topological Signal Processing Over Simplicial Complexes ». In : *IEEE Transactions on Signal Processing* 68 (2020), p. 2992-3007. ISSN : 1941-0476. DOI : [10.1109/TSP.2020.2981920](https://doi.org/10.1109/TSP.2020.2981920) (cf. p. 30, 104, 105).

- [BS20b] Sergio BARBAROSSA et Stefania SARDELLITTI. « Topological Signal Processing : Making Sense of Data Building on Multiway Relations ». In : *IEEE Signal Processing Magazine* 37.6 (nov. 2020), p. 174-183. ISSN : 1053-5888, 1558-0792. DOI : [10.1109/MSP.2020.3014067](https://doi.org/10.1109/MSP.2020.3014067). (Visité le 15/03/2023) (cf. p. [30](#), [104](#), [105](#)).
- [Bar+15] Guillaume BARBIER, Louis-Jean BOË, Guillaume CAPTIER et al. « Human Vocal Tract Growth : A Longitudinal Study of the Development of Various Anatomical Structures ». In : *Interspeech 2015 - 16th Annual Conference of the International Speech Communication Association*. Sept. 2015. (Visité le 15/05/2023) (cf. p. [21](#), [146](#)).
- [BPP21] Danielle BARNES, Luis POLANCO et Jose A. PEREA. « A Comparative Study of Machine Learning Methods for Persistence Diagrams ». In : *Frontiers in Artificial Intelligence* 4 (juill. 2021), p. 681174. ISSN : 2624-8212. DOI : [10.3389/frai.2021.681174](https://doi.org/10.3389/frai.2021.681174). (Visité le 02/11/2022) (cf. p. [101](#), [110](#), [124](#), [147](#)).
- [Bar+22b] Katrin D. BARTL-POKORNY, Florian B. POKORNY, Dunia GARRIDO et al. « Vocalisation Repertoire at the End of the First Year of Life : An Exploratory Comparison of Rett Syndrome and Typical Development ». In : *Journal of Developmental and Physical Disabilities* (mars 2022). ISSN : 1056-263X, 1573-3580. DOI : [10.1007/s10882-022-09837-w](https://doi.org/10.1007/s10882-022-09837-w). (Visité le 08/09/2022) (cf. p. [20](#), [21](#), [146](#)).
- [Bay+17] Atılım Günes BAYDIN, Barak A. PEARLMUTTER, Alexey Andreyevich RADUL et al. « Automatic Differentiation in Machine Learning : A Survey ». In : *The Journal of Machine Learning Research* 18.1 (jan. 2017), p. 5595-5637. ISSN : 1532-4435 (cf. p. [54](#)).
- [BCV13] Yoshua BENGIO, Aaron COURVILLE et Pascal VINCENT. « Representation Learning : A Review and New Perspectives ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (août 2013), p. 1798-1828. ISSN : 1939-3539. DOI : [10.1109/TPAMI.2013.50](https://doi.org/10.1109/TPAMI.2013.50) (cf. p. [24](#), [29](#), [41](#), [45](#)).
- [Ber22] Clément BERENFELD. « Statistical Inference on Unknown Manifolds ». These de Doctorat. Université Paris sciences et lettres, sept. 2022. (Visité le 04/07/2023) (cf. p. [28](#), [151](#)).
- [BS12] E. BERGELSON et D. SWINGLEY. « At 6-9 Months, Human Infants Know the Meanings of Many Common Nouns ». In : *Proceedings of the National Academy of Sciences* 109.9 (fév. 2012), p. 3253-3258. ISSN : 0027-8424, 1091-6490. DOI : [10.1073/pnas.1113380109](https://doi.org/10.1073/pnas.1113380109). (Visité le 20/06/2019) (cf. p. [19](#)).
- [Ber+19] Christian BERGLER, Hendrik SCHRÖTER, Rachael Xi CHENG et al. « ORCA-SPOT : An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning ». In : *Scientific Reports* 9.1 (déc. 2019), p. 10997. ISSN : 2045-

2322. DOI : [10.1038/s41598-019-47335-w](https://doi.org/10.1038/s41598-019-47335-w). (Visité le 23/09/2019) (cf. p. 24, 65).
- [BB20] Mattia G. BERGOMI et Adriano BARATÈ. « Homological Persistence in Time Series : An Application to Music Classification ». In : *Journal of Mathematics and Music* 14.2 (mai 2020), p. 204-221. ISSN : 1745-9737, 1745-9745. DOI : [10.1080/17459737.2020.1786745](https://doi.org/10.1080/17459737.2020.1786745). (Visité le 09/03/2022) (cf. p. 31, 101, 105).
- [BJ04] David M. BLEI et Michael I. JORDAN. « Variational Methods for the Dirichlet Process ». In : *Twenty-First International Conference on Machine Learning - ICML '04*. Banff, Alberta, Canada : ACM Press, 2004, p. 12. DOI : [10.1145/1015330.1015439](https://doi.org/10.1145/1015330.1015439). (Visité le 30/05/2023) (cf. p. 35, 139).
- [BJ06] David M. BLEI et Michael I. JORDAN. « Variational Inference for Dirichlet Process Mixtures ». In : *Bayesian Analysis* 1.1 (mars 2006), p. 121-143. ISSN : 1936-0975, 1931-6690. DOI : [10.1214/06-BA104](https://doi.org/10.1214/06-BA104). (Visité le 05/04/2023) (cf. p. 35, 139).
- [BKM17] David M. BLEI, Alp KUCUKELBIR et Jon D. MCAULIFFE. « Variational Inference : A Review for Statisticians ». In : *Journal of the American Statistical Association* 112.518 (avr. 2017), p. 859-877. ISSN : 0162-1459, 1537-274X. DOI : [10.1080/01621459.2017.1285773](https://doi.org/10.1080/01621459.2017.1285773). arXiv : [1601.00670](https://arxiv.org/abs/1601.00670). (Visité le 08/04/2019) (cf. p. 35).
- [BNJ03] David M. BLEI, Andrew Y. NG et Michael I. JORDAN. « Latent Dirichlet Allocation ». In : *Journal of Machine Learning Research* 3. Jan (2003), p. 993-1022. ISSN : ISSN 1533-7928. (Visité le 17/06/2023) (cf. p. 35).
- [Boë+13] Louis-Jean BOË, Pierre BADIN, Lucie MÉNARD et al. « Anatomy and Control of the Developing Human Vocal Tract : A Response to Lieberman ». In : *Journal of Phonetics* 41.5 (sept. 2013), p. 379-392. ISSN : 00954470. DOI : [10.1016/j.wocn.2013.04.001](https://doi.org/10.1016/j.wocn.2013.04.001). (Visité le 18/04/2019) (cf. p. 28).
- [Boë+17] Louis-Jean BOË, Frédéric BERTHOMMIER, Thierry LEGOU et al. « Evidence of a Vocalic Proto-System in the Baboon (*Papio Papio*) Suggests Pre-Hominin Speech Precursors ». In : *PLOS ONE* 12.1 (jan. 2017). Sous la dir. de David REBY, e0169321. ISSN : 1932-6203. DOI : [10.1371/journal.pone.0169321](https://doi.org/10.1371/journal.pone.0169321). (Visité le 17/04/2019) (cf. p. 27, 73).
- [BCM22] Jean-Daniel BOISSONNAT, Frédéric CHAZAL et Bertrand MICHEL. « Topological Data Analysis ». In : *Novel Mathematics Inspired by Industrial Challenges*. Sous la dir. de Michael GÜNTHER et Wil SCHILDERS. T. 38. Cham : Springer International Publishing, 2022, p. 247-269. ISBN : 978-3-030-96172-5 978-3-030-96173-2. DOI : [10.1007/978-3-030-96173-2_9](https://doi.org/10.1007/978-3-030-96173-2_9). (Visité le 14/04/2023) (cf. p. 105).
- [Bon+23a] Guillem BONAFOS, Jean-Marc FREYERMUTH, Pierre PUDLO et al. *French vowels*. Mai 2023. DOI : [10.5281/zenodo.7961904](https://doi.org/10.5281/zenodo.7961904). (Visité le 23/05/2023) (cf. p. 111).

- [Bon+23b] Guillem BONAFO, Pierre PUDLO, Jean-Marc FREYERMUTH et al. « Detecting Human and Non-Human Vocal Productions in Large Scale Audio Recordings ». In : (fév. 2023). DOI : [10.48550/arXiv.2302.07640](https://doi.org/10.48550/arXiv.2302.07640). arXiv : [2302.07640](https://arxiv.org/abs/2302.07640) [cs, eess, stat]. (Visité le 27/02/2023) (cf. p. 49, 147).
- [BGS07] C. BOUVEYRON, S. GIRARD et C. SCHMID. « High-Dimensional Data Clustering ». In : *Computational Statistics & Data Analysis* 52.1 (sept. 2007), p. 502-519. ISSN : 0167-9473. DOI : [10.1016/j.csda.2007.02.009](https://doi.org/10.1016/j.csda.2007.02.009). (Visité le 04/08/2023) (cf. p. 32).
- [BSD84] Bénédicte De BOYSSON-BARDIES, Laurent SAGART et Catherine DURAND. « Discernible Differences in the Babbling of Infants According to Target Language ». In : *Journal of Child Language* 11.1 (fév. 1984), p. 1-15. ISSN : 1469-7602, 0305-0009. DOI : [10.1017/S0305000900005559](https://doi.org/10.1017/S0305000900005559). (Visité le 30/09/2023) (cf. p. 20, 146).
- [BCd10] Eric BROCHU, Vlad M. CORA et Nando DE FREITAS. « A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning ». In : *arXiv :1012.2599 [cs]* (déc. 2010). arXiv : [1012.2599](https://arxiv.org/abs/1012.2599) [cs]. (Visité le 03/02/2021) (cf. p. 60, 70).
- [BK12] Patricia BROOKS et Vera KEMPE. *Language Development*. BPS Textbooks in Psychology. Chichester : BPS Blackwell, 2012. ISBN : 978-1-4443-3146-2 (cf. p. 146, 160).
- [Bro+23] Bradley CA BROWN, Anthony L. CATERINI, Brendan Leigh ROSS et al. « Verifying the Union of Manifolds Hypothesis for Image Data ». In : *The Eleventh International Conference on Learning Representations*. 2023. (Visité le 29/08/2023) (cf. p. 29).
- [Bub15] Peter BUBENIK. « Statistical Topological Data Analysis Using Persistence Landscapes ». In : *The Journal of Machine Learning Research* 16.1 (2015), p. 26 (cf. p. 100, 101, 110, 115).
- [Bub20] Peter BUBENIK. « The Persistence Landscape and Some of Its Properties ». In : *Topological Data Analysis*. Sous la dir. de Nils A. BAAS, Gunnar E. CARLSSON, Gereon QUICK et al. T. 15. Cham : Springer International Publishing, 2020, p. 97-117. ISBN : 978-3-030-43407-6 978-3-030-43408-3. DOI : [10.1007/978-3-030-43408-3_4](https://doi.org/10.1007/978-3-030-43408-3_4). (Visité le 27/09/2022) (cf. p. 100, 110, 115, 125).
- [Cab+21] Sandie CABON, Bertille MET-MONTOT, Fabienne POREE et al. « Automatic Extraction of Spontaneous Cries of Preterm Newborns in Neonatal Intensive Care Units ». In : *2020 28th European Signal Processing Conference (EUSIPCO)*. Amsterdam, Netherlands : IEEE, jan. 2021, p. 1200-1204. ISBN : 978-90-827970-5-3. DOI : [10.23919/Eusipco47968.2020.9287590](https://doi.org/10.23919/Eusipco47968.2020.9287590). (Visité le 19/12/2021) (cf. p. 21, 146, 166).

- [CMA18] Diana CAI, Michael MITZENMACHER et Ryan P ADAMS. « A Bayesian Non-parametric View on Count-Min Sketch ». In : *Advances in Neural Information Processing Systems*. T. 31. Curran Associates, Inc., 2018. (Visité le 18/09/2023) (cf. p. 35).
- [Cam+13] Trevor CAMPBELL, Miao LIU, Brian KULIS et al. *Dynamic Clustering via Asymptotics of the Dependent Dirichlet Process Mixture*. Nov. 2013. DOI : [10.48550/arXiv.1305.6659](https://doi.org/10.48550/arXiv.1305.6659). arXiv : [1305.6659](https://arxiv.org/abs/1305.6659) [cs, stat]. (Visité le 18/05/2023) (cf. p. 35, 165).
- [CCN21a] Antonio CANALE, Riccardo CORRADIN et Bernardo NIPOTI. *Importance Conditional Sampling for Pitman-Yor Mixtures*. Oct. 2021. arXiv : [1906.08147](https://arxiv.org/abs/1906.08147) [stat]. (Visité le 05/06/2023) (cf. p. 35, 139).
- [Cao97] Liangyue CAO. « Practical Method for Determining the Minimum Embedding Dimension of a Scalar Time Series ». In : *Physica D : Nonlinear Phenomena* 110.1 (déc. 1997), p. 43-50. ISSN : 0167-2789. DOI : [10.1016/S0167-2789\(97\)00118-8](https://doi.org/10.1016/S0167-2789(97)00118-8). (Visité le 30/04/2023) (cf. p. 107, 151).
- [Cao+19] Yueqi CAO, Shiqiang ZHANG, Fangjia YAN et al. « Unsupervised Environmental Sound Classification Based On Topological Persistence ». In : *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*. Déc. 2019, p. 1-5. DOI : [10.1109/ICSIDP47821.2019.9173135](https://doi.org/10.1109/ICSIDP47821.2019.9173135) (cf. p. 31).
- [Car09] Gunnar CARLSSON. « Topology and Data ». In : *Bulletin of the American Mathematical Society* 46.2 (2009), p. 255-308. ISSN : 0273-0979, 1088-9485. DOI : [10.1090/S0273-0979-09-01249-X](https://doi.org/10.1090/S0273-0979-09-01249-X). (Visité le 25/11/2022) (cf. p. 29, 83, 90, 101, 104, 124, 151).
- [Car+08] Gunnar CARLSSON, Tigran ISHKHANOV, Vin DE SILVA et al. « On the Local Behavior of Spaces of Natural Images ». In : *International Journal of Computer Vision* 76.1 (jan. 2008), p. 1-12. ISSN : 1573-1405. DOI : [10.1007/s11263-007-0056-x](https://doi.org/10.1007/s11263-007-0056-x). (Visité le 29/08/2023) (cf. p. 29, 165).
- [COO15] Mathieu CARRIÈRE, Steve Y. OUDOT et Maks OVSJANIKOV. « Stable Topological Signatures for Points on 3D Shapes ». In : *Computer Graphics Forum* 34.5 (août 2015), p. 1-12. ISSN : 01677055. DOI : [10.1111/cgf.12692](https://doi.org/10.1111/cgf.12692). (Visité le 04/11/2022) (cf. p. 30, 101, 110, 123, 124).
- [CBL20] Marisa CASILLAS, Penelope BROWN et Stephen C. LEVINSON. « Early Language Experience in a Tzeltal Mayan Village ». In : *Child Development* 91.5 (sept. 2020), p. 1819-1835. ISSN : 0009-3920, 1467-8624. DOI : [10.1111/cdev.13349](https://doi.org/10.1111/cdev.13349). (Visité le 19/12/2021) (cf. p. 166).
- [CK14] Sachin CHACHADA et C.-C. Jay KUO. « Environmental Sound Recognition : A Survey ». In : *APSIPA Transactions on Signal and Information Processing* 3.1 (2014). ISSN : 2048-7703, 2048-7703. DOI : [10.1017/ATSIP.2014.12](https://doi.org/10.1017/ATSIP.2014.12). (Visité le 14/06/2023) (cf. p. 112, 152).

- [Cha+13] Frederic CHAZAL, Vin DE SILVA, Marc GLISSE et al. *The Structure and Stability of Persistence Modules*. Mars 2013. DOI : [10.48550/arXiv.1207.3674](https://doi.org/10.48550/arXiv.1207.3674). arXiv : [1207.3674](https://arxiv.org/abs/1207.3674) [cs, math]. (Visité le 27/08/2023) (cf. p. 30, 98, 147).
- [Cha+15] Frederic CHAZAL, Brittany FASY, Fabrizio LECCI et al. « Subsampling Methods for Persistent Homology ». In : *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, juin 2015, p. 2143-2151. (Visité le 27/08/2023) (cf. p. 100).
- [Cha+09] Frédéric CHAZAL, David COHEN-STEINER, Marc GLISSE et al. « Proximity of Persistence Modules and Their Diagrams ». In : *Proceedings of the Twenty-Fifth Annual Symposium on Computational Geometry*. SCG '09. New York, NY, USA : Association for Computing Machinery, juin 2009, p. 237-246. ISBN : 978-1-60558-501-7. DOI : [10.1145/1542362.1542407](https://doi.org/10.1145/1542362.1542407). (Visité le 27/08/2023) (cf. p. 98).
- [Cha+18] Frédéric CHAZAL, Brittany FASY, Fabrizio LECCI et al. « Robust Topological Inference : Distance To a Measure and Kernel Distance ». In : *Journal of Machine Learning Research* 18.159 (2018), p. 1-40. ISSN : 1533-7928. (Visité le 28/08/2023) (cf. p. 100, 123).
- [Cha+14a] Frédéric CHAZAL, Brittany Terese FASY, Fabrizio LECCI et al. « Stochastic Convergence of Persistence Landscapes and Silhouettes ». In : *Proceedings of the Thirtieth Annual Symposium on Computational Geometry*. Kyoto Japan : ACM, juin 2014, p. 474-483. ISBN : 978-1-4503-2594-3. DOI : [10.1145/2582112.2582128](https://doi.org/10.1145/2582112.2582128). (Visité le 05/11/2021) (cf. p. 100, 101, 105, 110, 115, 125).
- [Cha+14b] Frédéric CHAZAL, Marc GLISSE, Catherine LABRUÈRE et al. « Convergence Rates for Persistence Diagram Estimation in Topological Data Analysis ». In : *Proceedings of the 31st International Conference on Machine Learning*. PMLR, jan. 2014, p. 163-171. (Visité le 27/08/2023) (cf. p. 99).
- [CM21] Frédéric CHAZAL et Bertrand MICHEL. « An Introduction to Topological Data Analysis : Fundamental and Practical Aspects for Data Scientists ». In : *Frontiers in Artificial Intelligence* 4 (2021). ISSN : 2624-8212. (Visité le 07/05/2023) (cf. p. 29, 83, 86, 99, 104, 105, 108, 123, 124, 151).
- [CLC19] Ming-Tso CHEN, Bo-Jun LI et Tai-Shih CHI. « CNN Based Two-stage Multi-resolution End-to-end Model for Singing Melody Extraction ». In : *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom : IEEE, mai 2019, p. 1005-1009. ISBN : 978-1-4799-8131-1. DOI : [10.1109/ICASSP.2019.8683630](https://doi.org/10.1109/ICASSP.2019.8683630). (Visité le 09/03/2021) (cf. p. 24).

- [Che+19] Shichuan CHEN, Shilian ZHENG, Lifeng YANG et al. « Deep Learning for Large-Scale Real-World ACARS and ADS-B Radio Signal Classification ». In : *IEEE Access* 7 (2019), p. 89256-89264. ISSN : 2169-3536. DOI : [10.1109/ACCESS.2019.2925569](https://doi.org/10.1109/ACCESS.2019.2925569) (cf. p. 25).
- [CSR04] Xin CHEN, Tricia STRIANO et Hannes RAKOCZY. « Auditory–Oral Matching Behavior in Newborns ». In : *Developmental Science* 7.1 (2004), p. 42-47. ISSN : 1467-7687. DOI : [10.1111/j.1467-7687.2004.00321.x](https://doi.org/10.1111/j.1467-7687.2004.00321.x). (Visité le 01/10/2023) (cf. p. 21).
- [Che+95] M. CHEOUR-LUHTANEN, K. ALHO, T. KUJALA et al. « Mismatch Negativity Indicates Vowel Discrimination in Newborns ». In : *Hearing Research* 82.1 (jan. 1995), p. 53-58. ISSN : 0378-5955. DOI : [10.1016/0378-5955\(94\)00164-1](https://doi.org/10.1016/0378-5955(94)00164-1) (cf. p. 18).
- [Cho+17a] Keunwoo CHOI, Gyorgy FAZEKAS, Mark SANDLER et al. « Convolutional Recurrent Neural Networks for Music Classification ». In : *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA : IEEE, mars 2017, p. 2392-2396. ISBN : 978-1-5090-4117-6. DOI : [10.1109/ICASSP.2017.7952585](https://doi.org/10.1109/ICASSP.2017.7952585). (Visité le 10/07/2022) (cf. p. 24).
- [Cho+17b] Keunwoo CHOI, György FAZEKAS, M. SANDLER et al. « Transfer Learning for Music Classification and Regression Tasks ». In : *International Society for Music Information Retrieval Conference*. Mars 2017. (Visité le 24/07/2023) (cf. p. 65).
- [CLG23] Idan COHEN, Ofir LINDENBAUM et Sharon GANNOT. *Unsupervised Acoustic Scene Mapping Based on Acoustic Features and Dimensionality Reduction*. Jan. 2023. DOI : [10.48550/arXiv.2301.00448](https://doi.org/10.48550/arXiv.2301.00448). arXiv : [2301.00448](https://arxiv.org/abs/2301.00448) [cs, eess]. (Visité le 29/08/2023) (cf. p. 29).
- [CEH07] David COHEN-STEINER, Herbert EDELSBRUNNER et John HARER. « Stability of Persistence Diagrams ». In : *Discrete & Computational Geometry* 37.1 (jan. 2007), p. 103-120. ISSN : 1432-0444. DOI : [10.1007/s00454-006-1276-5](https://doi.org/10.1007/s00454-006-1276-5). (Visité le 09/03/2023) (cf. p. 30, 98, 105, 147).
- [Coh+10] David COHEN-STEINER, Herbert EDELSBRUNNER, John HARER et Yuriy MILEYKO. « Lipschitz Functions Have L p -Stable Persistence ». In : *Foundations of Computational Mathematics* 10.2 (avr. 2010), p. 127-139. ISSN : 1615-3375, 1615-3383. DOI : [10.1007/s10208-010-9060-6](https://doi.org/10.1007/s10208-010-9060-6). (Visité le 04/11/2022) (cf. p. 101, 110, 114, 152).
- [CCN21b] Riccardo CORRADIN, Antonio CANALE et Bernardo NIPOTI. « **BNPmix** : An R Package for Bayesian Nonparametric Modeling via Pitman-Yor Mixtures ». In : *Journal of Statistical Software* 100.15 (2021). ISSN : 1548-7660. DOI : [10.18637/jss.v100.i15](https://doi.org/10.18637/jss.v100.i15). (Visité le 31/05/2023) (cf. p. 144).

- [Cyc+21] Margaret CYCHOSZ, Alejandrina CRISTIA, Elika BERGELSON et al. « Vocal Development in a Large-scale Crosslinguistic Corpus ». In : *Developmental Science* 24.5 (sept. 2021). ISSN : 1363-755X, 1467-7687. DOI : [10.1111/desc.13090](https://doi.org/10.1111/desc.13090). (Visité le 25/01/2022) (cf. p. 20, 23, 146, 160, 165).
- [Cyc+19] Meg CYCHOSZ, Amanda SEIDL, Elika BERGELSON et al. « BabbleCor : A Crosslinguistic Corpus of Babble Development in Five Languages ». In : (oct. 2019). DOI : [10.17605/OSF.IO/RZ4TX](https://doi.org/10.17605/OSF.IO/RZ4TX). (Visité le 08/06/2022) (cf. p. 23, 73, 76, 153, 160).
- [De +15] Pierpaolo DE BLASI, Stefano FAVARO, Antonio LIJOI et al. « Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process? » In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2 (fév. 2015), p. 212-229. ISSN : 1939-3539. DOI : [10.1109/TPAMI.2013.217](https://doi.org/10.1109/TPAMI.2013.217) (cf. p. 34).
- [de 19] Bart DE BOER. « Evolution of Speech : Anatomy and Control ». In : *Journal of Speech, Language, and Hearing Research* 62.8S (août 2019), p. 2932-2945. ISSN : 1092-4388, 1558-9102. DOI : [10.1044/2019_JSLHR-S-CSMC7-18-0293](https://doi.org/10.1044/2019_JSLHR-S-CSMC7-18-0293). (Visité le 23/03/2022) (cf. p. 22).
- [de +22] Thibault DE SURREL, Felix HENSEL, Mathieu CARRIÈRE et al. *RipsNet : A General Architecture for Fast and Robust Estimation of the Persistent Homology of Point Clouds*. Fév. 2022. arXiv : [2202.01725 \[cs\]](https://arxiv.org/abs/2202.01725). (Visité le 14/11/2022) (cf. p. 101).
- [DD19] Suman DEB et Samarendra DANDAPAT. « Emotion Classification Using Segmentation of Vowel-Like and Non-Vowel-Like Regions ». In : *IEEE Transactions on Affective Computing* 10.3 (juill. 2019), p. 360-373. ISSN : 1949-3045. DOI : [10.1109/TAFFC.2017.2730187](https://doi.org/10.1109/TAFFC.2017.2730187). (Visité le 28/09/2023) (cf. p. 111).
- [Dev22] TensorFlow DEVELOPERS. *TensorFlow*. Zenodo. Mai 2022. DOI : [10.5281/zenodo.6574269](https://doi.org/10.5281/zenodo.6574269). (Visité le 13/01/2023) (cf. p. 74).
- [DW22] Tamal K. DEY et Yusu WANG. *Computational Topology for Data Analysis*. First edition. New York : Cambridge University Press, 2022. ISBN : 978-1-00-909816-8 (cf. p. 83, 89, 108).
- [Die+04] C DIETRICH, G PALM, K RIEDE et al. « Classification of Bioacoustic Time Series Based on the Combination of Global and Local Decisions ». In : *Pattern Recognition* 37.12 (déc. 2004), p. 2293-2305. ISSN : 00313203. DOI : [10.1016/S0031-3203\(04\)00161-X](https://doi.org/10.1016/S0031-3203(04)00161-X). (Visité le 14/05/2020) (cf. p. 24, 64).
- [DL21] V. DIVOL et T. LACOMBE. « Understanding the topology and the geometry of the space of persistence diagrams via optimal partial transport ». In : *Journal of Applied and Computational Topology* 25 (2021), p. 1-53. DOI : <https://doi.org/10.1007/s41468-020-00061-z> (cf. p. 110).

- [Dog+18] E. M. DOGO, O. J. AFOLABI, N. I. NWULU et al. « A Comparative Analysis of Gradient Descent-Based Optimization Algorithms on Convolutional Neural Networks ». In : *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*. Déc. 2018, p. 92-99. DOI : [10.1109/CTEMS.2018.8769211](https://doi.org/10.1109/CTEMS.2018.8769211) (cf. p. 54).
- [Doi+19] Hirokazu DOI, Simone SULPIZIO, Gianluca ESPOSITO et al. « Inaudible Components of the Human Infant Cry Influence Haemodynamic Responses in the Breast Region of Mothers ». In : *The Journal of Physiological Sciences* 69.6 (nov. 2019), p. 1085-1096. ISSN : 1880-6562. DOI : [10.1007/s12576-019-00729-x](https://doi.org/10.1007/s12576-019-00729-x). (Visité le 22/05/2023) (cf. p. 22).
- [Don18] Mingwen DONG. « Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification ». In : *2018 Conference on Cognitive Computational Neuroscience*. Philadelphia, Pennsylvania, USA : Cognitive Computational Neuroscience, 2018. DOI : [10.32470/CCN.2018.1153-0](https://doi.org/10.32470/CCN.2018.1153-0). (Visité le 11/07/2022) (cf. p. 24).
- [Dor09] Robert M. DORAZIO. « On Selecting a Prior for the Precision Parameter of Dirichlet Process Mixture Models ». In : *Journal of Statistical Planning and Inference* 139.9 (sept. 2009), p. 3384-3390. ISSN : 0378-3758. DOI : [10.1016/j.jspi.2009.03.009](https://doi.org/10.1016/j.jspi.2009.03.009). (Visité le 27/09/2023) (cf. p. 153).
- [Doz16] Timothy DOZAT. « Incorporating Nesterov Momentum into Adam ». In : *ICLR Workshop*. 2016, p. 4 (cf. p. 54, 69).
- [DHS11] John DUCHI, Elad HAZAN et Yoram SINGER. « Adaptive Subgradient Methods for Online Learning and Stochastic Optimization ». In : *The Journal of Machine Learning Research* 12.null (juill. 2011), p. 2121-2159. ISSN : 1532-4435 (cf. p. 54).
- [EH09] Herbert EDELSBRUNNER et John HARER. *Computational Topology : An Introduction*. AMS Press, 2009 (cf. p. 83, 86, 92, 94, 108, 114, 152).
- [ELZ02] EDELSBRUNNER, LETSCHER et ZOMORODIAN. « Topological Persistence and Simplification ». In : *Discrete & Computational Geometry* 28.4 (nov. 2002), p. 511-533. ISSN : 0179-5376, 1432-0444. DOI : [10.1007/s00454-002-2885-2](https://doi.org/10.1007/s00454-002-2885-2). (Visité le 31/07/2023) (cf. p. 94).
- [Eng+17] Jesse ENGEL, Cinjon RESNICK, Adam ROBERTS et al. « Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders ». In : *arXiv :1704.01279 [cs]* (avr. 2017). arXiv : [1704.01279 \[cs\]](https://arxiv.org/abs/1704.01279). (Visité le 05/01/2021) (cf. p. 24).
- [EM06] Andrew ERRITY et John MCKENNA. « An Investigation of Manifold Learning for Speech Analysis ». In : *Interspeech 2006*. ISCA, sept. 2006, paper 1667-Thu2BuP8-. DOI : [10.21437/Interspeech.2006-628](https://doi.org/10.21437/Interspeech.2006-628). (Visité le 29/08/2023) (cf. p. 29).

- [Esc94] Michael D. ESCOBAR. « Estimating Normal Means with a Dirichlet Process Prior ». In : *Journal of the American Statistical Association* 89.425 (1994), p. 268-277. ISSN : 0162-1459. DOI : [10.2307/2291223](https://doi.org/10.2307/2291223). JSTOR : [2291223](https://www.jstor.org/stable/2291223). (Visité le 11/09/2023) (cf. p. [34](#), [139](#)).
- [Fag+19] Joël FAGOT, Louis-Jean BOË, Frederic BERTHOMIER et al. « The Baboon : A Model for the Study of Language Evolution ». In : *Journal of Human Evolution* 126 (jan. 2019), p. 39-50. ISSN : 00472484. DOI : [10.1016/j.jhevol.2018.10.006](https://doi.org/10.1016/j.jhevol.2018.10.006). (Visité le 02/09/2019) (cf. p. [27](#)).
- [Far+13] Clement FARABET, Camille COUPRIE, Laurent NAJMAN et al. « Learning Hierarchical Features for Scene Labeling ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (août 2013), p. 1915-1929. ISSN : 0162-8828, 2160-9292. DOI : [10.1109/TPAMI.2012.231](https://doi.org/10.1109/TPAMI.2012.231). (Visité le 09/03/2021) (cf. p. [24](#)).
- [Fas+14] Brittany Terese FASY, Fabrizio LECCI, Alessandro RINALDO et al. « Confidence Sets for Persistence Diagrams ». In : *The Annals of Statistics* 42.6 (déc. 2014). ISSN : 0090-5364. DOI : [10.1214/14-AOS1252](https://doi.org/10.1214/14-AOS1252). (Visité le 05/11/2021) (cf. p. [100](#), [110](#)).
- [FMN16] Charles FEFFERMAN, Sanjoy MITTER et Hariharan NARAYANAN. « Testing the Manifold Hypothesis ». In : *Journal of the American Mathematical Society* 29.4 (fév. 2016), p. 983-1049. ISSN : 0894-0347, 1088-6834. DOI : [10.1090/jams/852](https://doi.org/10.1090/jams/852). (Visité le 04/11/2022) (cf. p. [29](#), [151](#)).
- [Fer73] Thomas S. FERGUSON. « A Bayesian Analysis of Some Nonparametric Problems ». In : *The Annals of Statistics* 1.2 (mars 1973), p. 209-230. ISSN : 0090-5364, 2168-8966. DOI : [10.1214/aos/1176342360](https://doi.org/10.1214/aos/1176342360). (Visité le 17/05/2023) (cf. p. [34](#)).
- [Fer18] Massimo FERRI. « Why Topology for Machine Learning and Knowledge Extraction? » In : *Machine Learning and Knowledge Extraction* 1.1 (mai 2018), p. 115-120. ISSN : 2504-4990. DOI : [10.3390/make1010006](https://doi.org/10.3390/make1010006). (Visité le 07/11/2022) (cf. p. [104](#)).
- [FRB22] Tomer FIREAIZEN, Saar RON et Omer BOBROWSKI. « Alarm Sound Detection Using Topological Signal Processing ». In : *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Singapore, Singapore : IEEE, mai 2022, p. 211-215. ISBN : 978-1-66540-540-9. DOI : [10.1109/ICASSP43922.2022.9747228](https://doi.org/10.1109/ICASSP43922.2022.9747228). (Visité le 27/09/2022) (cf. p. [31](#), [101](#), [105](#), [110](#), [113](#), [152](#)).
- [Fla15] Patrick FLANDRIN. « Time-Frequency Filtering Based on Spectrogram Zeros ». In : *IEEE Signal Processing Letters* 22.11 (nov. 2015), p. 2137-2141. ISSN : 1070-9908, 1558-2361. DOI : [10.1109/LSP.2015.2463093](https://doi.org/10.1109/LSP.2015.2463093). (Visité le 16/03/2022) (cf. p. [106](#), [163](#)).

- [Fla18] Patrick FLANDRIN. *Explorations in Time-Frequency Analysis*. 1^{re} éd. Cambridge University Press, sept. 2018. ISBN : 978-1-108-36318-1 978-1-108-42102-7. DOI : [10 . 1017 / 9781108363181](https://doi.org/10.1017/9781108363181). (Visité le 05/09/2022) (cf. p. [106](#), [107](#)).
- [Fox+10] Emily B. FOX, Erik B. SUDDERTH, Michael I. JORDAN et al. « Bayesian Nonparametric Methods for Learning Markov Switching Processes ». In : *IEEE Signal Processing Magazine* 27.6 (nov. 2010), p. 43-54. ISSN : 1558-0792. DOI : [10 . 1109/MSP . 2010 . 937999](https://doi.org/10.1109/MSP.2010.937999) (cf. p. [35](#)).
- [GM14] Stephane GAIFFAS et Bertrand MICHEL. *Sparse Bayesian Unsupervised Learning*. Jan. 2014. DOI : [10 . 48550/arXiv . 1401 . 8017](https://doi.org/10.48550/arXiv.1401.8017). arXiv : [1401 . 8017 \[stat\]](https://arxiv.org/abs/1401.8017). (Visité le 23/08/2023) (cf. p. [32](#)).
- [Gem+17] Jort F. GEMMEKE, Daniel P. W. ELLIS, Dylan FREEDMAN et al. « Audio Set : An Ontology and Human-Labeled Dataset for Audio Events ». In : *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mars 2017, p. 776-780. DOI : [10 . 1109/ICASSP . 2017 . 7952261](https://doi.org/10.1109/ICASSP.2017.7952261) (cf. p. [48](#), [65](#)).
- [Geo23] Georgios P. GEORGIOU. « Comparison of the Prediction Accuracy of Machine Learning Algorithms in Crosslinguistic Vowel Classification ». In : *Scientific Reports* 13.1 (sept. 2023), p. 15594. ISSN : 2045-2322. DOI : [10 . 1038/s41598-023-42818-3](https://doi.org/10.1038/s41598-023-42818-3). (Visité le 28/09/2023) (cf. p. [111](#)).
- [GB12] Samuel J. GERSHMAN et David M. BLEI. « A Tutorial on Bayesian Nonparametric Models ». In : *Journal of Mathematical Psychology* 56.1 (fév. 2012), p. 1-12. ISSN : 0022-2496. DOI : [10 . 1016/j . jmp . 2011 . 08 . 004](https://doi.org/10.1016/j.jmp.2011.08.004). (Visité le 08/05/2023) (cf. p. [35](#), [133](#)).
- [Gha13] Zoubin GHAHRAMANI. « Bayesian Non-Parametrics and the Probabilistic Approach to Modelling ». In : *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences* 371.1984 (fév. 2013), p. 20110553. DOI : [10 . 1098/rsta . 2011 . 0553](https://doi.org/10.1098/rsta.2011.0553). (Visité le 15/05/2023) (cf. p. [34](#)).
- [Gv17] Subhashis GHOSAL et Aad VAN DER VAART. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge : Cambridge University Press, 2017. ISBN : 978-1-139-02983-4. DOI : [10 . 1017 / 9781139029834](https://doi.org/10.1017/9781139029834). (Visité le 21/05/2021) (cf. p. [33](#), [128](#), [131](#), [153](#)).
- [GK18] Marian GIDEA et Yuri KATZ. « Topological Data Analysis of Financial Time Series : Landscapes of Crashes ». In : *Physica A : Statistical Mechanics and its Applications* 491 (fév. 2018), p. 820-834. ISSN : 03784371. DOI : [10 . 1016/j . physa . 2017 . 09 . 028](https://doi.org/10.1016/j.physa.2017.09.028). arXiv : [1703 . 04385](https://arxiv.org/abs/1703.04385). (Visité le 11/10/2021) (cf. p. [30](#)).

- [Gil+17] Jill GILKERSON, Jeffrey A. RICHARDS, Steven F. WARREN et al. « Mapping the Early Language Environment Using All-Day Recordings and Automated Analysis ». In : *American Journal of Speech-Language Pathology* 26.2 (mai 2017), p. 248-265. DOI : [10.1044/2016_AJSLP-15-0169](https://doi.org/10.1044/2016_AJSLP-15-0169). (Visité le 22/05/2023) (cf. p. 22, 166).
- [Gir+14] Ross GIRSHICK, Jeff DONAHUE, Trevor DARRELL et al. « Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation ». In : *arXiv :1311.2524 [cs]* (oct. 2014). arXiv : [1311.2524 \[cs\]](https://arxiv.org/abs/1311.2524). (Visité le 09/03/2021) (cf. p. 24).
- [Glo+23] Pierre GLOAGUEN, Laetitia CHAPEL, Chloé FRIGUET et al. « Scalable Clustering of Segmented Trajectories within a Continuous Time Framework : Application to Maritime Traffic Data ». In : *Machine Learning* 112.6 (juin 2023), p. 1975-2001. ISSN : 1573-0565. DOI : [10.1007/s10994-021-06004-8](https://doi.org/10.1007/s10994-021-06004-8). (Visité le 12/08/2023) (cf. p. 35).
- [GBai] Xavier GLOROT et Yoshua BENGIO. « Understanding the Difficulty of Training Deep Feedforward Neural Networks ». In : *JMLR W&CP : Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*. T. 9. Chia Laguna Resort, Sardinia, Italy, 2010, Mai, p. 249-256 (cf. p. 42).
- [GBB11] Xavier GLOROT, Antoine BORDES et Yoshua BENGIO. « Deep Sparse Rectifier Neural Networks ». In : *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, juin 2011, p. 315-323. (Visité le 01/09/2023) (cf. p. 43).
- [GBC16] Ian GOODFELLOW, Yoshua BENGIO et Aaron COURVILLE. *Deep Learning*. Adaptive Computation and Machine Learning. Cambridge, Massachusetts London, England : The MIT Press, 2016. ISBN : 978-0-262-03561-3 (cf. p. 28, 40, 151).
- [Gor+22] Natàlia GORINA-CARETA, Teresa RIBAS-PRATS, Sonia ARENILLAS-ALCÓN et al. « Neonatal Frequency-Following Responses : A Methodological Framework for Clinical Applications ». In : *Seminars in Hearing* 43.3 (oct. 2022), p. 162-176. ISSN : 0734-0451. DOI : [10.1055/s-0042-1756162](https://doi.org/10.1055/s-0042-1756162). (Visité le 22/01/2023) (cf. p. 18).
- [Gra+20] Zeus GRACIA-TABUENCA, Juan Carlos DÍAZ-PATIÑO, Isaac ARELIO et al. « Topological Data Analysis Reveals Robust Alterations in the Whole-Brain and Frontal Lobe Functional Connectomes in Attention-Deficit/Hyperactivity Disorder ». In : *eNeuro* 7.3 (mai 2020), ENEURO.0543-19.2020. ISSN : 2373-2822. DOI : [10.1523/ENEURO.0543-19.2020](https://doi.org/10.1523/ENEURO.0543-19.2020). (Visité le 31/08/2023) (cf. p. 30).

- [Gu+17] Jiuxiang GU, Zhenhua WANG, Jason KUEN et al. « Recent Advances in Convolutional Neural Networks ». In : *arXiv:1512.07108 [cs]* (oct. 2017). arXiv : [1512.07108 \[cs\]](https://arxiv.org/abs/1512.07108). (Visité le 14/05/2020) (cf. p. 24).
- [Gua+10] Yue GUAN, Jennifer G. DY, Donglin NIU et al. « Variational Inference for Nonparametric Multiple Clustering ». In : *MultiClust Workshop, KDD-2010*. Citeseer, 2010, p. 67-125 (cf. p. 35).
- [Gug+14] Alessandra GUGLIELMI, Francesca IEVA, Anna M. PAGANONI et al. « Semiparametric Bayesian Models for Clustering and Classification in the Presence of Unbalanced In-Hospital Survival ». In : *Journal of the Royal Statistical Society Series C : Applied Statistics* 63.1 (jan. 2014), p. 25-46. ISSN : 0035-9254. DOI : [10.1111/rssc.12021](https://doi.org/10.1111/rssc.12021). (Visité le 17/09/2023) (cf. p. 35).
- [Guz+20] Andrey GUZHOV, Federico RAUE, Jörn HEES et al. « ESResNet : Environmental Sound Classification Based on Visual Domain Models ». In : *arXiv:2004.07301 [cs, eess]* (avr. 2020). arXiv : [2004.07301 \[cs, eess\]](https://arxiv.org/abs/2004.07301). (Visité le 09/03/2021) (cf. p. 25).
- [HB22] Shaked HAIM MEIROM et Omer BOBROWSKI. « Unsupervised Geometric and Topological Approaches for Cross-Lingual Sentence Representation and Comparison ». In : *Proceedings of the 7th Workshop on Representation Learning for NLP*. Dublin, Ireland : Association for Computational Linguistics, mai 2022, p. 173-183. DOI : [10.18653/v1/2022.repl4nlp-1.18](https://doi.org/10.18653/v1/2022.repl4nlp-1.18). (Visité le 31/08/2023) (cf. p. 31).
- [HR18] Boris HANIN et David ROLNICK. « How to Start Training : The Effect of Initialization and Architecture ». In : *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Red Hook, NY, USA : Curran Associates Inc., déc. 2018, p. 569-579. (Visité le 02/09/2023) (cf. p. 42).
- [HR03] Betty HART et Todd R RISLEY. « The Early Catastrophe ». In : (2003), p. 6 (cf. p. 166).
- [He+15] Kaiming HE, Xiangyu ZHANG, Shaoqing REN et al. « Delving Deep into Rectifiers : Surpassing Human-Level Performance on ImageNet Classification ». In : *arXiv:1502.01852 [cs]* (fév. 2015). arXiv : [1502.01852 \[cs\]](https://arxiv.org/abs/1502.01852). (Visité le 09/03/2021) (cf. p. 43, 69).
- [HG17] Dan HENDRYCKS et Kevin GIMPEL. *Adjusting for Dropout Variance in Batch Normalization and Weight Initialization*. Mars 2017. DOI : [10.48550/arXiv.1607.02488](https://doi.org/10.48550/arXiv.1607.02488). arXiv : [1607.02488 \[cs\]](https://arxiv.org/abs/1607.02488). (Visité le 02/09/2023) (cf. p. 42).
- [HMR21] Felix HENSEL, Michael MOOR et Bastian RIECK. « A Survey of Topological Machine Learning Methods ». In : *Frontiers in Artificial Intelligence* 4 (mai 2021), p. 681108. ISSN : 2624-8212. DOI : [10.3389/frai.2021.681108](https://doi.org/10.3389/frai.2021.681108). (Visité le 19/04/2022) (cf. p. 101, 110, 123, 147).

- [Her+17] Shawn HERSHEY, Sourish CHAUDHURI, Daniel P. W. ELLIS et al. « CNN Architectures for Large-Scale Audio Classification ». In : *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mars 2017, p. 131-135. DOI : [10.1109/ICASSP.2017.7952132](https://doi.org/10.1109/ICASSP.2017.7952132) (cf. p. 65).
- [Hjo+10] N.L. HJORT, C. HOLMES, P. MÜLLER et al. *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010. ISBN : 978-1-139-48460-2 (cf. p. 128, 153).
- [How+17] Andrew G. HOWARD, Menglong ZHU, Bo CHEN et al. « MobileNets : Efficient Convolutional Neural Networks for Mobile Vision Applications ». In : *arXiv :1704.04861 [cs]* (avr. 2017). arXiv : [1704.04861 \[cs\]](https://arxiv.org/abs/1704.04861). (Visité le 01/03/2022) (cf. p. 47-49, 65).
- [HOA22] Kevin H. HUANG, Peter ORBANZ et Morgane AUSTERN. *Quantifying the Effects of Data Augmentation*. Déc. 2022. DOI : [10.48550/arXiv.2202.09134](https://doi.org/10.48550/arXiv.2202.09134). arXiv : [2202.09134 \[cs, math, stat\]](https://arxiv.org/abs/2202.09134). (Visité le 18/09/2023) (cf. p. 56).
- [IS15] Sergey IOFFE et Christian SZEGEDY. « Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift ». In : *arXiv :1502.03167 [cs]* (mars 2015). arXiv : [1502.03167 \[cs\]](https://arxiv.org/abs/1502.03167). (Visité le 27/04/2020) (cf. p. 57, 69).
- [Ism+19] Hassan ISMAIL FAWAZ, Germain FORESTIER, Jonathan WEBER et al. « Deep Learning for Time Series Classification : A Review ». In : *Data Mining and Knowledge Discovery* 33.4 (juill. 2019), p. 917-963. ISSN : 1384-5810, 1573-756X. DOI : [10.1007/s10618-019-00619-1](https://doi.org/10.1007/s10618-019-00619-1). (Visité le 09/03/2021) (cf. p. 24).
- [Jar17] Alejandro JARA. « Theory and Computations for the Dirichlet Process and Related Models : An Overview ». In : *International Journal of Approximate Reasoning* 81 (fév. 2017), p. 128-146. ISSN : 0888-613X. DOI : [10.1016/j.ijar.2016.11.008](https://doi.org/10.1016/j.ijar.2016.11.008). (Visité le 15/05/2023) (cf. p. 34).
- [Jha17] Yuna JHANG. « Emergence of Functional Flexibility in Infant Vocalizations of the First 3 Months ». In : *Frontiers in Psychology* 8 (2017), p. 11 (cf. p. 20, 146, 160, 165).
- [Ji+13] Shuiwang JI, Wei XU, Ming YANG et al. « 3D Convolutional Neural Networks for Human Action Recognition ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (jan. 2013), p. 221-231. ISSN : 0162-8828, 2160-9292. DOI : [10.1109/TPAMI.2012.59](https://doi.org/10.1109/TPAMI.2012.59). (Visité le 15/01/2021) (cf. p. 24).
- [Jon01] Donald R. JONES. « A Taxonomy of Global Optimization Methods Based on Response Surfaces ». In : *Journal of Global Optimization* 21.4 (déc. 2001), p. 345-383. ISSN : 1573-2916. DOI : [10.1023/A:1012771025575](https://doi.org/10.1023/A:1012771025575). (Visité le 05/09/2023) (cf. p. 61).

- [KGW11] Maria KALLI, Jim E. GRIFFIN et Stephen G. WALKER. « Slice Sampling Mixture Models ». In : *Statistics and Computing* 21.1 (jan. 2011), p. 93-105. ISSN : 1573-1375. DOI : [10.1007/s11222-009-9150-y](https://doi.org/10.1007/s11222-009-9150-y). (Visité le 05/06/2023) (cf. p. 34, 139).
- [Kha+19] A. KHAMPARIA, D. GUPTA, N. G. NGUYEN et al. « Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network ». In : *IEEE Access* 7 (2019), p. 7717-7727. ISSN : 2169-3536. DOI : [10.1109/ACCESS.2018.2888882](https://doi.org/10.1109/ACCESS.2018.2888882) (cf. p. 25).
- [Kim+20] Kwangho KIM, Manzil ZAHEER, Frederic CHAZAL et al. « PLLay : Efficient Topological Layer Based on Persistence Landscapes ». In : *NeurIPS 2020 - 34th Conference on Neural Information Processing Systems*. Vancouver / Virtuel, Canada, 2020, p. 13 (cf. p. 101, 110, 125).
- [KB17] Diederik P. KINGMA et Jimmy BA. « Adam : A Method for Stochastic Optimization ». In : *arXiv :1412.6980 [cs]* (jan. 2017). arXiv : [1412.6980 \[cs\]](https://arxiv.org/abs/1412.6980). (Visité le 05/02/2020) (cf. p. 54, 59, 69).
- [KKN22] Lucas KOCK, Nadja KLEIN et David J. NOTT. « Variational Inference and Sparsity in High-Dimensional Deep Gaussian Mixture Models ». In : *Statistics and Computing* 32.5 (sept. 2022), p. 70. ISSN : 1573-1375. DOI : [10.1007/s11222-022-10132-z](https://doi.org/10.1007/s11222-022-10132-z). (Visité le 04/08/2023) (cf. p. 33).
- [KBT19] Yunus KORKMAZ, Aytuğ BOYACI et Türker TUNCER. « Turkish Vowel Classification Based on Acoustical and Decompositional Features Optimized by Genetic Algorithm ». In : *Applied Acoustics* 154 (nov. 2019), p. 28-35. ISSN : 0003-682X. DOI : [10.1016/j.apacoust.2019.04.027](https://doi.org/10.1016/j.apacoust.2019.04.027). (Visité le 28/09/2023) (cf. p. 111).
- [Kri+21] Aditi S. KRISHNAPRIYAN, Joseph MONTOYA, Maciej HARANCZYK et al. « Machine Learning with Persistent Homology and Chemical Word Embeddings Improves Prediction Accuracy and Interpretability in Metal-Organic Frameworks ». In : *Scientific Reports* 11.1 (avr. 2021), p. 8888. ISSN : 2045-2322. DOI : [10.1038/s41598-021-88027-8](https://doi.org/10.1038/s41598-021-88027-8). (Visité le 31/08/2023) (cf. p. 30).
- [KSH12] Alex KRIZHEVSKY, Ilya SUTSKEVER et Geoffrey E HINTON. « ImageNet Classification with Deep Convolutional Neural Networks ». In : *Advances in Neural Information Processing Systems*. T. 25. Curran Associates, Inc., 2012. (Visité le 02/09/2023) (cf. p. 47).
- [KM82] P. K. KUHL et A. N. MELTZOFF. « The Bimodal Perception of Speech in Infancy ». In : *Science (New York, N.Y.)* 218.4577 (déc. 1982), p. 1138-1141. ISSN : 0036-8075. DOI : [10.1126/science.7146899](https://doi.org/10.1126/science.7146899) (cf. p. 21).
- [KM96] P. K. KUHL et A. N. MELTZOFF. « Infant Vocalizations in Response to Speech : Vocal Imitation and Developmental Change ». In : *The Journal of the Acoustical Society of America* 100.4 Pt 1 (oct. 1996), p. 2425-2438. ISSN : 0001-4966. DOI : [10.1121/1.417951](https://doi.org/10.1121/1.417951) (cf. p. 21).

- [Kuh04] Patricia K. KUHL. « Early Language Acquisition : Cracking the Speech Code ». In : *Nature Reviews Neuroscience* 5.11 (nov. 2004), p. 831-843. ISSN : 1471-003X, 1471-0048. DOI : [10.1038/nrn1533](https://doi.org/10.1038/nrn1533). (Visité le 20/06/2019) (cf. p. 19, 146).
- [Kum17] Siddharth Krishna KUMAR. « On Weight Initialization in Deep Neural Networks ». In : *arXiv :1704.08863 [cs]* (mai 2017). arXiv : [1704.08863 \[cs\]](https://arxiv.org/abs/1704.08863). (Visité le 01/03/2021) (cf. p. 43, 69).
- [KHH19] Tsuyoshi KUNIHAMA, Carolyn T. HALPERN et Amy H. HERRING. « Non-Parametric Bayes Models for Mixed Scale Longitudinal Surveys ». In : *Journal of the Royal Statistical Society Series C : Applied Statistics* 68.4 (août 2019), p. 1091-1109. ISSN : 0035-9254. DOI : [10.1111/rssc.12348](https://doi.org/10.1111/rssc.12348). (Visité le 17/09/2023) (cf. p. 35).
- [Kus64] H. J. KUSHNER. « A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise ». In : *Journal of Basic Engineering* 86.1 (mars 1964), p. 97-106. ISSN : 0021-9223. DOI : [10.1115/1.3653121](https://doi.org/10.1115/1.3653121). (Visité le 05/09/2023) (cf. p. 61).
- [Lan+19] Sigrun LANG, Katrin D. BARTL-POKORNY, Florian B. POKORNY et al. « Canonical Babbling : A Marker for Earlier Identification of Late Detected Developmental Disorders? » In : *Current Developmental Disorders Reports* 6.3 (sept. 2019), p. 111-118. ISSN : 2196-2987. DOI : [10.1007/s40474-019-00166-w](https://doi.org/10.1007/s40474-019-00166-w). (Visité le 22/03/2022) (cf. p. 21, 165).
- [LeC+89] Y. LECUN, B. BOSER, J. S. DENKER et al. « Backpropagation Applied to Handwritten Zip Code Recognition ». In : *Neural Computation* 1.4 (déc. 1989), p. 541-551. ISSN : 0899-7667. DOI : [10.1162/neco.1989.1.4.541](https://doi.org/10.1162/neco.1989.1.4.541) (cf. p. 45, 46).
- [Lec85] Yann LECUN. « Une Procédure d'apprentissage Pour Réseau à Seuil Asymétrique (A Learning Scheme for Asymmetric Threshold Networks) ». In : *Proceedings of Cognitiva 85, Paris, France* (1985), p. 599-604 (cf. p. 53).
- [LB95] Yann LECUN et Yoshua BENGIO. « Convolutional Networks for Images, Speech, and Time-Series ». In : *The Handbook of Brain Theory and Neural Networks*. M.A. Arbib. MIT Press, 1995 (cf. p. 45).
- [LBH15] Yann LECUN, Yoshua BENGIO et Geoffrey HINTON. « Deep Learning ». In : *Nature* 521.7553 (mai 2015), p. 436-444. ISSN : 0028-0836, 1476-4687. DOI : [10.1038/nature14539](https://doi.org/10.1038/nature14539). (Visité le 04/04/2020) (cf. p. 24, 45, 47, 65).
- [LeC+12] Yann A. LECUN, Léon BOTTOU, Genevieve B. ORR et al. « Efficient Back-Prop ». In : *Neural Networks : Tricks of the Trade*. Sous la dir. de Grégoire MONTAVON, Geneviève B. ORR et Klaus-Robert MÜLLER. T. 7700. Berlin, Heidelberg : Springer Berlin Heidelberg, 2012, p. 9-48. ISBN : 978-3-642-35288-1 978-3-642-35289-8. DOI : [10.1007/978-3-642-35289-8_3](https://doi.org/10.1007/978-3-642-35289-8_3). (Visité le 22/03/2021) (cf. p. 54).

- [Lee+11] Hyekyoung LEE, Moo K. CHUNG, Hyejin KANG, Bung-Nyun KIM et al. « Discriminative Persistent Homology of Brain Networks ». In : *2011 IEEE International Symposium on Biomedical Imaging : From Nano to Macro*. Mars 2011, p. 841-844. DOI : [10.1109/ISBI.2011.5872535](https://doi.org/10.1109/ISBI.2011.5872535) (cf. p. 30).
- [Lee+14] Hyekyoung LEE, Moo K. CHUNG, Hyejin KANG et Dong Soo LEE. « Hole Detection in Metabolic Connectivity of Alzheimer's Disease Using k-Laplacian ». In : *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*. Sous la dir. de Polina GOLLAND, Nobuhiko HATA, Christian BARILLOT et al. Lecture Notes in Computer Science. Cham : Springer International Publishing, 2014, p. 297-304. ISBN : 978-3-319-10443-0. DOI : [10.1007/978-3-319-10443-0_38](https://doi.org/10.1007/978-3-319-10443-0_38) (cf. p. 30).
- [Lee+17a] Jongpil LEE, Taejun KIM, Jiyoung PARK et al. « Raw Waveform-based Audio Classification Using Sample-level CNN Architectures ». In : *arXiv:1712.00866 [cs, eess]* (déc. 2017). arXiv : [1712.00866 \[cs, eess\]](https://arxiv.org/abs/1712.00866). (Visité le 01/07/2020) (cf. p. 24).
- [Lee+17b] Jongpil LEE, Jiyoung PARK, Keunhyoung Luke KIM et al. « Sample-Level Deep Convolutional Neural Networks for Music Auto-tagging Using Raw Waveforms ». In : *arXiv:1703.01789 [cs]* (mai 2017). arXiv : [1703.01789 \[cs\]](https://arxiv.org/abs/1703.01789). (Visité le 09/03/2021) (cf. p. 24).
- [LDM10] Sue Ann S. LEE, Barbara DAVIS et Peter MACNEILAGE. « Universal Production Patterns and Ambient Language Influences in Babbling : A Cross-Linguistic Study of Korean- and English-learning Infants* ». In : *Journal of Child Language* 37.2 (mars 2010), p. 293-318. ISSN : 1469-7602, 0305-0009. DOI : [10.1017/S0305000909009532](https://doi.org/10.1017/S0305000909009532). (Visité le 03/10/2023) (cf. p. 20, 146).
- [Leh+15] L. H. LEHMAN, M. J. JOHNSON, S. NEMATI et al. « Bayesian Nonparametric Learning of Switching Dynamics in Cohort Physiological Time Series : Application in Critical Care Patient Monitoring ». In : *Advanced State Space Methods for Neural and Clinical Data*. Sous la dir. de Zhe CHEN. 1^{re} éd. Cambridge University Press, sept. 2015, p. 257-282. ISBN : 978-1-107-07919-9 978-1-139-94143-3. DOI : [10.1017/CB09781139941433.012](https://doi.org/10.1017/CB09781139941433.012). (Visité le 18/09/2023) (cf. p. 35).
- [LNZ22] Jeremy. LEVY, Alexander. NAITSAT et Yehoshua Y ZEEVI. « Classification of audio signals using spectrogram surfaces and extrinsic distortion measures ». In : *EURASIP Journal on Advances in Signal Processing* 10 (2022), p. 83-103. ISSN : 1573-1405. DOI : [10.1186/s13634-022-00933-9](https://doi.org/10.1186/s13634-022-00933-9) (cf. p. 106).
- [Li+21] Xiao LI, Michele GUINDANI, Chaan S. NG et al. « A Bayesian Nonparametric Model for Textural Pattern Heterogeneity ». In : *Journal of the Royal Statistical Society Series C : Applied Statistics* 70.2 (mars 2021), p. 459-480. ISSN : 0035-9254. DOI : [10.1111/rssc.12469](https://doi.org/10.1111/rssc.12469). (Visité le 17/09/2023) (cf. p. 35).

- [LCK19] Xinyu LI, Venkata CHEBIYYAM et Katrin KIRCHHOFF. « Multi-Stream Network With Temporal Attention For Environmental Sound Classification ». In : *arXiv :1901.08608 [cs, eess]* (jan. 2019). arXiv : [1901 . 08608 \[cs, eess\]](https://arxiv.org/abs/1901.08608). (Visité le 09/03/2021) (cf. p. 25).
- [LSG19] Yuelin LI, Elizabeth SCHOFIELD et Mithat GÖNEN. « A Tutorial on Dirichlet Process Mixture Modeling ». In : *Journal of Mathematical Psychology* 91 (août 2019), p. 128-144. ISSN : 0022-2496. DOI : [10 . 1016/ j . jmp . 2019 . 04 . 004](https://doi.org/10.1016/j.jmp.2019.04.004) (cf. p. 35).
- [Lic+22] Guilherme LICHAND, Onicio LEAL NETO, John PHUKA et al. *The Early Childhood Development Replication Crisis, and How Wearable Technologies Could Help Overcome It*. SSRN Scholarly Paper. Rochester, NY, juill. 2022. DOI : [10 . 2139/ ssrn . 4162049](https://doi.org/10.2139/ssrn.4162049). (Visité le 17/09/2023) (cf. p. 22).
- [LJY16] Jen-Yu LIU, Shyh-Kang JENG et Yi-Hsuan YANG. « Applying Topological Persistence in Convolutional Neural Network for Music Audio Signals ». In : *arXiv :1608.07373 [cs]* (août 2016). arXiv : [1608 . 07373 \[cs\]](https://arxiv.org/abs/1608.07373). (Visité le 07/05/2021) (cf. p. 31, 101, 105, 110, 125).
- [Lo84] Albert Y. LO. « On a Class of Bayesian Nonparametric Estimates : I. Density Estimates ». In : *The Annals of Statistics* 12.1 (1984), p. 351-357. ISSN : 0090-5364. JSTOR : [2241054](https://www.jstor.org/stable/2241054). (Visité le 11/09/2023) (cf. p. 34).
- [Loc+23] Marguerite LOCKHART-BOURON, Andrey ANIKIN, Katarzyna PISANSKI et al. « Infant Cries Convey Both Stable and Dynamic Information about Age and Identity ». In : *Communications Psychology* 1.1 (oct. 2023), p. 1-15. ISSN : 2731-9121. DOI : [10 . 1038/ s44271 - 023 - 00022 - z](https://doi.org/10.1038/s44271-023-00022-z). (Visité le 06/11/2023) (cf. p. 21, 146).
- [Log+19] Jessica A. R. LOGAN, Laura M. JUSTICE, Melike YUMUŞ et al. « When Children Are Not Read to at Home : The Million Word Gap ». In : *Journal of developmental and behavioral pediatrics : JDBP* 40.5 (juin 2019), p. 383-386. ISSN : 1536-7312. DOI : [10 . 1097/ DBP . 0000000000000657](https://doi.org/10.1097/DBP.0000000000000657) (cf. p. 166).
- [Mac01] Steven N MACEACHERN. « Decision Theoretic Aspects of Dependent Nonparametric Processes ». In : *Bayesian Methods with Applications to Science, Policy, and Official Statistics* (2001), p. 551-560 (cf. p. 35).
- [Mac16] Steven N. MACEACHERN. « Nonparametric Bayesian Methods : A Gentle Introduction and Overview ». In : *Communications for Statistical Applications and Methods* 23.6 (nov. 2016), p. 445-466. ISSN : 2287-7843. DOI : [10 . 5351/ CSAM . 2016 . 23 . 6 . 445](https://doi.org/10.5351/CSAM.2016.23.6.445). (Visité le 17/09/2023) (cf. p. 34).
- [Mam+09] Birgit MAMPE, Angela D. FRIEDERICI, Anne CHRISTOPHE et al. « Newborns' Cry Melody Is Shaped by Their Native Language ». In : *Current biology : CB* 19.23 (déc. 2009), p. 1994-1997. ISSN : 1879-0445. DOI : [10 . 1016/ j . cub . 2009 . 09 . 064](https://doi.org/10.1016/j.cub.2009.09.064) (cf. p. 20, 146).

- [Man+16] Vikash MANSINGHKA, Patrick SHAFTO, Eric JONAS et al. « CrossCat : A Fully Bayesian Nonparametric Method for Analyzing Heterogeneous, High Dimensional Data ». In : *Journal of Machine Learning Research* 17.138 (2016), p. 1-49. ISSN : 1533-7928. (Visité le 17/09/2023) (cf. p. 35).
- [Mar19] Charles C. MARGOSSIAN. « A Review of Automatic Differentiation and Its Efficient Implementation ». In : *WIREs Data Mining and Knowledge Discovery* 9.4 (2019), e1305. ISSN : 1942-4795. DOI : [10.1002/widm.1305](https://doi.org/10.1002/widm.1305). (Visité le 03/09/2023) (cf. p. 54).
- [MMN22] Vasileios MAROULAS, Cassie Putman MICUCCI et Farzana NASRIN. « Bayesian Topological Learning for Classifying the Structure of Biological Networks ». In : *Bayesian Analysis* 17.3 (sept. 2022). ISSN : 1936-0975. DOI : [10.1214/21-BA1270](https://doi.org/10.1214/21-BA1270). (Visité le 16/01/2023) (cf. p. 30, 101).
- [MMO19] Vasileios MAROULAS, Joshua L. MIKE et Christopher OBALLE. « Nonparametric Estimation of Probability Density Functions of Random Persistence Diagrams ». In : *Journal of Machine Learning Research* 20.151 (2019), p. 1-49. ISSN : 1533-7928. (Visité le 03/01/2023) (cf. p. 30, 101, 147).
- [MNO20] Vasileios MAROULAS, Farzana NASRIN et Christopher OBALLE. « A Bayesian Framework for Persistent Homology ». In : *SIAM Journal on Mathematics of Data Science* 2.1 (jan. 2020), p. 48-74. ISSN : 2577-0187. DOI : [10.1137/19M1268719](https://doi.org/10.1137/19M1268719). (Visité le 23/07/2022) (cf. p. 30, 101, 110).
- [MHB15] Brian MCFEE, Eric J HUMPHREY et Juan P BELLO. « A Software Framework for Musical Data Augmentation ». In : *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015* (2015), p. 248-254 (cf. p. 56, 67).
- [MRB15] James MCINERNEY, Rajesh RANGANATH et David BLEI. « The Population Posterior and Bayesian Modeling on Streams ». In : *Advances in Neural Information Processing Systems*. T. 28. Curran Associates, Inc., 2015. (Visité le 17/09/2023) (cf. p. 35).
- [MHM20] Leland MCINNES, John HEALY et James MELVILLE. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. Sept. 2020. arXiv : [1802.03426](https://arxiv.org/abs/1802.03426) [cs, stat]. (Visité le 10/10/2022) (cf. p. 29, 112, 151, 154).
- [Meg+19a] Khadidja MEGUELATI, Benedicte FONTEZ, Nadine HILGERT et al. « Dirichlet Process Mixture Models Made Scalable and Effective by Means of Massive Distribution ». In : *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. Limassol Cyprus : ACM, avr. 2019, p. 502-509. ISBN : 978-1-4503-5933-7. DOI : [10.1145/3297280.3297327](https://doi.org/10.1145/3297280.3297327). (Visité le 30/05/2023) (cf. p. 35).

- [Meg+19b] Khadidja MEGUELATI, Benedicte FONTEZ, Nadine HILGERT et al. « High Dimensional Data Clustering by Means of Distributed Dirichlet Process Mixture Models ». In : *2019 IEEE International Conference on Big Data (Big Data)*. Los Angeles, CA, USA : IEEE, déc. 2019, p. 890-899. ISBN : 978-1-72810-858-2. DOI : [10.1109/BigData47090.2019.9006065](https://doi.org/10.1109/BigData47090.2019.9006065). (Visité le 24/05/2023) (cf. p. 35).
- [Meh+88] Jacques MEHLER, Peter JUSZYK, Ghislaine LAMBERTZ et al. « A Precursor of Language Acquisition in Young Infants ». In : *Cognition* 29.2 (juill. 1988), p. 143-178. ISSN : 0010-0277. DOI : [10.1016/0010-0277\(88\)90035-2](https://doi.org/10.1016/0010-0277(88)90035-2). (Visité le 01/10/2023) (cf. p. 18).
- [Meh+17] Soroush MEHRI, Kundan KUMAR, Ishaan GULRAJANI et al. « SampleRNN : An Unconditional End-to-End Neural Audio Generation Model ». In : *arXiv :1612.07837 [cs]* (fév. 2017). arXiv : [1612.07837 \[cs\]](https://arxiv.org/abs/1612.07837). (Visité le 13/04/2021) (cf. p. 24).
- [Mic15] Bertrand MICHEL. « A Statistical Approach to Topological Data Analysis ». Thesis. UPMC Université Paris VI, nov. 2015. (Visité le 23/08/2023) (cf. p. 29, 99).
- [Mil+22] Manuel MILLING, Florian B. POKORNY, Katrin D. BARTL-POKORNY et al. « Is Speech the New Blood? Recent Progress in AI-Based Disease Detection From Audio in a Nutshell ». In : *Frontiers in Digital Health* 4 (mai 2022), p. 886615. ISSN : 2673-253X. DOI : [10.3389/fdgth.2022.886615](https://doi.org/10.3389/fdgth.2022.886615). (Visité le 08/09/2022) (cf. p. 23, 146).
- [Mol+22] Marco MOLINARI, Andrea CREMASCHI, Maria DE IORIO et al. « Bayesian Nonparametric Modelling of Multiple Graphs with an Application to Ethnic Metabolic Differences ». In : *Journal of the Royal Statistical Society Series C: Applied Statistics* 71.5 (nov. 2022), p. 1181-1204. ISSN : 0035-9254. DOI : [10.1111/rssc.12570](https://doi.org/10.1111/rssc.12570). (Visité le 17/09/2023) (cf. p. 35).
- [Moo+20] Michael MOOR, Max HORN, Bastian RIECK et al. « Topological Autoencoders ». In : *arXiv :1906.00722 [Cs, Math, Stat]*. T. PMLR 119. 2020, p. 7045-7054. arXiv : [1906.00722 \[cs, math, stat\]](https://arxiv.org/abs/1906.00722). (Visité le 16/03/2022) (cf. p. 101, 124, 161, 165).
- [MW18] Lydia MORGAN et Yvonne E. WREN. « A Systematic Review of the Literature on Early Vocalizations and Babbling Patterns in Young Children ». In : *Communication Disorders Quarterly* 40.1 (nov. 2018), p. 3-14. ISSN : 1525-7401, 1538-4837. DOI : [10.1177/1525740118760215](https://doi.org/10.1177/1525740118760215). (Visité le 08/09/2022) (cf. p. 20, 146).
- [Mos13] Karl MOSLER. « Depth Statistics ». In : *Robustness and Complex Data Structures : Festschrift in Honour of Ursula Gather*. Sous la dir. de Claudia BECKER, Roland FRIED et Sonja KUHN. Berlin, Heidelberg : Springer, 2013, p. 17-34. ISBN : 978-3-642-35494-6. DOI : [10.1007/978-3-642-35494-6_2](https://doi.org/10.1007/978-3-642-35494-6_2). (Visité le 27/09/2023) (cf. p. 154).

- [MM22] Karl MOSLER et Pavlo MOZHAROVSKIY. « Choosing Among Notions of Multivariate Depth Statistics ». In : *Statistical Science* 37.3 (août 2022), p. 348-368. ISSN : 0883-4237, 2168-8745. DOI : [10.1214/21-STS827](https://doi.org/10.1214/21-STS827). (Visité le 27/09/2023) (cf. p. 154).
- [MQ04] Peter MÜLLER et Fernando A. QUINTANA. « Nonparametric Bayesian Data Analysis ». In : *Statistical Science* 19.1 (fév. 2004). ISSN : 0883-4237. DOI : [10.1214/088342304000000017](https://doi.org/10.1214/088342304000000017). (Visité le 15/05/2023) (cf. p. 34).
- [Mur22] Kevin P. MURPHY. *Probabilistic Machine Learning: An Introduction*. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts : The MIT Press, 2022. ISBN : 978-0-262-04682-4 (cf. p. 40).
- [Mur23] Kevin P. MURPHY. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023 (cf. p. 40, 128).
- [Nas+19] Farzana NASRIN, Christopher OBALLE, David L. BOOTHE et al. « Bayesian Topological Learning for Brain State Classification ». In : *arXiv:1912.08348 [cs, stat]* (déc. 2019). arXiv : [1912.08348 \[cs, stat\]](https://arxiv.org/abs/1912.08348). (Visité le 06/05/2021) (cf. p. 30, 105, 147).
- [Nea00] Radford M. NEAL. « Markov Chain Sampling Methods for Dirichlet Process Mixture Models ». In : *Journal of Computational and Graphical Statistics* 9.2 (2000), p. 249-265. ISSN : 1061-8600. DOI : [10.2307/1390653](https://doi.org/10.2307/1390653). JSTOR : [1390653](https://www.jstor.org/stable/1390653). (Visité le 05/04/2023) (cf. p. 34, 139, 141, 142, 153).
- [Ngu+18] Thi Thu Thuy NGUYEN, Tien Thanh NGUYEN, Alan Wee-Chung LIEW et al. « Variational Inference Based Bayes Online Classifiers with Concept Drift Adaptation ». In : *Pattern Recognition* 81 (sept. 2018), p. 280-293. ISSN : 00313203. DOI : [10.1016/j.patcog.2018.04.007](https://doi.org/10.1016/j.patcog.2018.04.007). (Visité le 24/07/2022) (cf. p. 24, 64).
- [Ngu+16] Tien Thanh NGUYEN, Thi Thu Thuy NGUYEN, Xuan Cuong PHAM et al. « A Novel Combining Classifier Method Based on Variational Inference ». In : *Pattern Recognition* 49 (jan. 2016), p. 198-212. ISSN : 00313203. DOI : [10.1016/j.patcog.2015.06.016](https://doi.org/10.1016/j.patcog.2015.06.016). (Visité le 27/07/2022) (cf. p. 24).
- [NLC11] Monica NICOLAU, Arnold J. LEVINE et Gunnar CARLSSON. « Topology Based Data Analysis Identifies a Subgroup of Breast Cancers with a Unique Mutational Profile and Excellent Survival ». In : *Proceedings of the National Academy of Sciences* 108.17 (avr. 2011), p. 7265-7270. DOI : [10.1073/pnas.1102826108](https://doi.org/10.1073/pnas.1102826108). (Visité le 26/08/2023) (cf. p. 30).
- [NSW08] Partha NIYOGI, Stephen SMALE et Shmuel WEINBERGER. « Finding the Homology of Submanifolds with High Confidence from Random Samples ». In : *Discrete & Computational Geometry* 39.1 (mars 2008), p. 419-441. ISSN : 1432-0444. DOI : [10.1007/s00454-008-9053-2](https://doi.org/10.1007/s00454-008-9053-2). (Visité le 29/08/2023) (cf. p. 29).

- [Oik+19] Tuomas OIKARINEN, Karthik SRINIVASAN, Olivia MEISNER et al. « Deep Convolutional Network for Animal Sound Classification and Source Attribution Using Dual Audio Recordings ». In : *The Journal of the Acoustical Society of America* 145.2 (fév. 2019), p. 654-662. ISSN : 0001-4966. DOI : [10.1121/1.5087827](https://doi.org/10.1121/1.5087827). (Visité le 20/04/2021) (cf. p. 24, 65).
- [Oll+98] D. K. OLLER, R. E. EILERS, A. R. NEAL et al. « Late Onset Canonical Babbling : A Possible Early Marker of Abnormal Development ». In : *American journal of mental retardation : AJMR* 103.3 (nov. 1998), p. 249-263. ISSN : 0895-8017. DOI : [10.1352/0895-8017\(1998\)103<0249:LOCBAP>2.0.CO;2](https://doi.org/10.1352/0895-8017(1998)103<0249:LOCBAP>2.0.CO;2) (cf. p. 20, 21, 146).
- [Oll+10] D. K. OLLER, P. NIYOGI, S. GRAY et al. « Automated Vocal Analysis of Naturalistic Recordings from Children with Autism, Language Delay, and Typical Development ». In : *Proceedings of the National Academy of Sciences of the United States of America* 107.30 (juill. 2010), p. 13354-13359. ISSN : 1091-6490. DOI : [10.1073/pnas.1003882107](https://doi.org/10.1073/pnas.1003882107) (cf. p. 22, 146).
- [Oll00] D. Kimbrough OLLER. *The Emergence of the Speech Capacity*. The Emergence of the Speech Capacity. Mahwah, NJ, US : Lawrence Erlbaum Associates Publishers, 2000, p. xvii, 428. ISBN : 978-0-8058-2628-9 978-0-8058-2629-6 (cf. p. 19, 146, 165).
- [Oll+13] D. Kimbrough OLLER, Eugene H. BUDER, Heather L. RAMSDELL et al. « Functional Flexibility of Infant Vocalization and the Emergence of Language ». In : *Proceedings of the National Academy of Sciences* 110.16 (avr. 2013), p. 6318-6323. ISSN : 0027-8424, 1091-6490. DOI : [10.1073/pnas.1300337110](https://doi.org/10.1073/pnas.1300337110). (Visité le 28/04/2022) (cf. p. 20, 146, 160).
- [Oll+99] D. Kimbrough OLLER, Rebecca E EILERS, A. Rebecca NEAL et al. « Precursors to Speech in Infancy : The Prediction of Speech and Language Disorders ». In : *Journal of Communication Disorders* 32.4 (juill. 1999), p. 223-245. ISSN : 0021-9924. DOI : [10.1016/S0021-9924\(99\)00013-1](https://doi.org/10.1016/S0021-9924(99)00013-1). (Visité le 30/09/2023) (cf. p. 20, 21, 146).
- [Oor+16] Aaron van den OORD, Sander DIELEMAN, Heiga ZEN et al. « WaveNet : A Generative Model for Raw Audio ». In : *arXiv:1609.03499 [cs]* (sept. 2016). arXiv : [1609.03499 \[cs\]](https://arxiv.org/abs/1609.03499). (Visité le 23/04/2020) (cf. p. 24, 46, 47).
- [OT10] Peter ORBANZ et Yee Whye TEH. « Bayesian Nonparametric Models ». In : *Encyclopedia of Machine Learning*. Sous la dir. de Claude SAMMUT et Geoffrey I. WEBB. Boston, MA : Springer US, 2010, p. 81-89. ISBN : 978-0-387-30164-8. DOI : [10.1007/978-0-387-30164-8_66](https://doi.org/10.1007/978-0-387-30164-8_66). (Visité le 24/05/2023) (cf. p. 128, 134).

- [Ott+17] Nina OTTER, Mason A PORTER, Ulrike TILLMANN et al. « A Roadmap for the Computation of Persistent Homology ». In : *EPJ Data Science* 6.1 (déc. 2017), p. 17. ISSN : 2193-1127. DOI : [10.1140/epjds/s13688-017-0109-5](https://doi.org/10.1140/epjds/s13688-017-0109-5). (Visité le 04/10/2022) (cf. p. 104).
- [PSY20] Kamalesh PALANISAMY, Dipika SINGHANIA et Angela YAO. « Rethinking CNN Models for Audio Classification ». In : *arXiv :2007.11154 [cs, eess]* (nov. 2020). arXiv : [2007.11154](https://arxiv.org/abs/2007.11154) [cs, eess]. (Visité le 14/01/2021) (cf. p. 24, 65).
- [PY10] Sinno Jialin PAN et Qiang YANG. « A Survey on Transfer Learning ». In : *IEEE Transactions on Knowledge and Data Engineering* 22.10 (oct. 2010), p. 1345-1359. ISSN : 1041-4347. DOI : [10.1109/TKDE.2009.191](https://doi.org/10.1109/TKDE.2009.191). (Visité le 01/10/2020) (cf. p. 48).
- [PKL18] Yagya Raj PANDEYA, Dongwhoon KIM et Joonwhoan LEE. « Domestic Cat Sound Classification Using Learned Features from Deep Neural Nets ». In : *Applied Sciences* 8.10 (oct. 2018), p. 1949. ISSN : 2076-3417. DOI : [10.3390/app8101949](https://doi.org/10.3390/app8101949). (Visité le 04/08/2020) (cf. p. 24).
- [PR08] Omiros PAPASPILIOPOULOS et Gareth O. ROBERTS. « Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Models ». In : *Biometrika* 95.1 (2008), p. 169-186. ISSN : 0006-3444. JSTOR : [20441450](https://www.jstor.org/stable/20441450). (Visité le 11/09/2023) (cf. p. 34, 139).
- [PD07] Sylvain PARIS et Fredo DURAND. « A Topological Approach to Hierarchical Segmentation Using Mean Shift ». In : *2007 IEEE Conference on Computer Vision and Pattern Recognition*. Juin 2007, p. 1-8. DOI : [10.1109/CVPR.2007.383228](https://doi.org/10.1109/CVPR.2007.383228) (cf. p. 30).
- [PW23] Sanjana PATIL et Kiran WANI. « Gear Fault Detection Using Noise Analysis and Machine Learning Algorithm with YAMNet Pretrained Network ». In : *Materials Today : Proceedings*. 2nd International Conference and Exposition on Advances in Mechanical Engineering (ICoAME 2022) 72 (jan. 2023), p. 1322-1327. ISSN : 2214-7853. DOI : [10.1016/j.matpr.2022.09.307](https://doi.org/10.1016/j.matpr.2022.09.307). (Visité le 13/07/2023) (cf. p. 65).
- [Pat+19] Vic PATRANGENARU, Peter BUBENIK, Robert L. PAIGE et al. « Challenges in Topological Object Data Analysis ». In : *Sankhya A* 81.1 (fév. 2019), p. 244-271. ISSN : 0976-836X, 0976-8378. DOI : [10.1007/s13171-018-0137-7](https://doi.org/10.1007/s13171-018-0137-7). (Visité le 06/04/2022) (cf. p. 30, 105, 115, 125).
- [Pau+11] Rhea PAUL, Yael FUERST, Gordon RAMSAY et al. « Out of the Mouths of Babes : Vocal Production in Infant Siblings of Children with ASD : Vocalizations in Infant Siblings ». In : *Journal of Child Psychology and Psychiatry* 52.5 (mai 2011), p. 588-598. ISSN : 00219630. DOI : [10.1111/j.1469-7610.2010.02332.x](https://doi.org/10.1111/j.1469-7610.2010.02332.x). (Visité le 16/12/2019) (cf. p. 21, 146).

- [Ped22] Matteo PEDONE. « Covariate-Dependent Bayesian Models for Heterogeneous Populations ». Thèse de doct. Università di Firenze, 2022. (Visité le 17/09/2023) (cf. p. 35, 166).
- [PAS23] Matteo PEDONE, Raffaele ARGIENTO et Francesco C. STINGO. *Personalized Treatment Selection via Product Partition Models with Covariates*. Sept. 2023. DOI : [10.48550/arXiv.2210.06030](https://doi.org/10.48550/arXiv.2210.06030). arXiv : [2210.06030](https://arxiv.org/abs/2210.06030) [stat]. (Visité le 17/09/2023) (cf. p. 35, 166).
- [Pd15] Cássio M.M. PEREIRA et Rodrigo F. DE MELLO. « Persistent Homology for Time Series and Spatial Data Clustering ». In : *Expert Systems with Applications* 42.15-16 (sept. 2015), p. 6026-6038. ISSN : 09574174. DOI : [10.1016/j.eswa.2015.04.010](https://doi.org/10.1016/j.eswa.2015.04.010). (Visité le 29/11/2022) (cf. p. 30, 105, 114, 152).
- [Pic15] Karol J. PICZAK. « Environmental Sound Classification with Convolutional Neural Networks ». In : *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. Boston, MA, USA : IEEE, sept. 2015, p. 1-6. ISBN : 978-1-4673-7454-5. DOI : [10.1109/MLSP.2015.7324337](https://doi.org/10.1109/MLSP.2015.7324337). (Visité le 23/03/2021) (cf. p. 25).
- [Pij+11] Bryan C. PIJANOWSKI, Almo FARINA, Stuart H. GAGE et al. « What Is Soundscape Ecology? An Introduction and Overview of an Emerging New Science ». In : *Landscape Ecology* 26.9 (nov. 2011), p. 1213-1232. ISSN : 0921-2973, 1572-9761. DOI : [10.1007/s10980-011-9600-8](https://doi.org/10.1007/s10980-011-9600-8). (Visité le 15/07/2022) (cf. p. 65, 66).
- [PY97] Jim PITMAN et Marc YOR. « The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator ». In : *The Annals of Probability* 25.2 (avr. 1997), p. 855-900. ISSN : 0091-1798, 2168-894X. DOI : [10.1214/aop/1024404422](https://doi.org/10.1214/aop/1024404422). (Visité le 01/06/2023) (cf. p. 34).
- [Pop+21] Phillip POPE, Chen ZHU, Ahmed ABDELKADER et al. *The Intrinsic Dimension of Images and Its Impact on Learning*. Avr. 2021. DOI : [10.48550/arXiv.2104.08894](https://doi.org/10.48550/arXiv.2104.08894). arXiv : [2104.08894](https://arxiv.org/abs/2104.08894) [cs, stat]. (Visité le 29/08/2023) (cf. p. 29).
- [Pra19] Yosef PRAT. « Animals Have No Language, and Humans Are Animals Too ». In : *Perspectives on Psychological Science* 14.5 (sept. 2019), p. 885-893. ISSN : 1745-6916, 1745-6924. DOI : [10.1177/1745691619858402](https://doi.org/10.1177/1745691619858402). (Visité le 29/03/2021) (cf. p. 24, 75).
- [Qui+16] Fernando A. QUINTANA, Wesley O. JOHNSON, L. Elaine WAETJEN et al. « Bayesian Nonparametric Longitudinal Data Analysis ». In : *Journal of the American Statistical Association* 111.515 (juill. 2016), p. 1168-1181. ISSN : 0162-1459, 1537-274X. DOI : [10.1080/01621459.2015.1076725](https://doi.org/10.1080/01621459.2015.1076725). (Visité le 22/11/2022) (cf. p. 35, 166).

- [Qui+22] Fernando A. QUINTANA, Peter MÜLLER, Alejandro JARA et al. « The Dependent Dirichlet Process and Related Models ». In : *Statistical Science* 37.1 (fév. 2022), p. 24-41. ISSN : 0883-4237, 2168-8745. DOI : [10.1214/20-STS819](https://doi.org/10.1214/20-STS819). (Visité le 27/06/2023) (cf. p. 35, 161, 166).
- [RG00] Carl RASMUSSEN et Zoubin GHAMRANI. « Occam' s Razor ». In : *Advances in Neural Information Processing Systems*. T. 13. MIT Press, 2000. (Visité le 21/06/2023) (cf. p. 34).
- [RW06] Carl Edward RASMUSSEN et Christopher K. I. WILLIAMS. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, Mass : MIT Press, 2006. ISBN : 978-0-262-18253-9 (cf. p. 61).
- [Ren+16] Shaoqing REN, Kaiming HE, Ross GIRSHICK et al. « Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks ». In : *arXiv :1506.01497 [cs]* (jan. 2016). arXiv : [1506.01497 \[cs\]](https://arxiv.org/abs/1506.01497). (Visité le 09/03/2021) (cf. p. 24).
- [Rib+19] Teresa RIBAS-PRATS, Laura ALMEIDA, Jordi COSTA-FAIDELLA et al. « The Frequency-Following Response (FFR) to Speech Stimuli : A Normative Dataset in Healthy Newborns ». In : *Hearing Research* 371 (jan. 2019), p. 28-39. ISSN : 1878-5891. DOI : [10.1016/j.heares.2018.11.001](https://doi.org/10.1016/j.heares.2018.11.001) (cf. p. 18).
- [Rom+18] Rachel R. ROMEO, Julia A. LEONARD, Sydney T. ROBINSON et al. « Beyond the 30-Million-Word Gap : Children's Conversational Exposure Is Associated With Language-Related Brain Function ». In : *Psychological Science* 29.5 (mai 2018), p. 700-710. ISSN : 0956-7976. DOI : [10.1177/0956797617742725](https://doi.org/10.1177/0956797617742725). (Visité le 30/09/2023) (cf. p. 166).
- [Roy+15] Brandon C. ROY, Michael C. FRANK, Philip DECAMP et al. « Predicting the Birth of a Spoken Word ». In : *Proceedings of the National Academy of Sciences* 112.41 (oct. 2015), p. 12663-12668. ISSN : 0027-8424, 1091-6490. DOI : [10.1073/pnas.1419773112](https://doi.org/10.1073/pnas.1419773112). (Visité le 22/03/2022) (cf. p. 22).
- [RHW86] David E. RUMELHART, Geoffrey E. HINTON et Ronald J. WILLIAMS. « Learning Representations by Back-Propagating Errors ». In : *Nature* 323.6088 (oct. 1986), p. 533-536. ISSN : 0028-0836, 1476-4687. DOI : [10.1038/323533a0](https://doi.org/10.1038/323533a0). (Visité le 04/09/2023) (cf. p. 53).
- [SAN96] J. R. SAFFRAN, R. N. ASLIN et E. L. NEWPORT. « Statistical Learning by 8-Month-Old Infants ». In : *Science* 274.5294 (déc. 1996), p. 1926-1928. ISSN : 0036-8075, 1095-9203. DOI : [10.1126/science.274.5294.1926](https://doi.org/10.1126/science.274.5294.1926). (Visité le 04/06/2019) (cf. p. 18).
- [Saf+99] Jenny R SAFFRAN, Elizabeth K JOHNSON, Richard N ASLIN et al. « Statistical Learning of Tone Sequences by Human Infants and Adults ». In : *Cognition* 70.1 (fév. 1999), p. 27-52. ISSN : 0010-0277. DOI : [10.1016/S0010-0277\(98\)00075-4](https://doi.org/10.1016/S0010-0277(98)00075-4). (Visité le 30/09/2023) (cf. p. 18).

- [Saf03] Jenny R. SAFFRAN. « Statistical Language Learning : Mechanisms and Constraints ». In : *Current Directions in Psychological Science* 12.4 (août 2003), p. 110-114. ISSN : 0963-7214. DOI : [10.1111/1467-8721.01243](https://doi.org/10.1111/1467-8721.01243). (Visité le 03/10/2023) (cf. p. 19).
- [SB17] Justin SALAMON et Juan Pablo BELLO. « Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification ». In : *IEEE Signal Processing Letters* 24 (mars 2017), p. 279-283. DOI : [10.1109/LSP.2017.2657381](https://doi.org/10.1109/LSP.2017.2657381). arXiv : [1608.04363](https://arxiv.org/abs/1608.04363). (Visité le 05/08/2020) (cf. p. 56).
- [Sal+21] Andrew SALCH, Adam REGALSKI, Hassan ABDALLAH et al. « From Mathematics to Medicine : A Practical Primer on Topological Data Analysis (TDA) and the Development of Related Analytic Tools for the Functional Discovery of Latent Structure in fMRI Data ». In : *PLOS ONE* 16.8 (août 2021). Sous la dir. de Federico GIOVE, e0255859. ISSN : 1932-6203. DOI : [10.1371/journal.pone.0255859](https://doi.org/10.1371/journal.pone.0255859). (Visité le 16/12/2021) (cf. p. 30, 101, 105, 147).
- [San+17] Nicole SANDERSON, Elliott SHUGERMAN, Samantha MOLNAR et al. « Computational Topology Techniques for Characterizing Time-Series Data ». In : t. 10584. 2017, p. 284-296. DOI : [10.1007/978-3-319-68765-0_24](https://doi.org/10.1007/978-3-319-68765-0_24). arXiv : [1708.09359](https://arxiv.org/abs/1708.09359) [cs]. (Visité le 29/11/2022) (cf. p. 31).
- [SLR18] Alexander SCHINDLER, Thomas LIDY et Andreas RAUBER. « Multi-Temporal Resolution Convolutional Neural Networks for Acoustic Scene Classification ». In : *arXiv :1811.04419 [cs, eess]* (nov. 2018). arXiv : [1811.04419](https://arxiv.org/abs/1811.04419) [cs, eess]. (Visité le 09/03/2021) (cf. p. 24).
- [SZ00] Robert SERFLING et Yijun ZUO. « General Notions of Statistical Depth Function ». In : *The Annals of Statistics* 28.2 (avr. 2000), p. 461-482. ISSN : 0090-5364, 2168-8966. DOI : [10.1214/aos/1016218226](https://doi.org/10.1214/aos/1016218226). (Visité le 27/09/2023) (cf. p. 154).
- [SDB16] Lee M. SEVERSKY, Shelby DAVIS et Matthew BERGER. « On Time-Series Topological Data Analysis : New Data and Opportunities ». In : *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Las Vegas, NV, USA : IEEE, juin 2016, p. 1014-1022. ISBN : 978-1-5090-1437-8. DOI : [10.1109/CVPRW.2016.131](https://doi.org/10.1109/CVPRW.2016.131). (Visité le 07/11/2022) (cf. p. 30, 105).
- [Sha+23] Mohammad Moez SHAHRAMNIA, Akram AHMADI, Arezoo SAFFARIYAN et al. « Speech Sound Production, Speech Intelligibility, and Oral-Motor Outcomes of Preterm Children : Are They Different from Full Term Children? » In : *Applied Neuropsychology : Child* 12.1 (jan. 2023), p. 17-25. ISSN : 2162-2965. DOI : [10.1080/21622965.2021.2017940](https://doi.org/10.1080/21622965.2021.2017940). (Visité le 30/08/2023) (cf. p. 166).

- [Sha+16] Bobak SHAHRIARI, Kevin SWERSKY, Ziyu WANG et al. « Taking the Human Out of the Loop : A Review of Bayesian Optimization ». In : *Proceedings of the IEEE* 104.1 (jan. 2016), p. 148-175. ISSN : 1558-2256. DOI : [10.1109/JPROC.2015.2494218](https://doi.org/10.1109/JPROC.2015.2494218) (cf. p. 60, 70).
- [Shi+17] Yuta SHINYA, Masahiko KAWAI, Fusako NIWA et al. « Fundamental Frequency Variation of Neonatal Spontaneous Crying Predicts Language Acquisition in Preterm and Term Infants ». In : *Frontiers in Psychology* 8 (2017). ISSN : 1664-1078. (Visité le 22/05/2023) (cf. p. 21, 166).
- [STT21] Maciej SKORSKI, Alessandro TEMPERONI et Martin THEOBALD. « Revisiting Weight Initialization of Deep Neural Networks ». In : *Proceedings of The 13th Asian Conference on Machine Learning*. PMLR, nov. 2021, p. 1192-1207. (Visité le 02/09/2023) (cf. p. 42).
- [SLA12] Jasper SNOEK, Hugo LAROCHELLE et Ryan P ADAMS. « Practical Bayesian Optimization of Machine Learning Algorithms ». In : *Advances in Neural Information Processing Systems*. T. 25. Curran Associates, Inc., 2012. (Visité le 11/08/2023) (cf. p. 60, 70).
- [Sri+10] Niranjan SRINIVAS, Andreas KRAUSE, Sham KAKADE et al. « Gaussian Process Optimization in the Bandit Setting : No Regret and Experimental Design ». In : *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Madison, WI, USA : Omnipress, juin 2010, p. 1015-1022. ISBN : 978-1-60558-907-7. (Visité le 05/09/2023) (cf. p. 61).
- [Sri+14] Nitish SRIVASTAVA, Geoffrey HINTON, Alex KRIZHEVSKY et al. « Dropout : A Simple Way to Prevent Neural Networks from Overfitting ». In : *Journal of Machine Learning Research* 15 (2014), p. 1929-1958 (cf. p. 57, 58, 69).
- [Sto22] Dan STOWELL. « Computational Bioacoustics with Deep Learning : A Review and Roadmap ». In : *PeerJ* 10 (mars 2022), e13152. ISSN : 2167-8359. DOI : [10.7717/peerj.13152](https://doi.org/10.7717/peerj.13152). (Visité le 28/04/2022) (cf. p. 26, 65, 66, 162).
- [Sto+18] Dan STOWELL, Yannis STYLIANOU, Mike WOOD et al. « Automatic Acoustic Detection of Birds through Deep Learning : The First Bird Audio Detection Challenge ». In : *arXiv :1807.05812 [cs, eess]* (juill. 2018). arXiv : [1807.05812 \[cs, eess\]](https://arxiv.org/abs/1807.05812). (Visité le 26/04/2021) (cf. p. 24, 65).
- [SVP19] Nicola STRISCIUGLIO, Mario VENTO et Nicolai PETKOV. « Learning Representations of Sound Using Trainable COPE Feature Extractors ». In : *Pattern Recognition* 92 (août 2019), p. 25-36. ISSN : 00313203. DOI : [10.1016/j.patcog.2019.03.016](https://doi.org/10.1016/j.patcog.2019.03.016). (Visité le 29/07/2022) (cf. p. 24, 65).
- [Sud06] Erik B. (Erik Blaine) SUDDERTH. « Graphical Models for Visual Object Recognition and Tracking ». Thesis. Massachusetts Institute of Technology, 2006. (Visité le 30/05/2023) (cf. p. 128, 132).

- [Sue18] Jerome SUEUR. *Sound Analysis and Synthesis with R*. 1st edition. New York, NY : Springer Berlin Heidelberg, 2018. ISBN : 978-3-319-77645-3 (cf. p. 27, 112, 152).
- [Sze+15] Christian SZEGEDY, WEI LIU, YANGQING JIA et al. « Going Deeper with Convolutions ». In : *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA : IEEE, juin 2015, p. 1-9. ISBN : 978-1-4673-6964-0. DOI : [10 . 1109 / CVPR . 2015 . 7298594](https://doi.org/10.1109/CVPR.2015.7298594). (Visité le 20/07/2022) (cf. p. 47).
- [Tak81] Floris TAKENS. « Detecting Strange Attractors in Turbulence ». In : *Dynamical Systems and Turbulence, Warwick 1980*. Sous la dir. de David RAND et Lai-Sang YOUNG. T. 898. Berlin, Heidelberg : Springer Berlin Heidelberg, 1981, p. 366-381. ISBN : 978-3-540-11171-9 978-3-540-38945-3. DOI : [10 . 1007 / BFb0091924](https://doi.org/10.1007/BFb0091924). (Visité le 07/05/2023) (cf. p. 107, 151).
- [Tan+18] Chuanqi TAN, Fuchun SUN, Tao KONG et al. « A Survey on Deep Transfer Learning ». In : *Artificial Neural Networks and Machine Learning – ICANN 2018*. Sous la dir. de Věra KŮRKOVÁ, Yannis MANOLOPOULOS, Barbara HAMMER et al. T. 11141. Cham : Springer International Publishing, 2018, p. 270-279. ISBN : 978-3-030-01423-0 978-3-030-01424-7. DOI : [10 . 1007 / 978-3-030-01424-7_27](https://doi.org/10.1007/978-3-030-01424-7_27). (Visité le 12/05/2022) (cf. p. 48).
- [Teh+06] Yee Whye TEH, Michael I JORDAN, Matthew J BEAL et al. « Hierarchical Dirichlet Processes ». In : *Journal of the American Statistical Association* 101.476 (déc. 2006), p. 1566-1581. ISSN : 0162-1459. DOI : [10 . 1198 / 016214506000000302](https://doi.org/10.1198/016214506000000302). (Visité le 17/06/2023) (cf. p. 35, 165).
- [TJ10] Yee Whye TEH et Michael I. JORDAN. « Hierarchical Bayesian Nonparametric Models with Applications ». In : *Bayesian Nonparametrics*. Sous la dir. de Chris HOLMES, Nils Lid HJORT, Peter MÜLLER et al. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge : Cambridge University Press, 2010, p. 158-207. ISBN : 978-0-521-51346-3. DOI : [10 . 1017 / CB09780511802478 . 006](https://doi.org/10.1017/CB09780511802478.006). (Visité le 25/05/2023) (cf. p. 165).
- [TCS22] Alberto TENA, Francesc CLARIÀ et Francesc SOLSONA. « Automated Detection of COVID-19 Cough ». In : *Biomedical Signal Processing and Control* 71 (jan. 2022), p. 103175. ISSN : 1746-8094. DOI : [10 . 1016 / j . bspc . 2021 . 103175](https://doi.org/10.1016/j.bspc.2021.103175). (Visité le 13/07/2023) (cf. p. 65).
- [TdL00] J. B. TENENBAUM, V. DE SILVA et J. C. LANGFORD. « A Global Geometric Framework for Nonlinear Dimensionality Reduction ». In : *Science (New York, N.Y.)* 290.5500 (déc. 2000), p. 2319-2323. ISSN : 0036-8075. DOI : [10 . 1126 / science . 290 . 5500 . 2319](https://doi.org/10.1126/science.290.5500.2319) (cf. p. 28).
- [] *TensorFlow Hub*. <https://tfhub.dev/google/yamnet/1>. (Visité le 26/04/2022) (cf. p. 65, 67).

- [ter+21] Sita M. TER HAAR, Ahana A. FERNANDEZ, Maya GRATIER et al. « Cross-Species Parallels in Babbling : Animals and Algorithms ». In : *Philosophical Transactions of the Royal Society B : Biological Sciences* 376.1836 (oct. 2021), p. 20200239. ISSN : 0962-8436, 1471-2970. DOI : [10.1098/rstb.2020.0239](https://doi.org/10.1098/rstb.2020.0239). (Visité le 31/05/2022) (cf. p. 20, 146, 160).
- [TUW91] Esther THELEN, Beverly D. ULRICH et Peter H. WOLFF. « Hidden Skills : A Dynamic Systems Analysis of Treadmill Stepping during the First Year ». In : *Monographs of the Society for Research in Child Development* 56.1 (1991), p. i-103. ISSN : 0037-976X. DOI : [10.2307/1166099](https://doi.org/10.2307/1166099). JSTOR : [1166099](https://www.jstor.org/stable/1166099). (Visité le 01/10/2023) (cf. p. 21).
- [Tin20] Raphaël TINARRAGE. « Topological Inference from Measures and Vector Bundles ». Thèse de doct. Université Paris-Saclay, oct. 2020. (Visité le 10/03/2023) (cf. p. 28, 29).
- [TJ12] Ruth TINCOFF et Peter W. JUSCZYK. « Six-Month-Olds Comprehend Words That Refer to Parts of the Body ». In : *Infancy* 17.4 (2012), p. 432-444. ISSN : 1532-7078. DOI : [10.1111/j.1532-7078.2011.00084.x](https://doi.org/10.1111/j.1532-7078.2011.00084.x). (Visité le 30/09/2023) (cf. p. 19).
- [TH17] Yuji TOKOZUME et Tatsuya HARADA. « Learning Environmental Sounds with End-to-End Convolutional Neural Network ». In : *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA : IEEE, mars 2017, p. 2721-2725. ISBN : 978-1-5090-4117-6. DOI : [10.1109/ICASSP.2017.7952651](https://doi.org/10.1109/ICASSP.2017.7952651). (Visité le 09/03/2021) (cf. p. 25).
- [Tra+15] Du TRAN, Lubomir BOURDEV, Rob FERGUS et al. « Learning Spatiotemporal Features with 3D Convolutional Networks ». In : *arXiv :1412.0767 [cs]* (oct. 2015). arXiv : [1412.0767 \[cs\]](https://arxiv.org/abs/1412.0767). (Visité le 15/01/2021) (cf. p. 24).
- [Tro+23] Ilya TROFIMOV, Daniil CHERNIAVSKII, Eduard TULCHINSKII et al. « Learning Topology-Preserving Data Representations ». In : *The Eleventh International Conference on Learning Representations*. Fév. 2023. (Visité le 25/05/2023) (cf. p. 29, 101, 151, 161, 165).
- [VR21] M. VASHKEVICH et Yu. RUSHKEVICH. « Classification of ALS Patients Based on Acoustic Analysis of Sustained Vowel Phonations ». In : *Biomedical Signal Processing and Control* 65 (mars 2021), p. 102350. ISSN : 1746-8094. DOI : [10.1016/j.bspc.2020.102350](https://doi.org/10.1016/j.bspc.2020.102350). (Visité le 28/09/2023) (cf. p. 111).
- [VZT12] Josef VAVRINA, Petr ZETOCHA et Jana TUCKOVA. « Detection of Degree of Developmental Dysphasia Based on Methods of Vowel Analysis ». In : *2012 35th International Conference on Telecommunications and Signal Processing (TSP)*. Juill. 2012, p. 503-507. DOI : [10.1109/TSP.2012.6256345](https://doi.org/10.1109/TSP.2012.6256345). (Visité le 28/09/2023) (cf. p. 111).

- [WG18] Sara WADE et Zoubin GHARAMANI. « Bayesian Cluster Analysis : Point Estimation and Credible Balls (with Discussion) ». In : *Bayesian Analysis* 13.2 (juin 2018), p. 559-626. ISSN : 1936-0975, 1931-6690. DOI : [10.1214/17-BA1073](https://doi.org/10.1214/17-BA1073). (Visité le 06/06/2023) (cf. p. [143](#), [153](#)).
- [Was18] Larry WASSERMAN. « Topological Data Analysis ». In : *Annual Review of Statistics and Its Application* 5.1 (2018), p. 501-532. DOI : [10.1146/annurev-statistics-031017-100045](https://doi.org/10.1146/annurev-statistics-031017-100045) (cf. p. [104](#), [124](#), [152](#)).
- [Wer88] WERBOS. « Backpropagation : Past and Future ». In : *IEEE 1988 International Conference on Neural Networks*. Juill. 1988, 343-353 vol.1. DOI : [10.1109/ICNN.1988.23866](https://doi.org/10.1109/ICNN.1988.23866) (cf. p. [53](#)).
- [WRS21] Kathleen WERMKE, Michael P. ROBB et Philip J. SCHLUTER. « Melody Complexity of Infants' Cry and Non-Cry Vocalisations Increases across the First Six Months ». In : *Scientific Reports* 11.1 (déc. 2021), p. 4137. ISSN : 2045-2322. DOI : [10.1038/s41598-021-83564-8](https://doi.org/10.1038/s41598-021-83564-8). (Visité le 19/12/2021) (cf. p. [21](#), [146](#)).
- [Wil+17] Conor J. WILD, Annika C. LINKE, Leire ZUBIAURRE-ELORZA et al. « Adult-like Processing of Naturalistic Sounds in Auditory Cortex by 3- and 9-Month Old Infants ». In : *NeuroImage* 157 (août 2017), p. 623-634. ISSN : 10538119. DOI : [10.1016/j.neuroimage.2017.06.038](https://doi.org/10.1016/j.neuroimage.2017.06.038). (Visité le 22/01/2023) (cf. p. [18](#)).
- [XW14] Kelin XIA et Guo-Wei WEI. « Persistent Homology Analysis of Protein Structure, Flexibility, and Folding ». In : *International Journal for Numerical Methods in Biomedical Engineering* 30.8 (août 2014), p. 814-844. ISSN : 2040-7947. DOI : [10.1002/cnm.2655](https://doi.org/10.1002/cnm.2655) (cf. p. [30](#)).
- [Xia+18] Xianjun XIA, Roberto TOGNERI, Ferdous SOHEL et al. « Random Forest Classification Based Acoustic Event Detection Utilizing Contextual-Information and Bottleneck Features ». In : *Pattern Recognition* 81 (sept. 2018), p. 1-13. ISSN : 00313203. DOI : [10.1016/j.patcog.2018.03.025](https://doi.org/10.1016/j.patcog.2018.03.025). (Visité le 24/07/2022) (cf. p. [24](#), [64](#)).
- [XDR21] Xiaoqi XU, Nicolas DROUGARD et Raphaëlle N. ROY. « Topological Data Analysis as a New Tool for EEG Processing ». In : *Frontiers in Neuroscience* 15 (nov. 2021), p. 761703. ISSN : 1662-453X. DOI : [10.3389/fnins.2021.761703](https://doi.org/10.3389/fnins.2021.761703). (Visité le 02/12/2022) (cf. p. [30](#), [105](#), [123](#), [124](#), [147](#)).
- [YC21] Peter Tsung-Wen YEN et Siew Ann CHEONG. « Using Topological Data Analysis (TDA) and Persistent Homology to Analyze the Stock Markets in Singapore and Taiwan ». In : *Frontiers in Physics* 9 (2021). ISSN : 2296-424X. (Visité le 31/08/2023) (cf. p. [30](#)).
- [Zei12] Matthew D. ZEILER. *ADADELTA : An Adaptive Learning Rate Method*. Déc. 2012. arXiv : [1212.5701 \[cs\]](https://arxiv.org/abs/1212.5701). (Visité le 04/09/2023) (cf. p. [54](#)).


- [Zha+17a] Chiyuan ZHANG, Samy BENGIO, Moritz HARDT et al. « Understanding Deep Learning Requires Rethinking Generalization ». In : *arXiv:1611.03530 [cs]* (fév. 2017). arXiv : [1611.03530 \[cs\]](#). (Visité le 03/03/2021) (cf. p. 56).
- [ZWP22] Michael Minyi ZHANG, Sinead A. WILLIAMSON et Fernando PEREZ-CRUZ. « Accelerated Parallel Non-conjugate Sampling for Bayesian Non-parametric Models ». In : *Statistics and Computing* 32.3 (juin 2022), p. 50. ISSN : 0960-3174, 1573-1375. DOI : [10.1007/s11222-022-10108-z](#). arXiv : [1705.07178 \[stat\]](#). (Visité le 11/09/2023) (cf. p. 35).
- [Zha+16] Weibin ZHANG, Wenkang LEI, Xiangmin XU et al. « Improved Music Genre Classification with Convolutional Neural Networks ». In : *Interspeech 2016*. Sept. 2016, p. 3304-3308. DOI : [10.21437/Interspeech.2016-1236](#). (Visité le 09/03/2021) (cf. p. 24).
- [Zha+17b] Ying ZHANG, Mohammad PEZESHKI, Philemon BRAKEL et al. « Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks ». In : *arXiv:1701.02720 [cs, stat]* (jan. 2017). arXiv : [1701.02720 \[cs, stat\]](#). (Visité le 12/03/2021) (cf. p. 24).
- [Zho+18] Wenda ZHOU, Victor VEITCH, Morgane AUSTERN et al. « Non-Vacuous Generalization Bounds at the ImageNet Scale : A PAC-Bayesian Compression Approach ». In : *International Conference on Learning Representations*. Sept. 2018. (Visité le 18/09/2023) (cf. p. 56).
- [ZEH16] Zhenyao ZHU, Jesse H. ENGEL et Awni HANNUN. « Learning Multiscale Features Directly From Waveforms ». In : *arXiv:1603.09509 [cs]* (avr. 2016). arXiv : [1603.09509 \[cs\]](#). (Visité le 09/03/2021) (cf. p. 24).
- [ZC05] Afra ZOMORODIAN et Gunnar CARLSSON. « Computing Persistent Homology ». In : *Discrete & Computational Geometry* 33.2 (fév. 2005), p. 249-274. ISSN : 1432-0444. DOI : [10.1007/s00454-004-1146-y](#). (Visité le 01/08/2023) (cf. p. 108).
- [Zom05] Afra J. ZOMORODIAN. *Topology for Computing*. 1^{re} éd. Cambridge University Press, jan. 2005. ISBN : 978-0-521-83666-1 978-0-521-13609-9 978-0-511-54694-5. DOI : [10.1017/CB09780511546945](#). (Visité le 15/08/2023) (cf. p. 83, 104, 151).
- [Zub+18] Leire ZUBIAURRE-ELORZA, Annika C. LINKE, Charlotte HERZMANN et al. « Auditory Structural Connectivity in Preterm and Healthy Term Infants during the First Postnatal Year ». In : *Developmental Psychobiology* 60.3 (2018), p. 256-264. ISSN : 1098-2302. DOI : [10.1002/dev.21610](#). (Visité le 22/01/2023) (cf. p. 18).

ANNEXES

A. Documents à destination des parents

Page 1 - Ajouter le titre de la page

VOLONTAIRE POUR AIDER LA RECHERCHE ?



BABYVOC

Nous voulons comprendre les phases précoces du développement du langage chez le bébé.

La méthodologie consiste en des enregistrements audio réalisés à domicile. Aucun danger sanitaire. Données recueillies anonymes.



PARTICIPEZ À LA RECHERCHE !

SOYEZ ACTEUR D'UNE ÉTUDE POUR COMPRENDRE LE DÉVELOPPEMENT DU LANGAGE CHEZ L'ENFANT.

Sans risque pour l'enfant et en étant le moins contraignant possible pour la famille, nous cherchons à décrire la trajectoire développementale des vocalisations émises par l'enfant lors de ses **12 premiers mois**.

Un micro est laissé à la famille **trois jours par mois**, puis récupéré par le chercheur. Les parents n'ont pas de déplacements à faire et sont maîtres des moments enregistrés.

Il est possible de se retirer de la recherche à tout moment, sans justification. Les enregistrements sont anonymisés.

Pour avoir plus d'informations, rendez-vous sur le site de l'étude en scannant le QR-code, ou directement sur <https://lpc.univ-amu.fr/fr/babyvoc>.

Si vous avez des questions : lpc-babyvoc@univ-amu.fr



Étude réalisée dans le cadre d'une thèse en partenariat avec l'Institut de Mathématiques de Marseille, le Laboratoire de Psychologie Cognitive et Résurgences R&D, sous la convention CIFRE n°2019/1534. Validé par le comité d'éthique de l'université d'Aix-Marseille.



Notice d'information

Titre de l'étude:

ETUDE DESPRODUCTIONS VOCALES SPONTANÉES DU BÉBÉ HUMAIN

Madame, Monsieur,

L'investigateur principal, le Dr. Arnaud Rey, chercheur au CNRS, vous a proposé de participer au protocole de recherche intitulé :

«Étude des productions vocales spontanées du bébé humain ».

Nous vous proposons de lire attentivement cette notice d'information qui a pour but de répondre aux questions que vous seriez susceptible de vous poser avant de prendre votre décision de participation.

Vous pourrez à tout moment vous adresser au Dr. Rey pour lui poser toutes les questions complémentaires que vous souhaitez.

Objectif de la recherche

Cette recherche vise à mieux comprendre l'évolution et les caractéristiques des productions vocales chez le bébé humain de 0 à 12 mois.

Quelle est la méthodologie et comment se déroule l'expérimentation ?

La méthodologie employée consiste en des enregistrements audios réalisés à domicile. Un micro placé sur un pied et raccordé à un système d'enregistrement numérique est placé à proximité pour enregistrer le bébé et ses productions vocales. Ce système est mis en route par le parent au moment où il laisse son enfant dans son lieu de repos et il est arrêté quand le parent vient le rechercher. Ces enregistrements sont réalisés une fois par mois pendant les 12 premiers mois de l'enfant, à raison de 3 jours et deux nuits d'enregistrement consécutifs par mois. Ce système d'enregistrement est le même que celui qui est vendu dans le commerce comme moyen d'enregistrement du son. Il ne présente donc aucun danger sanitaire.

Quelles sont les contraintes et désagréments ?

Ce système est sans contrainte et sans désagrément pour l'enfant. Pour les parents, il leur est simplement demandé de bien penser à allumer le système d'enregistrement lorsqu'ils quittent la chambre de l'enfant et à l'éteindre lorsqu'ils viennent le rechercher.

Quels sont vos droits en tant que participant(e) à cette recherche ?

Vous pouvez refuser de participer à cette recherche sans avoir à vous justifier. De même vous pouvez vous retirer à tout moment, sans justification et sans conséquence pour vous et votre enfant.

Que deviendront les données enregistrées ?

A l'issue de chaque enregistrement audio, les données seront stockées sur un serveur sécurisé et dédié à cette recherche. Chaque fichier sera identifié par un code qui ne portera pas le nom de l'enfant afin d'en garantir l'anonymat. Enfin, aucune propagation de ces données ne sera possible et tout le travail d'analyse des données se fera sur le serveur dédié sans possibilité d'exporter ces données. Seuls les chercheurs impliqués dans cette étude auront accès aux enregistrements.

Cette recherche relève de l'application du Code de la Santé Publique (Titre II du Livre Premier relatif aux recherches biomédicales). Ces informations sont consultables sur le site Internet de Legifrance (www.legifrance.gouv.fr)

**L'investigateur principal de cette étude est le Dr Arnaud Rey.
Cette étude est réalisée par le Laboratoire de Psychologie Cognitive (UMR 7290 CNRS et Aix-Marseille Université).**

1-Conformément aux dispositions de loi relative à l'informatique et aux libertés (loi n°78-17 du 6 janvier 1978 modifiée par la loi n°2004-801 du 6 août 2004) vous disposez d'un droit d'accès, de rectification et d'opposition relatif au traitement de vos données personnelles. Ces droits s'exercent auprès du Dr. Arnaud Rey.

Formulaire du recueil de consentement (en 2 exemplaires)

Titre de l'étude:

**ETUDE DES PRODUCTIONS VOCALES
DU BÉBÉ HUMAIN**

Le Dr Arnaud Rey, chercheur CNRS au laboratoire de Psychologie Cognitive à Marseille (UMR 7290) et investigateur principal, m'a proposé de participer à la recherche intitulée :

«Étude des productions vocales du bébé humain »

dont le responsable de traitement est le Laboratoire de Psychologie Cognitive.

J'ai pris connaissance de la note d'information m'expliquant le protocole de recherche mentionné ci-dessus. J'ai pu poser toutes les questions que je voulais et j'ai reçu des réponses adaptées.

J'ai noté que les données recueillies lors de cette recherche demeureront strictement confidentielles et anonymes. Elles ne sortiront pas de l'Union Européenne. Elles seront conservées jusqu'à deux ans après la dernière publication. Le destinataire des données est le doctorant en charge du projet de thèse et ses directeurs de thèse. Le Délégué à la Protection des Données est Emilie MASSON (dpd.demandes@cnrs.fr).

J'accepte le traitement informatisé des données nominatives qui me concernent en conformité avec l'article 6.e du RGPD ; car le traitement de données est nécessaire à l'exécution d'une mission d'intérêt public (recherche scientifique), ainsi qu'avec les dispositions de la loi n°2004-801 du 6 août 2004 relative à la protection des personnes et modifiant la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. J'ai noté que je pourrai exercer mon droit d'accès et de rectification garanti par les articles 39 et 40 de cette loi en m'adressant auprès du Dr Arnaud Rey. Je peux introduire une réclamation auprès de la CNIL.

J'ai compris que je pouvais refuser de participer à cette étude sans conséquence pour moi, et que je pourrai retirer mon consentement à tout moment (avant et en cours d'étude) sans avoir à me justifier et sans conséquence. Je peux pour cela envoyer un mail ou téléphoner au doctorant Guillem Bonafos (06.98.03.49.99. ou guillem.bonafos@univ-amu.fr).

Compte tenu des informations qui m'ont été transmises, j'accepte librement et volontairement de participer à la recherche intitulée : « Étude des productions vocales du bébé humain ».

Mon consentement ne décharge pas l'investigateur et le promoteur de leurs responsabilités à mon égard.

Bibliographie – A. Documents à destination des parents

Fait à le
En deux exemplaires originaux

Participant à la recherche
principal

Investigateur

Noms Prénoms :

Nom Prénom :

Signature :
(Précédée de la mention : *Lu, compris et approuvé*)

Signature :