Caractérisation du répertoire vocal des chimpanzés par apprentissage profond

Nicolas Audebert, Marion Laporte

Caractérisation du répertoire vocal des chimpanzés par apprentissage profond

Nicolas Audebert[1] Marion Laporte[2]

[1] CEDRIC, Conservatoire national des arts et métiers, 2 rue Conté 75003 Paris, France
[2] Muséum national d'histoire naturelle, 17 place du Trocadéro et du 11 novembre 75116 Paris, France
nicolas.audebert@cnam.fr, marion1laporte@gmail.com

1 Introduction

Dans les questionnements sur l'émergence du langage, les primates sont généralement considérés comme de bon modèles évolutifs. Récemment, des travaux ont par exemple mis en évidence la capacité de certains primates à produire des vocalisations semblables à des voyelles [2, 4, 6]. Cependant, beaucoup de primates vivent dans des environnements à faible visibilité, ce qui les amènent à privilégier des signaux vocaux à une fréquence fondamentale élevée, plus adaptée à la propagation longue portée [10]. C'est notamment le cas pour les chimpanzés, habitués aux environnements de jungle ou de savanne [5], chez qui ces vocalisations représentent deux-tiers des communications vocales des mâles [1]. Ces vocalisations à longue distance ne peuvent toutefois pas véhiculer d'information liée aux formants, contrairement à la prononciation humaine des voyelles. À l'inverse, ce sont les modulations de fréquence fondamentale qui seront préservées par la distance tout en étant robuste au bruit ambiant. Pourtant, aucune étude ne s'est encore penchée sur l'importance de cette fondamentale dans l'interprétation des vocalisations des chimpanzés. Dans nos travaux, nous nous proposons de confronter la catégorisation des vocalisations de chimpanzés communément admise chez les primatologues à la fouille statistique de données. En particulier, cette taxonomie divise les vocalisations en groupes séparés et bien distincts. Toutefois, les observations sur le terrain laissent à penser que le répertoire vocal des chimpanzés présente un caractère gradé, c'est-à-dire continu, passant aisément d'une modalité à une autre. Nous proposons donc dans cet article d'analyser un jeu de données de vocalisations de chimpanzés acquises *in situ*. Notre objectif est la reconnaissance du *type* de la vocalisation, c'est-à-dire la classification supervisée : pouvons-nous construire un modèle statistique capable de prédire le type de vocalisation à partir du signal ? Si oui, à partir de quelles informations fréquentielles ? Et quelles conclusions sur le répertoire vocal des chimpanzés pouvons-nous en tirer ? En particulier, nous souhaitons comparer deux types de modèles : les modèles s'intéressant au spectre du signal, comme ceux existant [11], que nous confronterons à la classification à partir de la fréquence fondamentale seule, afin d'isoler l'importance de la modulation de la f_0 dans la communication entre chimpanzés.

2 Classification automatique des vocalisations

2.1 Jeu de données

Nous avons collecté *in situ* plus de 750 heures d'enregistrements de chimpanzés dans leur environnement naturel en Ouganda. Le corpus final est constitué de plus de 6000 vocalisations de 30 individus (12 mâles et 18 femelles, dont respectivement 11 et 12 adultes). Le profil fréquentiel correspondant à chaque vocalisation a été extrait manuellement à l'aide du logiciel Winpitch [8], tel qu'illustré dans la Figure 1. Les vocalisations sont classées dans une des six catégories d'intérêt : grognement (GR, *grunt*), aboiement (BK, *bark*), aboiement doux (SB, *soft bark*), inhalation (IN) et hululement (HT, *hoot*). [1] Le nombre total de classes du jeu de données est porté à huit, considérant que les hululements se séparent en trois sous-types : hauts (HH, *high hoot*), moyens (HT) et bas (LH, *low hoot*). La durée de chaque vocalisation varie entre environ 20 ms pour les plus courtes et 1,5 s pour les plus longues.

2.2 Classifieurs

Notre objectif est la classification *supervisée* des vocalisations : un modèle statistique peut-il reproduire la catégorisation experte des primatologues ? Nous considérons trois approches : temporelle, fréquentielle et spectro-temporelle. Nous conservons $2/3$ du jeu de données pour l'apprentissage et $1/3$ pour l'évaluation.

Classification par profil fréquentiel L'évolution de la fréquence fondamentale f_0 est modélisée comme une série temporelle univariée, échantillonnée avec un pas de 10 ms. Nous considérons un classifieur kNN utilisant la déformation temporelle dynamique (DTW) comme métrique, ainsi qu'une SVM basée sur le noyau GAK [3] (dérivé de la DTW).

1. Environ 2000 vocalisations additionnelles, provenant de catégories annexes ou non identifiées, pourraient être exploitées à l'avenir.

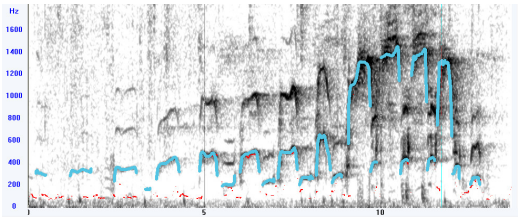

FIGURE 1 – Spectrogramme annoté avec le profil fréquentiel (fondamentale f_0, en bleu clair).

Entrée	Méthode	BK	SB	GR	IN	SC	LH	HT	HH	Acc.
P. fréq.	KNN + DTW	0.76	0.52	0.83	0.70	0.82	0.19	0.29	0.32	69.2
P. fréq.	SVM + GAK	0.91	0.73	0.85	0.81	0.64	0.73	0.66	0.00	79.6
MFCC	SVM + RBF	0.81	0.56	0.74	0.79	0.83	0.51	0.51	0.63	72.8
Spectre	CNN 2D	0.89	0.67	0.84	0.84	0.86	0.72	0.71	0.61	81.5

TABLE 1 – Performances des classifieurs de vocalisation selon le type d'entrée (scores F_1 par classe et *accuracy* globale).

Classification par MFCC Nous calculons pour chaque vocalisation les coefficients MFCC, dérivés du spectre de Mel et classiquement utilisés en traitement de la parole humaine. En considérant que les vocalisations des grands singes peuvent présenter des caractéristiques proches, les MFCC représentent un choix standard pour la classification. Nous calculons 40 coefficients pour chaque fenêtre temporelle d'environ 20 ms. Le vecteur de caractéristiques final comporte la moyenne et l'écart-type de chaque coefficient sur toute la durée de la vocalisation. Ce vecteur de dimension 80 est passé à une SVM RBF.

Classification du spectrogramme Les CNN 2D typiquement utilisés pour la classification d'images obtiennent des performances largement satisfaisantes pour la classification de spectrogramme, en dépit de la nature différente des signaux [7]. Les images de spectrogrammes sont obtenues par passage dans le domaine de Fourier, puis en décibel, sur une fenêtre de 20 ms (320 points à une fréquence d'échantillonnage $f_s = 16\,\mathrm{kHz}$) et un pas de 3 ms avec fenêtrage de Hann. Les images subissent un *padding* afin de maintenir leurs dimensions constantes à 150×320. Nous utilisons en première approche EfficientNet-B3 [9] pré-entraîné sur ImageNet puis optimisé par descente de gradient adaptative (Adam) avec une taille de batch de 10 et pas d'apprentissage $1{\times}10^{-4}$ pour 36 000 itérations.

2.3 Discussion

Le Tableau 1 détaille les performances des classifieurs selon la représentation des vocalisations choisie en entrée. Les performances de classification sur le profil fréquentiel seul sont raisonnablement élevées : plus de 75% des vocalisations sont identifiables à partir de l'évolution de la f_0, confirmant que cette modulation de la fondamentale est un facteur fortement différenciant dans le répertoire du chimpanzé. Les performances sont comparables sur les coefficients MFCC. Bien que ceux-ci ne représentent plus l'information temporelle, ils synthétisent suffisamment bien l'information spectrale pour permettre d'identifier presque 80% des vocalisations. Comme attendu, les performances les plus élevées sont obtenues pour le CNN 2D appliqué au spectrogramme : il s'agit de l'entrée comportant l'intégralité de l'information. Cela confirme d'une part que seule l'étude spectro-temporelle permet d'entièrement caractériser une vocalisation et, d'autre part, le potentiel des réseaux profonds convolutifs issus de la reconnaissance d'images pour l'analyse des sons.
La difficulté à classer les *soft barks* renforce l'hypothèse d'une gradation du répertoire vocal des chimpanzés. Ces vocalisations sont en effet confondues par tous les modèles avec des *barks* ou des *grunts*. Ces deux groupes représentent vraisemblablement les extrémités d'un cluster plus grand, reliées par des *soft barks* aux caractéristiques sonores plus ou moins prononcées. Les hululements présentent le même schéma : plus de 50% des erreurs sur le tryptique LH/HT/HH sont des confusions avec un autre type de hululement, et ce quel que soit le modèle. Le CNN identifie mieux la fréquence de la vocalisation, son intensité et ses harmoniques grâce à au spectrogramme, ce qui fiabilise le classement par rapport aux MFCC ou au profil fréquentiel. En revanche, les confusions résiduelles indiquent que les hululements ne sont peut-être pas séparables sans ambiguité.

3 Conclusion

Notre étude confirme la place de la modulation de la fréquence fondamentale comme principal facteur caractéristique d'une vocalisation chez les chimmpanzés, permettant d'identifier 80% des vocalisations. Sans surprise, la catégorisation des vocalisations est plus précise lorsque tout le spectrogramme est pris en compte. En particulier, l'information spectrale seule représentée par les moments d'ordre 1 des coefficients MFCC ne suffit pas à mieux reconnaître une vocalisation par rapport à la f_0. En revanche, les CNN 2D sont bien adaptés à la reconnaissance des vocalisations et permettent notamment de lever des ambiguités de classification sur des sous-catégories fines (*soft barks* et variantes du *hoot*) que le profil fréquentiel seul ne parvient pas à expliquer. Le CNN 2D est donc un candidat naturel pour l'extraction de représentations sur des vocalisations.
À terme, nous souhaitons prolonger cette étude par une analyse non-supervisée. En particulier, nous avons jusqu'ici écarté certaines vocalisations difficiles à inscrire dans la catégorisation décrite précédemment. Un *clustering* combiné à une visualisation en dimension réduite, par exemple via t-SNE, permettrait de représenter le répertoire vocal des chimpanzés et de mieux comprendre les liens entre les différentes vocalisations, ainsi que la réalité statistique de la catégorisation experte construite par les primatologues. D'autres extracteurs de caractéristiques sont envisageables. Le CNN 1D sur le profil fréquentiel, voire sur la forme d'onde brute, pourrait ainsi être une alternative avantageuse au CNN image utilisé ici.

Remerciements Ces travaux appartienent au projet *Apesvoice*, financé par Sorbonne Université (Émergence 2021-2022). Nous remercions l'équipe *Origins of Speech* de l'ISCD et tout particulièrement Amélie Viallet, Louis-Jean Boë et Pascal Perrier pour leurs conseils.

Références

[1] A. C. ARCADI. « Vocal responsiveness in male wild chimpanzees : implications for the evolution of language ». In : *Journal of Human Evolution* 39.2 (août 2000), p. 205-223.

[2] Louis-Jean BOË et al. « Evidence of a Vocalic Proto-System in the Baboon (Papio papio) Suggests Pre-Hominin Speech Precursors ». In : *PLOS ONE* 12.1 (11 jan. 2017).

[3] Marco CUTURI. « Fast global alignment kernels ». In : *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. 28 juin 2011, p. 929-936.

[4] W. Tecumseh FITCH et al. « Monkey vocal tracts are speech-ready ». In : *Science Advances* 2.12 (sept. 2016).

[5] Jane GOODALL. *The chimpanzees of Gombe : patterns of behavior*. Belknap Press of Harvard University Press, 1986.

[6] Sven GRAWUNDER et al. « Chimpanzee vowel-like sounds and voice quality suggest formant space expansion through the hominoid lineage ». In : *Philosophical Transactions of the Royal Society B : Biological Sciences* (jan. 2022).

[7] Shawn HERSHEY 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.

[8] Philippe MARTIN. « WinPitch LTL, un logiciel multimédia d'enseignement de la prosodie ». In : *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication* (Vol. 8, n° 2 2 déc. 2005), p. 95-108.

[9] Mingxing TAN et Quoc LE. « EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks ». In : *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 24 mai 2019, p. 6105-6114.

[10] Ingo R. TITZE et Anil PALAPARTHI. « Radiation efficiency for long-range vocal communication in mammals and birds ». In : *The Journal of the Acoustical Society of America* 143.5 (mai 2018), p. 2813-2824.

[11] Philip WADEWITZ et al. « Characterizing Vocal Repertoires—Hard vs. Soft Classification Approaches ». In : *PLOS ONE* 10.4 (27 avr. 2015).