Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Statistiques des sons naturels et hypothèse du codage efficace pour la perception de la musique et de la parole : Mise en place d'une méthodologie d'évaluation.

Agnieszka Duniec[1] Olivier Crouzet[1,2] Elisabeth Delais-Roussarie[1]
(1) Laboratoire de Linguistique de Nantes, LLING – UMR6310, Université de Nantes / CNRS
chemin de la Censive du Tertre, 44312 Nantes Cedex, France
(2) ENT Department - University Medical Center Groningen, Rijksuniversiteit Groningen, Pays-Bas
agnieszka.duniec@etu.univ-nantes.fr, olivier.crouzet@univ-nantes.fr,
elisabeth.delais-roussarie@univ-nantes.fr

RÉSUMÉ

L'hypothèse du *codage efficace* prédit que les systèmes perceptifs sont optimalement adaptés aux propriétés statistiques des signaux naturels. Ce caractère optimal a été récemment évalué sur la base d'analyses statistiques réalisées sur des décompositions spectrales de signaux de parole représentés comme des modulations d'énergie. Ces travaux pourraient trouver des applications directes dans l'amélioration du codage des signaux acoustiques par des implants cochléaires. Cependant, les recherches sur la perception de la musique par des personnes sourdes portant un implant cochléaire mettent en avant des limites qui semblent discordantes avec les performances observées concernant certaines propriétés fondamentales de la parole. Nous comparons les résultats d'analyses statistiques de signaux musicaux avec ceux qui ont été réalisés sur de la parole dans le but d'évaluer les impacts respectifs de ces deux gammes de signaux sonores pour évaluer leurs contributions à cette proposition théorique. Des résultats préliminaires et les perspectives futures sont discutés.

ABSTRACT

Natural sound statistics and the efficient coding hypothesis for music and speech perception : setting-up an evaluation methodology.

The *efficient coding* hypothesis predicts that perceptual systems are optimally adapted to natural signal statistics. Such optimal characterization has recently been evaluated on the basis of statistical analyses that were performed on spectral decompositions of speech signals. Speech signals were decomposed into energy envelopes and these results may find applications in the improvement of acoustic signal coding for cochlear implants. However, research on music perception in cochlear implanted deaf listeners sheds light on potential limits associated with music perception that seem to be in contradiction with how some of the fundamental properties of speech sounds are processed. Our aim is to compare the statistical analysis of natural music signals with previous work on speech in order to evaluate their respective contributions to this theoretical proposal. Preliminary results along with future perspectives are discussed.

MOTS-CLÉS : perception, implants cochléaires, statistiques des signaux naturels, hypothèse du codage efficace.

KEYWORDS: perception, cochlear implants, natural signal statistics, efficient coding hypothesis.

Volume 1 : Journées d'Études sur la Parole, pages 181–189.
HAL : hal-02798534.

1 Introduction

La perception de l'environnement sonore est un mécanisme particulièrement complexe. Les travaux princeps de Bregman (1994) ont mis en évidence que les mécanismes impliqués dans l'analyse des scènes auditives reposent en partie sur des processus cognitifs généraux. Si l'on considère en général que des mécanismes fondamentaux d'analyse auditive doivent être impliqués dans les traitements liés aux différents systèmes de communication (langage oral et musique par exemple), il semble par contre raisonnable de considérer que ces deux catégories de signaux restent au moins en partie fondamentalement différentes du point de vue de leur organisation sonore et des informations acoustiques qui les composent (rythmique, mélodique, harmonique).

En outre, on observe des performances très différentes entre la perception de la parole et celle de la musique par des auditeurs sourds portant un implant cochléaire. Si une grande partie des informations associées aux sons de parole peut donner lieu à des performances de reconnaissance satisfaisantes dans des environnements non-dégradés (Bouton *et al.*, 2012), la capacité de ces auditeurs à apprécier ou à identifier un certain nombre d'aspects liés à la musique (mélodie, accords par exemple) reste limitée (Galvin *et al.*, 2009). Ainsi, parole et musique semblent porter des informations qui ne sont pas propices à donner lieu à des traitements perceptifs similaires lorsqu'on les étudie du point de vue de personnes sourdes portant un implant cochléaire.

1.1 L'hypothèse du codage efficace

Proposée initialement par Smith & Lewicki (2006) pour le système auditif sur la base de travaux ayant plutôt porté leur attention sur le système visuel (Simoncelli & Olshausen, 2001), la théorie du « codage efficace » (*efficient coding hypothesis*) postule que les systèmes perceptifs sont optimalement adaptés aux propriétés des signaux naturels de manière à transmettre un maximum d'information en recourant à une consommation minimale de ressources. Cette hypothèse a des implications fortes concernant (1) la nature des représentations sensorielles impliquées dans la perception de notre environnement et (2) les mécanismes perceptifs qui sont mis en œuvre : On s'attend à trouver des correspondances entre propriétés physiques des signaux et caractéristiques des systèmes sensoriels.

Même si cette hypothèse peut paraître aller de soi, elle débouche sur une prédiction qui n'est pas si évidente, laquelle consiste à envisager une réduction maximale de la « granularité » des représentations perceptives permettant aux observateurs d'être parfaitement efficaces sans mettre en oeuvre de représentations « trop » fines qui demanderaient plus de complexité de traitement que nécessaire. Ainsi, les systèmes sensoriels pourraient procéder à une analyse peu précise si elle est « optimale ». L'une des prédictions qui découlent de l'hypothèse du codage efficace est que les signaux naturels développés par les espèces animales auraient évolué de manière à être compatibles avec cette caractéristique et ne requerraient donc pas de contenir plus d'information que nécessaire. De ce point de vue, aussi bien les stimuli naturels que les systèmes sensoriels seraient *optimalement* économes en information, tant que cette économie garantit une analyse perceptive efficace. C'est de ces prédictions que l'*hypothèse du codage efficace* tire son nom.

Si l'on considère cette hypothèse à l'aune des travaux réalisés ces 30 dernières années sur la perception de la parole par des personnes sourdes portant un implant cochléaire aussi bien que par des personnes normo-entendantes soumises à des simulations d'implants cochléaires (parole vocodée à canaux), la plupart des résultats observés semblent aller dans ce sens. Ainsi, certaines informations linguistiques

sont accessibles avec une résolution spectrale très limitée (4 bandes de fréquence suffisent à percevoir avec une performance élevée le voisement ou le mode d'articulation d'une consonne, Shannon *et al.*, 1995). Par contre, d'autres informations acoustiques semblent beaucoup plus problématiques : informations tonales associées à la prosodie de la phrase ou aux tons phonémiques (Milczynski *et al.*, 2012; Gaudrain *et al.*, 2008; Everhardt *et al.*, 2020), genre du locuteur (Fuller *et al.*, 2014), nasalité vocalique (Borel, 2015) par exemple. De même, la perception de la musique, du genre du locuteur, ou de la parole en environnement bruité semblent résister à des conditions plus fines de résolution spectrale (Galvin *et al.*, 2009; Fuller *et al.*, 2014). Les travaux actuels sur la parole vocodée montrent de manière générale qu'un accroissement du nombre de canaux spectraux n'est pas suffisant pour améliorer significativement les performances de perception de ces informations et une grande partie des enjeux théoriques actuels est sous-tendue par cette limite.

1.2 Analyse statistique de signaux naturels

Les signaux naturels présentent des propriétés statistiques qui ont conduit certains auteurs à explorer plus précisément le rôle que jouent ces régularités dans leur reconnaissance. Par exemple, les « textures sonores » (bruit du vent, vol d'insectes…) sont associées à des propriétés statistiques spécifiques qui sont corrélées entre les bandes de fréquence (McDermott & Simoncelli, 2011). La synthèse de sons qui respectent ces régularités corrélées conduit à une reconnaissance de ces textures, ce qui semble indiquer que certains types de « statistiques des signaux naturels » jouent un rôle crucial dans les mécanismes d'identification perceptive.

Certains travaux ont également décrit l'existence de similarités entre des propriétés statistiques observées dans des langues orales et dans de la musique. Ces similitudes porteraient notamment sur la structure énergétique du spectre des harmoniques (Schwartz *et al.*, 2003) ainsi que sur la taille des intervalles tonaux (Han *et al.*, 2011). Ces observations tendent à argumenter en faveur de l'existence de propriétés structurelles parallèles dans la parole et la musique.

Les travaux qui reposent sur l'*hypothèse du codage efficace* s'inspirent de principes relativement proches tirés des travaux sur les *statistiques des signaux naturels* : les signaux de communication seraient caractérisés par des propriétés sonores statistiques qui seraient régulières malgré la diversité apparente des réalisations acoustiques. Ces travaux se sont notamment centrés sur le caractère optimal (1) du nombre de canaux spectraux pour représenter des langues orales mais aussi (2) de la localisation des frontières entre ces canaux.

Ming & Holt (2009) ont montré que, sans changer le nombre de canaux spectraux (6 en l'occurrence) les changements de localisation des frontières spectrales en parole vocodée ont des effets massifs sur les taux de reconnaissance de mots et de segments phonétiques. Ueda & Nakajima (2017), capitalisant sur ces résultats, ont développé une méthode d'analyse inspirée des travaux de Plomp *et al.* (1967) sur les voyelles : ils étendent cette approche à l'étude d'un corpus de phrases. Ils procèdent, sur la base de signaux acoustiques de parole codés sur environ 100 canaux de représentation spectrale répartis en « bandes critiques » étroites, à diverses Analyses en Composantes Principales (ACP, en anglais *PCA*) portant sur les enveloppes d'énergie de ces canaux et varient le nombre de facteurs associés à la sortie de l'ACP (2, 3, 4, 5, 6). Leur travail aboutit à la conclusion que 4 facteurs suffiraient à représenter optimalement des signaux de parole, et ce pour chacune des 8 langues de leur échantillon. Ils constatent par ailleurs que les 3 frontières fréquentielles découlant de chacune des ACP à 4 facteurs réalisées sur ces 8 langues sont parfaitement appariées (env. 540, 1720, 3300 Hz), ce qui les amène à conclure que les langues seraient de manière générale fondées sur des indices qui seraient

parfaitement adaptés à un traitement perceptif « parcimonieux » de la parole.

Récemment, Grange & Culling (2018) ont répliqué l'étude de Ueda & Nakajima (2017) en la mettant en rapport avec des données de perception de parole vocodée (simulations d'implants cochléaires) et ont abouti à des conclusions assez similaires. Leurs résultats suggèrent néanmoins que, pour rendre compte de manière appropriée des propriétés acoustiques de la parole vocodée, il faudrait 6 à 7 canaux spectraux pour représenter optimalement ces signaux. Cette limite correspond dans leurs données, à un point d'inflexion au-delà duquel la performance de reconnaissance mesurée chez les auditeurs ne s'améliore plus.

Si Ming & Holt (2009) se positionnent en faveur d'un traitement efficace relevant de représentations équivalentes quels que soient les signaux envisagés (parole, musique, sons de l'environnement), les données de la littérature concernant les patients sourds qui utilisent un implant cochléaire pourraient amener à nuancer cette position. Ainsi, les performances observées aussi bien chez des auditeurs normo-entendants écoutant des signaux vocodés que chez des patients sourds portant un implant cochléaire sont systématiquement meilleures pour de la parole que pour de la musique (Galvin *et al.*, 2009; Crew *et al.*, 2015), notamment si l'on compare les performances mesurées dans le silence en environnement non-réverbérant. Du point de vue de l'hypothèse du codage efficace, on pourrait être amené à envisager que parole et musique requièrent des niveaux de résolution spectrale très différents pour que leur analyse perceptive soit appropriée. Si tel était le cas, une telle constatation aurait un impact crucial sur les fondements ou la compréhension de cette hypothèse du codage efficace.

L'objet de notre travail est d'évaluer cette contradiction potentielle en mettant en place une série d'analyses qui chercheront dans un premier temps à évaluer les *propriétés statistiques de signaux naturels* de musique et à les comparer à des répliques des analyses réalisées par Ueda & Nakajima (2017) sur de la parole naturelle.

2 Analyse des propriétés statistiques de signaux de musique

2.1 Méthode

L'ensemble des analyses acoustiques et statistiques est réalisé dans l'environnement Matlab. Les scripts d'analyse sont disponibles sur un dépôt github (https://github.com/crouzet-o/naturalSignalStats).

2.1.1 Base de données d'enregistrements musicaux

Dans un souci de réplicabilité des analyses et des résultats qui en découleront, nous avons choisi d'utiliser des extraits musicaux issus d'une base de données en *open source* : *FMA* (*Free Music Archive*, Defferrard *et al.*, 2017). FMA offre la possibilité d'accéder légalement à une bibliothèque d'enregistrements de musique sous licence libre. Elle est constituée de 4 versions, lesquelles contiennent de 8000 à 106574 morceaux musicaux qui sont disponibles soit en extraits de 30 s (les 3 premières versions) soit dans leur intégralité (la 4ème version).

Tous les morceaux musicaux sont au format MP3 et sont associés à des informations qualitatives (*tags*) : numéro d'identification du morceau, titre, artiste, genre (et sous-genres), ainsi qu'à des traits

musicaux (*features*) automatiquement déterminés par la bibliothèque librosa (McFee *et al.*, 2015, traits spectraux, rythmiques…).

Les données présentées ici concernent la base de données la moins volumineuse. Elle est composée de 8000 extraits de 30 secondes de 8 genres musicaux différents en format MP3 (taux de compression entre 128 et 256 kbits/s, fréq. d'échantillonnage 44.1 kHz). Les 8 genres musicaux sont en proportions équilibrées dans cette version de la base.

2.1.2 Paramètrage acoustique des signaux

Préalablement à l'analyse statistique des signaux, nous procédons à une paramétrisation acoustique équivalente à celles qui ont été utilisées dans les travaux précédents (Ueda & Nakajima, 2017; Grange & Culling, 2018). On notera que les travaux antérieurs ayant porté sur de la parole, ils se sont restreints à des fréquences supérieures d'environ 8000 Hz. En ce qui nous concerne, nous manipulons ce paramètre afin de comparer les résultats obtenus en fonction de la limite supérieure de fréquence, laquelle pourrait comporter des informations acoustiques essentielles pour les signaux de musique.

Nous avons analysé une durée totale de signal audio équivalente à celle qui a été étudiée pour les langues les plus fournies de l'échantillon étudié par Ueda & Nakajima (2017, env. 4000 s). Pour celà, nous extrayons pour chaque enregistrement musical disponible les 10 premières secondes. Au total 471 stimuli musicaux ont été exploités, parmi lesquels 71 n'étaient pas lisibles par l'algorithme de décompression MP3 utilisé. L'échantillon final est composé de 400 enregistrements audio fournissant une durée totale de 4000 s (soit environ 1h) d'audio.

Les enregistrements sélectionnés sont ensuite convertis en monophonique par combinaison des deux canaux stéréophoniques et concaténés les uns aux autres. Les enveloppes de modulation temporelle des signaux sont extraites à partir d'un banc de filtres dont la largeur croît de manière logarithmique avec la fréquence centrale (canaux de largeur $\frac{1}{4}$ d'ERB, ce qui correspond à 106 canaux spectraux allant jusqu'à la fréquence supérieure maximale de 8000 Hz et à 129 canaux pour une fréquence maximale de 22000 Hz). Ces enveloppes subissent une rectification demi-onde puis un filtrage passe-bas avec une fréquence de coupure de 50 Hz. Les signaux d'enveloppe résultants sont ensuite élevés au carré et convertis en notes centrées réduites (*z-scores*). Cette chaîne de paramétrage permet de procéder à une analyse des co-modulations entres les bandes de fréquence sur une base d'analyse des corrélations entre les informations d'enveloppe.

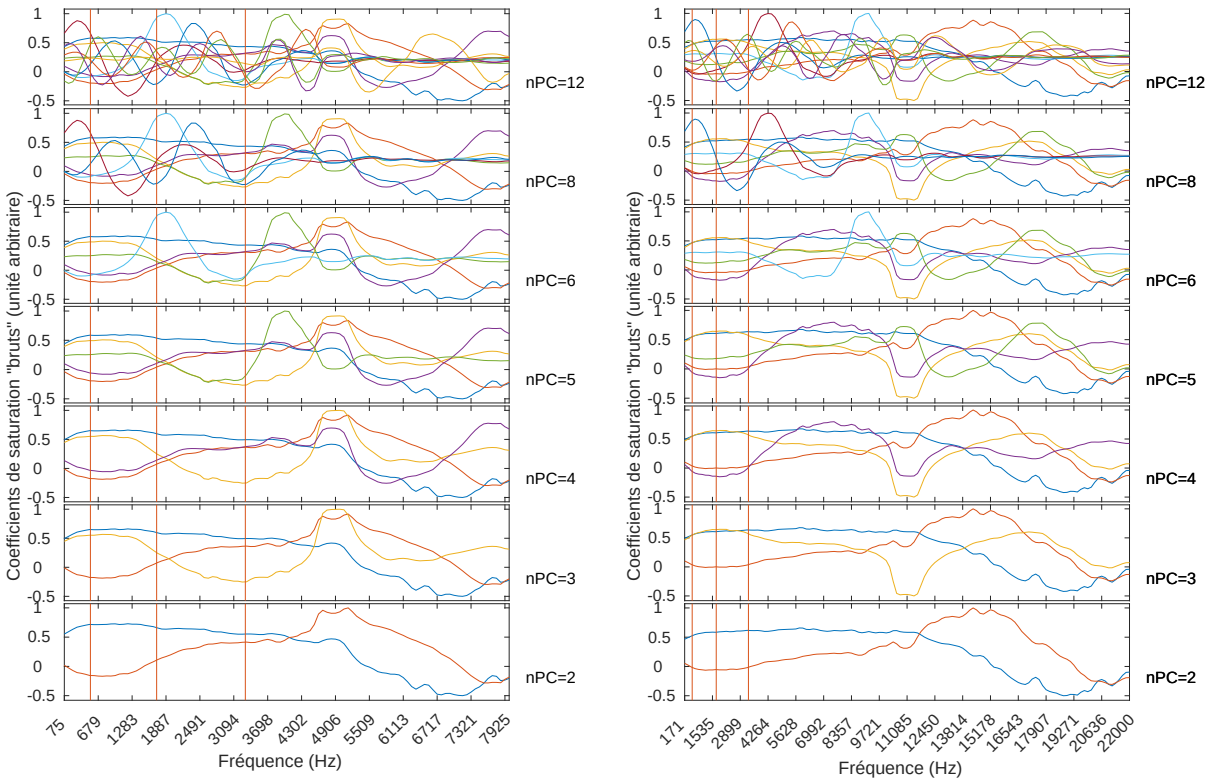
Le signal résultant, composé de 106 / 129 canaux en fonction de la fréquence maximale, correspond aux modulations temporelles de l'enveloppe de chaque canal au cours du temps. Cette matrice de modulations d'amplitude est alors transférée vers un outil statistique d'analyse en composantes principales (*Principal Components Analysis*).

2.1.3 Analyse en Composantes Principales

L'Analyse en Composantes Principales est une méthode descriptive d'analyse de données qui permet une étude simultanée de plus de 2 dimensions (analyse multivariée). L'objectif est de représenter l'essentiel de l'information contenue dans un tableau de données quantitatif en réduisant le nombre de facteurs explicatifs. Le principe est de transformer des variables liées (ayant des corrélations statistiques) en nouvelles variables synthétiques (composantes principales) en perdant le moins d'information possible. Cette analyse permet non seulement de réduire le nombre de variables à

mesurer, et ainsi améliorer la caractérisation des données, mais aussi d'identifier les facteurs non corrélés utiles pour procéder à une analyse discriminante. Concrètement, les variables initiales sont représentées dans un nouvel espace de facteurs définis par les vecteurs propres de la matrice de corrélations. L'hypothèse sous-jacente à l'application de cette méthode sur des signaux sonores est que certains canaux spectraux contiendraient des informations redondantes et qu'il serait alors économe de restreindre l'analyse perceptive à une séparation en zones de fréquences étant maximalement informatives (donc minimalement redondantes). En celà, l'ACP nous permettrait d'identifier les canaux de fréquence optimaux pour différencier de manière parcimonieuse les enregistrements d'un corpus.

L'analyse préliminaire des résultats porte essentiellement sur la description des graphiques indiquant les valeurs des coefficients de saturation pour chaque composante principale en fonction du canal de fréquence (Fig. 1), ce qui permet de décrire l'empan de fréquences qui corrèle avec une même composante.



(a) Fréq. Max. : 8000 Hz

(b) Fréq. Max. : 22000 Hz

FIGURE 1 – Coefficients de saturation (*factor loadings*) issus des ACP réalisées sur un échantillon des signaux de musique de la base de données *FMA* (*Free Music Archive*, Defferrard *et al.*, 2017), durée totale 4000 s, fréquence maximale supérieure de (a) 8000 Hz, (b) 22000 Hz

2.2 Résultats

Les travaux antérieurs ayant porté sur de la parole, ils se sont restreints à des fréquences supérieures autour de 8000 Hz. En ce qui nous concerne, nous manipulons ce paramètre afin de comparer les

résultats obtenus en fonction de la limite supérieure de fréquence, à savoir 22000 Hz car cette gamme supérieure pourrait comporter des informations acoustiques essentielles pour les signaux de musique.

On peut dans un premier temps observer que la prise en compte des fréquences supérieures à 8000 Hz se justifie totalement puisque la deuxième composante principale (en rouge sur la fig. 1) s'étend approximativement sur la gamme 12000 à 16000 Hz. La figure 1a représente la distribution des coefficients de saturation des composantes principales sur la gamme de fréquences s'étendant jusqu'à 8000 Hz en fonction du nombre de facteurs étudiés. Le nombre de facteurs augmente de bas en haut (de 2 à 12). La figure 1b représente l'analyse réalisée sur une plus large gamme de fréquences mais pour les mêmes signaux et le même nombre de composantes principales.

Dans l'analyse réalisée par Ueda & Nakajima (2017), les frontières entre canaux optimaux sont placées aux croisements des courbes de saturation, délimitant ainsi des zones de fréquences. Les frontières que nous avons utilisées dans nos analyses sont celles de Ueda & Nakajima (2017, les traits verticaux de couleur orange sur les graphiques de la fig. 1). Dans l'analyse réalisée par les auteurs, ces frontières délimitent quatre zones de fréquence principales : '*low*' centrée à 300 Hz, '*mid-low*' centrée autour de 1000 Hz, '*mid-high*' centrée à 2200 Hz et '*high*' centrée autour de 4500 Hz. Dans l'analyse que nous avons réalisée jusqu'à 22000 Hz, les 3 zones inférieures identifiées par les auteurs ('*low*', '*mid-low*' et '*mid-high*') semblent peu représentées. Même la zone supérieure ('*high*') autour de 4500 Hz n'apparaît qu'à partir de la 8ème CP même si la zone de hautes fréquences de notre analyse (à partir de 5000 Hz) est largement représentée par les 5 à 6 premières CP. Néanmoins, la 3ème CP peut-être décrite comme s'étendant sur les 3 zones inférieures identifiées par Ueda & Nakajima (2017). Ces observations reflètent également les données obtenues en se limitant à la gamme de fréquences inférieure à 8000 Hz. Dans nos analyses, un découpage plus fin associé aux 3 zones de basse fréquence identifiées par Ueda & Nakajima (2017) apparaît seulement à partir de la huitième CP, particulièrement si l'on s'intéresse à l'analyse jusqu'à 22000 Hz.

3 Discussion

À la lueur de ces premières explorations, ont peut identifier deux points centraux : (1) si les données musicales ne coincident pas avec les données de parole décrites dans la littérature, certains principes similaires semblent ressortir (répartition des coefficients de saturation sur des gammes de fréquence spécifiques ; correspondance possible mais encore à l'état de conjecture entre les zones de fréquence décrites en parole et en musique, cependant sur des composantes très différentes –composantes de plus grande contribution / de plus bas niveau pour la parole que pour la musique). (2) il semble probable (et pas surprenant) que les signaux de musique requièrent un plus grand nombre de composantes principales pour être caractérisés que les signaux de parole mais il restera à identifier dans quelle mesure les frontières fréquentielles pertinentes pourraient entrer en correspondance.

Les résultats présentés ici sont préliminaires et constituent seulement une ébauche des analyses à venir. Nous développons actuellement une amélioration des scripts d'analyse afin de mettre en place une rotation orthogonale des facteurs qui permettra de faciliter la description des résultats issus des ACP. En outre, à partir des principes d'extraction de données présentés ici, nous nous orienterons vers des analyses plus objectives (1) des composantes principales à prendre en compte à partir des mesures d'inertie (2) ainsi que des frontières de fréquence correspondantes.

Du point de vue des perspectives générales de ce travail, l'hypothèse du codage efficace prédit qu'un

nombre réduit de dimensions devrait permettre de caractériser les signaux de musique au même titre que les signaux de parole. Cette hypothèse pourra donc être évaluée de manière *interne* : en étudiant les résultats obtenus pour des signaux de musique indépendamment des résultats antérieurs sur la parole. Nous chercherons à établir dans quelle mesure l'analyse statistique de signaux musicaux permet de retrouver certains *patterns* observés pour la parole : nombre réduit de canaux optimaux par rapport au nombre total de canaux de codage en enveloppe d'énergie, correspondance des résultats quel que soit le genre musical, comparaison des résultats en fonction de certains paramètres qualitatifs (présence majoritaire d'informations à large bande comme des percussions vs. caractéristiques tonales).

Parallèlement à cette évaluation interne, certains résultats devront nécessairement être discutés dans une perspective comparative articulant les données observées sur la parole et sur la musique : il est tout à fait possible que les détails de ces analyses fassent ressortir (1) que les signaux de musique requièrent un plus grand nombre de canaux spectraux que la parole, et (2) que les frontières entre les canaux (à nombre de canaux équivalent ou pas) soient divergentes si l'on compare les analyses qui reposent sur la parole et celles qui reposent sur la musique.

Afin de pouvoir s'assurer que d'éventuelles divergences de résultats entre d'une part nos analyses réalisées sur de la musique et, d'autre part ceux de Ueda & Nakajima (2017) et Grange & Culling (2018), ne seraient pas le fait de différences fines qui interviendraient dans la mise en œuvre des algorithmes, nous comparerons nos résultats obtenus sur des signaux de musique avec une base de données de parole. Ceci permettra de s'assurer que nous répliquons les observations des travaux antérieurs (Ueda & Nakajima, 2017; Grange & Culling, 2018). Dans le souci de s'approcher au mieux des conditions des études précédentes, nous utiliserons une base de données de phrases collectées en laboratoire et contenant un nombre suffisament large de phrases différentes pour chaque locuteur.

Les résultats observés constitueront une source précieuse d'information, d'une part pour évaluer les fondements de l'hypothèse du codage efficace et ses impacts sur la modélisation perceptive des signaux naturels, d'autre part pour envisager dans quelle mesure cette hypothèse pourrait conduire à développer des solutions permettant d'améliorer le codage de signaux sonores par des dispositifs comme les implants cochléaires.

Remerciements

Ce travail a reçu le soutien du programme Recherche – Formation – Innovation « Ouest Industries Créatives » (RFI-OIC, Région Pays de la Loire) par une allocation doctorale attribuée à AD.

Références

BOREL S. (2015). *Perception auditive, visuelle et audiovisuelle des voyelles nasales par les adultes devenus sourds. Lecture labiale, implant cochléaire, implant du tronc cérébral.* Thèse de doctorat, Université de la Sorbonne Nouvelle – Paris 3.

BOUTON S., SERNICLAES W., BERTONCINI J. & COLÉ P. (2012). Perception of Speech Features by French-Speaking Children With Cochlear Implants. *Journal of Speech Language and Hearing Research*, **55**(1), 139–153. DOI : 10.1044/1092-4388(2011/10-0330).

BREGMAN A. S. (1994). *Audirory scene analysis : the perceptual organization of sound.* A Bradford book, Cambridge, Mass. : MIT Press., 2nd édition.

CREW J. D., GALVIN J. J. & FU Q.-J. (2015). Melodic contour identification and sentence recognition using sung speech. *The Journal of the Acoustical Society of America*, **3**(138).

DEFFERRARD M., BENZI K., VANDERGHEYNST P. & BRESSON X. (2017). Fma : Dataset for music analysis. *18th International Society for Music Information Retrieval Conference*.

EVERHARDT M. K., SARAMPALIS A., COLER M., BAŞKENT D. & LOWIE W. (2020). Meta-Analysis on the Identification of Linguistic and Emotional Prosody in Cochlear Implant Users and Vocoder Simulations :. *Ear and Hearing*, p. in press. DOI : 10.1097/AUD.0000000000000863.

FULLER C. D., GAUDRAIN E., CLARKE J. N., GALVIN J. J., FU Q.-J., FREE R. H. & BAŞKENT D. (2014). Gender categorization is abnormal in cochlear implant users. *Journal of the Association for Research in Otolaryngology,*, **6**(15).

GALVIN J. J., FU Q.-J. & SHANNON R. V. (2009). Melodic contour identification and music perception by cochlear implant users. *Annals of the New York Academy of Sciences*, **1**(1169), 518–533.

GAUDRAIN E., GRIMAULT N., HEALY E. W. & BÉRA J.-C. (2008). Streaming of vowel sequences based on fundamental frequency in a cochlear-implant simulation. *The Journal of the Acoustical Society of America*, **124**(5), 3076–87. DOI : 10.1121/1.2988289.

GRANGE J. & CULLING J. (2018). The factor analysis of speech : Limitations and opportunities for cochlear implants. *Acta Acustica united with Acustica*, **104**, 835–838.

HAN S. E., SUNDARARAJAN J., BOWLING D. L., LAKE J. & PURVES D. (2011). Co-Variation of Tonality in the Music and Speech of Different Cultures. *PLOS ONE*, **6**(5), e20160. DOI : 10.1371/journal.pone.0020160.

MCDERMOTT J. H. & SIMONCELLI E. P. (2011). Sound Texture Perception via Statistics of the Auditory Periphery : Evidence from Sound Synthesis. *Neuron*, **71**(5), 926–940. DOI : 10.1016/j.neuron.2011.06.032.

MCFEE B., RAFFEL C., LIANG D., ELLIS D., MCVICAR M., BATTENBERG E. & NIETO O. (2015). librosa : Audio and Music Signal Analysis in Python. In *The 14th Python in Science Conference (scipy 2015)*, p. 18–24, Austin, Texas. DOI : 10.25080/Majora-7b98e3ed-003.

MILCZYNSKI M., CHANG J. E., WOUTERS J. & VAN WIERINGEN A. (2012). Perception of Mandarin Chinese with cochlear implants using enhanced temporal pitch cues. *Hearing Research*, **285**(1–2), 1–12. DOI : 10.1016/j.heares.2012.02.006.

MING V. L. & HOLT L. L. (2009). Efficient coding in human auditory perception. *The Journal of the Acoustical Society of America*, **3**(126), 1312–1320.

PLOMP R., POLS L. C. W. & VAN DE GEER J. P. (1967). Dimensional analysis of vowel spectra. *The Journal of the Acoustical Society of America*, **3**(41), 707–712.

SCHWARTZ D. A., HOWE C. Q. & PURVES D. (2003). The Statistical Structure of Human Speech Sounds Predicts Musical Universals. *Journal of Neuroscience*, **23**(18), 7160–7168. DOI : 10.1523/JNEUROSCI.23-18-07160.2003.

SHANNON R., ZENG F., KAMATH V., WYGONSKI J. & EKELID M. (1995). Speech recognition with primarily temporal cues. *Science*, **270**, 303–304.

SIMONCELLI E. P. & OLSHAUSEN B. A. (2001). Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, **24**(1), 1193–1216. DOI : 10.1146/annurev.neuro.24.1.1193.

SMITH E. C. & LEWICKI M. S. (2006). Efficient auditory coding. *Nature*, **7079**, 978–982.

UEDA K. & NAKAJIMA Y. (2017). An acoustic key to eight languages/dialects : Factor analyses of critical-band-filtered speech. *Scientific Reports*, **7**, 42468. DOI : 10.1038/srep42468.