Analyse Factorielle de signaux musicaux : comparaison avec les données de parole dans la perspective de l'hypothèse de codage efficace.

Agnieszka Duniec, Olivier Crouzet, Elisabeth Delais-Roussarie

XXXIVe Journées d'Études sur la Parole – JEP 2022
13-17 June 2022, Île de Noirmoutier, France

Analyse Factorielle de signaux musicaux : comparaison avec les données de parole dans la perspective de l'hypothèse du codage efficace et de l'application aux implants cochléaires.

Agnieszka Duniec[1] Olivier Crouzet[1] 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
agnieszka.duniec@etu.univ-nantes.fr, olivier.crouzet@univ-nantes.fr

RÉSUMÉ

Certaines études supportent l'hypothèse d'une relation entre les propriétés statistiques des modulations spectro-temporelles de la parole et les mécanismes d'analyse perceptive. Ces travaux ont des conséquences pour la modélisation des mécanismes perceptifs mis en place à travers des implants cochléaires et montrent que les gammes de fréquence dont les modulations d'amplitude sont corrélées semblent correspondre à des frontières spectrales qui pourraient être envisagées comme optimales sur le plan perceptif. L'objet du présent travail est de présenter une première comparaison entre les résultats récents sur la parole et l'analyse comparable d'une base de données d'enregistrements musicaux afin d'évaluer la généralisabilité de cette hypothèse concernant des types variés de signaux acoustiques de communication. Selon nos analyses réalisées sur une base de données d'enregistrements de musique d'ensemble, une partie des observations est comparable. Nous envisageons les implications et limites de cette comparaison concernant la perception sonore pour le langage et la musique.

ABSTRACT

Factor Analysis of music signals : comparison with speech data from the *efficient coding hypothesis* viewpoint

Recent studies have led to support the hypothesis of a relationship between statistical properties of spectro-temporal modulations in speech and mechanisms of perceptual analysis. These results have strong implications for the modelling of perceptual mechanisms associated with cochlear implants and suggest that frequency channels exhibiting correlations in amplitude modulation would correspond to optimal spectral boundaries in terms of perceptual analysis. We provide a first comparison of previous results on speech data with a comparable database of ensemble music recordings in order to provide elements for the evaluation of the generalizability of these observations for various types of acoustic communication signals. According to our analysis of a music database, some aspects of the results are similar. We discuss the implications and limits of this comparison in terms of sound perception for language and music.

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

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

10.21437/JEP.2022-75

1 Introduction

L'*hypothèse du codage efficace* (Smith & Lewicki, 2006) tire ses origines des travaux sur les *statistiques des signaux naturels* (Simoncelli & Olshausen, 2001; McDermott & Simoncelli, 2011). Du point de vue de cette approche, les signaux acoustiques de communication seraient caractérisés par des propriétés statistiques régulières, lesquelles seraient au fondement des mécanismes d'analyse perceptive malgré la diversité apparente des réalisations sonores. Pour une présentation succincte des enjeux liés à l'hypothèse du codage efficace et aux liens qu'établit cette hypothèse avec les travaux sur les statistiques de signaux naturels, cf. par exemple Duniec *et al.* (2020). Diverses études récentes ont cherché à évaluer la validité de cette hypothèse pour la perception de la parole en lien avec les mécanismes de codage des informations sonores mis en oeuvre dans les implants cochléaires (Ming & Holt, 2009; Ueda & Nakajima, 2017; Grange & Culling, 2018).

1.1 Hypothèse du codage efficace et implant cochléaire

Ming & Holt (2009) ont mesuré les performances de reconnaissance de parole vocodée, souvent décrite comme simulant les informations diffusées par les implants cochléaires, chez des auditeurs normo-entendants. Ils 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 sur les taux de reconnaissance de mots et de segments phonétiques. Les performances sont meilleures si les localisations de ces frontières concordent avec des positions spectrales dérivées de modélisations issues de la théorie de l'information et correspondent donc à une « perspective efficace ».

Dans une toute autre perspective, Ueda & Nakajima (2017), 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 et procèdent, sur la base de signaux acoustiques de parole codés sur environ 20 canaux de représentation spectraleà des Analyses en Composantes Principales (ACP) portant sur les enveloppes d'énergie de ces canaux. Il font varier 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 seraient 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 » (ou efficace) de la parole.

Récemment, Grange & Culling (2018) ont répliqué l'étude de Ueda & Nakajima (2017) en modifiant légèrement l'algorithme d'analyse statistique (accroissement du nombre de canaux spectraux entrés dans l'ACP à environ 100 canaux notamment, estimation de la contribution de chacune des 20 premières composantes principales issues de l'ACP à travers les valeurs propres –*eigenvalues*). Ils ont ensuite évalué ces données à la lueur des performances observées en perception de parole vocodée (simulations d'implants cochléaires) et ont abouti à des conclusions assez similaires aux travaux précédents. 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 le graphique des valeurs propres (*scree plot*) qu'ils présentent, à un point d'inflexion au-delà duquel les valeurs propres semblent augmenter plus lentement. Cette limite est aussi associée dans les courbes de performance en fonction du nombre de canaux vocodés, à une amélioration moins marquée des performances observées à partir de 8 canaux

spectraux. Ces deux mesures (l'une statistique issue de signaux naturels, l'autre comportementale issue de signaux vocodés) seraient donc cohérentes et suggèreraient que cette limite de 7/8 canaux pourrait refléter une version optimale de la représentation perceptive des signaux de parole.

1.2 Généralisation à d'autres signaux sonores

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. Une telle constatation aurait un impact crucial sur les fondements ou la compréhension de cette hypothèse du codage efficace.

Les travaux antérieurs réalisés dans ce cadre (Ueda & Nakajima, 2017; Grange & Culling, 2018) se sont uniquement penchés sur des ensembles de données relativement limités en termes de généralisation (*clean speech* : parole lue sans sources sonores supplémentaires dans tous les cas, données correspondant à un unique locuteur dans le cas du travail de Grange & Culling, 2018). Les données de (Ming & Holt, 2009) portent sur une base de données multilocuteurs mais se concentrent aussi sur de la parole lue produite isolément. On peut se poser la question du statut de signaux naturels tout aussi essentiels pour l'espèce humaine : la parole bruitée ou correspondant à des mélanges de type *cocktail-party*, des signaux de parole dans lesquels apparaissent des variations acoustiques liées à des changements de locuteurs, la musique…

L'objet du présent travail est d'évaluer le degré de généralisabilité des travaux antérieurs en mettant en place une série d'analyses qui fourniront des éléments de comparaison des *propriétés statistiques de signaux naturels de musique* et de les comparer aux données obtenues sur la parole dans la littérature (Ueda & Nakajima, 2017; Grange & Culling, 2018).

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 OSF [1]. La base de données utilisée est la *Free Music Archive* (Defferrard *et al.*, 2017) qui contient des enregistrements de musique d'ensemble dans des styles divers. Pour une description plus précise de cette base de données, cf. Duniec *et al.* (2020).

Nous avons analysé deux échantillons de musique dont les durées totales de signal audio diffèrent. Le premier échantillon est d'une durée équivalente à celle qui a été étudiée pour les langues les

1. https://osf.io/r5bxk/?view_only=1eed0a88f17448818e653c57b8246643

plus fournies de l'échantillon analysé par Ueda & Nakajima (2017, 4000 s, env. 1h de données, 400 échantillons aléatoires de 10 s). C'est également la durée que nous avons estimée concernant l'analyse présentée par Grange & Culling (2018). L'autre échantillon de musique dure 6 h en tout et contient 2160 échantillons alétoires de 10 s. Cette durée totale de 6 h a été choisie afin de trouver un compromis optimal entre contraintes computationnelles (occupation des données en mémoire pour l'analyse statistique) et maximisation de la taille de l'échantillon. Le traitement des données a été réalisé sur un ordinateur personnel équipé de 32 Go de RAM.

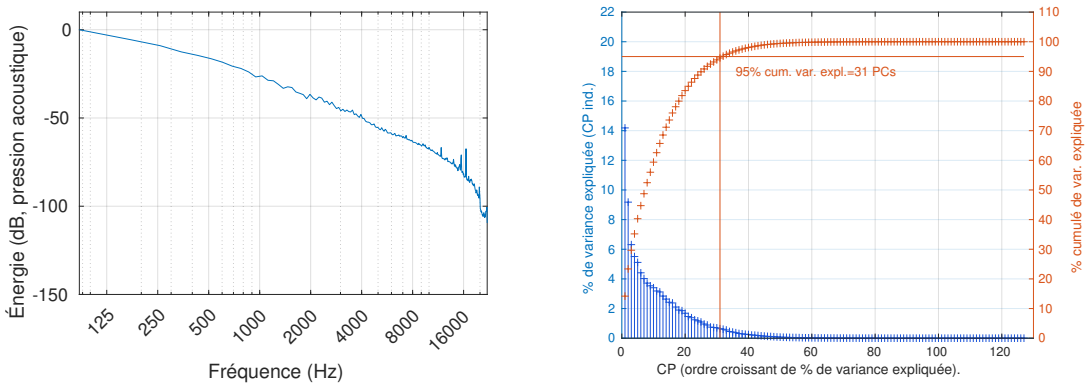
Pour chacun des deux ensembles de données, nous extrayons pour chaque enregistrement musical sélectionné un échantillon aléatoire de 10 s dont la position temporelle à l'intérieur du signal d'origine est sélectionnée au hasard. Au total, pour le premier ensemble de données, 400 stimuli musicaux ont été exploités, parmi lesquels aucune erreur de lecture par l'algorithme de décompression MP3 utilisé n'a été constatée. L'échantillon final est composé de 400 enregistrements audio fournissant une durée totale de 4000 s (soit environ 1 h) d'audio. Pour le second ensemble de données, 2162 stimuli musicaux ont été exploités, parmi lesquels 2 erreurs de décompression ont été constatées. L'échantillon final est composé de 2160 enregistrements audio fournissant une durée totale de 21600 s (soit 6 h) d'audio. Les résultats observés étant très proches, et pour des raisons de place, nous ne présentons ici que les données obtenues sur les 6 h de musique.

2.1.1 Paramètrage acoustique des signaux

Préalablement à l'analyse statistique des signaux, nous procédons à une paramétrisation acoustique comparable 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 logiquement restreints à des fréquences maximales d'environ 8 kHz. Nous avons estimé le spectre d'amplitude à long terme de l'échantillon de musique utilisé afin de déterminer une limite raisonnable de fréquence haute pour l'analyse et avons sélectionné la limite de 20 kHz, qui correspond à une diminution assez marquée de l'énergie relative pour les fréquences supérieures (cf Fig. 1a). Le spectre à long-terme sur la partie inférieure à 20 kHz présente une pente moyenne d'environ -14 dB / octave.

Il est très probable que cette limite observée ainsi que la forme globale du spectre à long-terme soient en partie liées à l'usage du format mp3 dans la base FMA et pas uniquement à des propriétés de la structure spectrale des enregistrements d'origine ou des signaux naturels produits lors de ces performances. Même si une base de données non-compressées aurait été bien évidemment plus intéressante du point de vue des formats informatiques utilisés, les enregistrements disponibles sont sous licence libre et ce point a guidé notre démarche afin de garantir les possibilités ultérieures d'évaluation de la reproductibilité de nos résultats dans le cadre d'une approche de science ouverte. En outre, notre analyse repose sur l'extraction de *corrélations des variations d'amplitude* entre canaux spectraux. Ces variations d'amplitude étant centrées-réduites pour leur Analyse en Composantes Principales, ceci rend la méthode relativement indépendante des altérations énergétiques des composantes spectrales telles que celles qui sont générées par l'algorithme de compression MP3.

Les portions de 10 s des enregistrements sélectionnés sont ensuite converties en format monophonique par combinaison des deux canaux stéréophoniques et concaténées les unes aux autres. Le signal concaténé est soumis à un filtrage passe-bas (fréquence de coupure 20 kHz) et sous-échantillonné à 40 kHz. 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, Moore & Glasberg, 1983, ce qui correspond à 127 canaux spectraux allant jusqu'à la



(a) Spectre à Long-Terme

(b) % de variance expliquée

FIGURE 1 – Résultats observés sur la concaténation de 6 h d'échantillons de 10 s de musique issus de la base de données *FMA* (*Free Music Archive*, Defferrard *et al.*, 2017). (a) spectre à long-terme du signal analysé, (b) Graphe des valeurs propres (% de variance expliquée) issu de l'Analyse Factorielle (en rouge : % associés à chaque composante principale individuelle — en bleu : % cumulés). La composante principale pour laquelle on atteint un cumul égal à 95% est indiquée.

fréquence supérieure maximale de 20 kHz). Ces enveloppes subissent une rectification demi-onde puis un filtrage passe-bas avec une fréquence de coupure de 50 Hz (fréquence d'échantillonnage 100 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 d'énergie entre les bandes de fréquence (corrélation entre les enveloppes d'énergie des canaux).

La matrice de données résultante, composée de 127 canaux de fréquence, correspond aux modulations temporelles de l'enveloppe de chaque canal spectral au cours du temps. Elle est transférée vers un outil statistique d'Analyse en Composantes Principales afin de procéder à une Analyse Factorielle.

2.1.2 Analyse Factorielle

L'Analyse Factorielle est une méthode descriptive d'analyse de données qui repose sur la technique d'Analyse en Composantes Principales (ACP). Elle 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 (statistiquement corrélées entre elles) en nouvelles variables *synthétiques*. Les composantes principales sont donc regroupées en facteurs abstraits. 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'Analyse Factorielle permettrait d'identifier les canaux de fréquence optimaux pour différencier de manière parcimonieuse les propriétés sonores distinctives d'un corpus.

Après application de l'ACP, les N composantes principales considérées (2 à 32 selon les cas) sont soumises à une rotation *varimax* qui a pour fonction de maximiser l'indépendance des facteurs et

de fournir en sortie un vecteur de coefficients de saturation (*factor loading*) pour chaque facteur retenu. Ces vecteurs représentent quelles zones de fréquence du spectre sont modulées de manière maximalement indépendante des autres zones (cf. Fig. 2).

3 Résultats

Pour l'analyse des résultats, nous avons adopté une approche comparative aux études précédentes tout en identifiant des critères supplémentaires d'évaluation des données qui n'auraient pas été mentionnés dans ces travaux. Nous décrivons dans un premier temps les résultats associés aux pourcentages de variance expliquée par les facteurs retenus. Nous nous penchons ensuite sur les courbes de coefficients de saturation représentant les regroupements de fréquences sonores.

3.1 Résultats : pourcentages de variance expliquée

Seul le travail de Grange & Culling (2018) donne des indications concernant les pourcentages de variance expliquée (valeurs propres –*eigenvalues*– normalisées) par les facteurs retenus. Ces données ont été extraites de la représentation graphique qu'en donnent les auteurs et devront être confirmées par une réplication de l'analyse sur de la parole comparable ainsi que par une vérification auprès des auteurs. Le graphe des valeurs propres (ang. *scree plot*) pour nos données est présenté dans la Fig. 1b. Elles sont normalisées afin de représenter les pourcentages de variance expliquée par chaque composante principale et sont associées à leurs sommes cumulées.

La première observation concernant les données de musique est que, contrairement à la description de Grange & Culling (2018) pour la parole, il n'apparaît pas de *point d'inflexion* marqué dans le graphe des valeurs propres. Ceci peut s'expliquer de deux manières. D'une part les signaux de parole utilisés par les auteurs correspondent à des enregistrements monolocuteurs : les informations liées aux fréquences de résonance peuvent donc prendre une part plus importante que les autres sources d'information dans la différenciation entre les signaux (variations de hauteur tonale –*pitch*– notamment, changements de qualité de la voix, altérations des résonances liées à des modifications de débit…), ce qui pourrait expliquer le point de rupture / d'inflexion observé en termes de quantité d'information entre d'une part des informations liées aux résonances qui caractériseraient les catégories phonémiques, et de l'autre des informations liées à des propriétés plus fines qui caractériseraient des aspects de l'intonation ou des propriétés vocales adoptées par le locuteur. En outre, ces informations de résonance sont caractéristiques d'un locuteur unique dans l'étude de Grange & Culling (2018). L'introduction de locuteurs distincts conduirait nécessairement à une dispersion plus marquée des fréquences de résonance associées. Il n'est donc pas garanti qu'un tel point d'inflexion émergerait en intégrant des données multilocuteurs dans l'analyse.

Dans l'étude de Grange & Culling (2018), seules les valeurs propres correspondant aux 22 premières composantes principales sont représentées. On note que sur leurs données de parole, le pourcentage cumulé maximal d'explication de la variance atteint seulement 66.4% à la 22^{e} composante principale et que le point d'inflexion sur lequel se fondent les auteurs pour mettre en relation les données perceptives avec les données statistiques (7 ou 8 canaux / composantes principales) correspond à des pourcentages cumulés d'explication de la variance respectivement égaux à 33.6% et 37.2% (pour 7 et 8 composantes principales). Ces mesures semblent parfaitement prévisibles pour de la parole : on sait par exemple qu'une part non négligeable des informations correspondant à des

oppositions phonémiques est accessible avec peu de canaux en parole vocodée ou chez des patients utilisant un implant cochléaire dans le silence (Shannon *et al.*, 1995; Bouton *et al.*, 2012) mais que les performances sont altérées pour certaines oppositions ou dans des conditions non-optimales de transmission du signal.

Considérant les pourcentages de variance expliquée à travers les données de musique, et si l'on se restreint aux 22 premiers facteurs afin de pouvoir comparer avec la parole, on atteint déjà un cumul d'environ 86 % de variance expliquée à la 22e CP : 20 points de plus que pour la parole monolocuteur. Concernant les résultats observés autour de 7 ou 8 facteurs, les valeurs propres cumulées atteignent 48.7 et 52.4% pour la musique : 10 à 15 points de plus que pour les données de parole. Nous considérons ces différences dans la discussion.

3.2 Résultats : coefficients de saturation

Ueda & Nakajima (2017) et Grange & Culling (2018) proposent des éléments d'analyse des coefficients de saturation. Nous procédons de même pour un nombre variable de composantes principales retenues (3, 4, 7, 8, 12, 16, 24 et 32). Celles-ci sont soumises à une rotation orthogonale *varimax* qui a pour fonction de maximiser l'indépendance des facteurs et de fournir en sortie un vecteur de coefficients de saturation (ang. *factor loading*) pour chaque facteur. Tout comme dans les travaux antérieurs, ces valeurs sont représentées sous forme de courbes qui font ressortir des gammes de fréquences corrélées entre elles (Fig. 2) ce qui permet de décrire l'empan des fréquences qui sont associées au sein d'un même facteur synthétique.

Dans l'analyse réalisée par Ueda & Nakajima (2017), les frontières entre canaux decrits comme optimaux sont placées aux croisements des courbes de saturation, délimitant ainsi des zones de fréquences. Dans la Fig. 2, nous avons représenté les frontières proposées par les auteurs à travers des traits verticaux de couleur orange. L'analyse proposée par les auteurs identifie 3 frontières qui délimitent quatre zones de fréquence principales.Une exploration visuelle rapide du graphique permet de repérer que l'organisation des courbes de saturation semble se stabiliser au-moins globalement à partir de 12 facteurs ou en tout cas au-dessus de 8. Au-delà, on constate une amélioration progressive de la différenciation des fréquences mais pas de réorganisation massive. Les données sont en cours d'analyses complémentaires afin de déterminer par le calcul les frontières spectrales séparant ces différentes composantes en fonction du nombre de facteurs.

4 Discussion

Concernant l'analyse des pourcentages de variance expliquée à travers les signaux de musique, les résultats sont en partie non conformes à nos attentes initiales. Si l'on se restreint aux 22 premiers facteurs, et en fondant nos prédictions sur les différences d'organisation sonore importantes qui caractérisent la parole monolocuteur et des enregistrements de musique d'ensemble, nous nous attendions clairement à ce que nos données fassent ressortir des phénomènes impliquant une élévation du nombre de facteurs ou des valeurs propres plus faibles pour un même nombre de facteurs. Or on observe l'inverse : on atteint déjà un cumul d'environ 86 % de variance expliquée à la 22e CP : 20 points de plus que pour la parole monolocuteur et les données relevées pour 7/8 facteurs donnent des valeurs supérieures de 10 à 15 points.

Il est probable qu'une grande partie de la variance exprimée ici repose sur des informations rythmiques associées à des événements acoustiques à spectre large ainsi qu'à une différenciation entre des gammes spectrales de timbres instrumentaux (gamme tonale plutôt aiguë / medium / grave). Cette suggestion peut être au-moins en partie confirmée par les données de la Fig. 2 associées à un nombre réduit de facteurs (entre 3 et 8).

Il est possible que les *poids* des contributions respectives des premières composantes principales ne s'organisent pas de la même manière pour la musique (dans l'ordre : rythme, timbres globaux puis tonalité) et pour la parole isolée (un ordre possible serait : timbre associé aux segments puis intonation et rythme), ce qui pourrait expliquer ces divergences entre nos attentes initiales et les résultats.

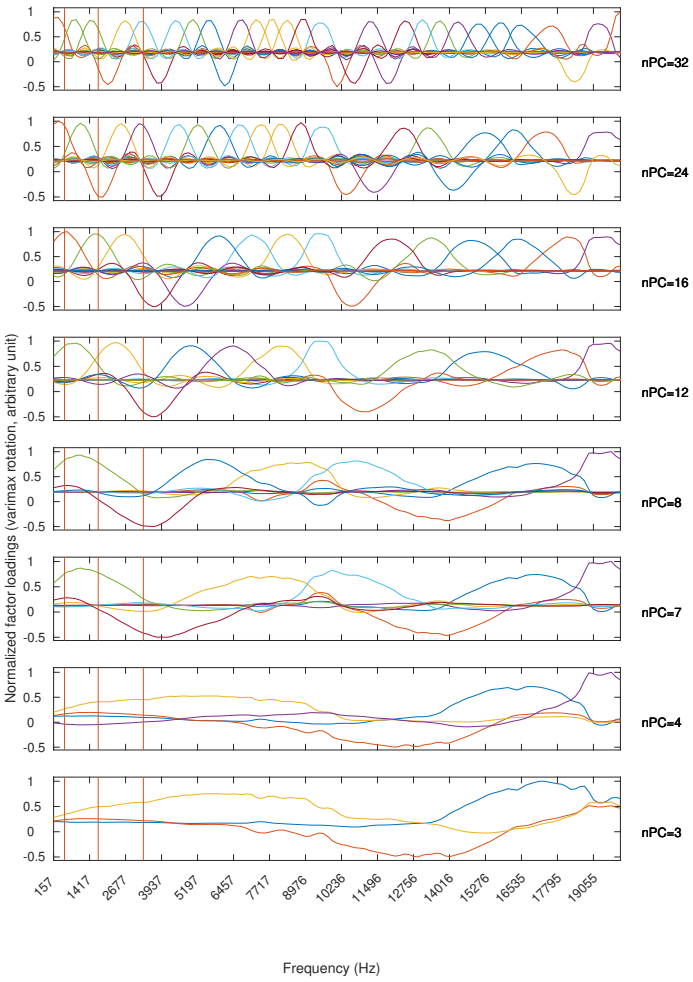


FIGURE 2 – Coefficients de saturation (ang. *factor loadings*) après rotation varimax issus des ACP réalisées sur un échantillon de 6 h de musique issu de la base de données *FMA* (*Free Music Archive*, Defferrard *et al.*, 2017), fréquence maximale supérieure de 20 kHz.

Si l'on observe les données de courbes de saturation (Fig. 2), on voit clairement que les frontières optimales proposées par Ueda & Nakajima (2017) pour la parole ne correspondent à aucune répartition comparable des pics associés à ces courbes pour une décomposition en 4 facteurs. Même si cette observation était prévisible étant données les propriétés spectrales de la parole telles qu'elles peuvent être mises en relation avec les hypothèses de Ueda & Nakajima (2017, concentration de l'information principale dans des fréquences inférieures à 8 kHz), nous nous orientons maintenant vers l'évaluation de la correspondance entre les valeurs de fréquence des frontières issues de la parole et de la musique afin de pouvoir envisager les perspectives offertes en termes de représentations perceptives.

Nous finalisons l'automatisation du calcul des frontières de fréquence et vérifierons les résultats antérieurs sur une base de données de parole ainsi que sur des données alternatives (mélanges de locuteurs, enregistrements mono-instrumentaux). Des études perceptives permettront d'évaluer la validité de ces frontières pour la reconnaissance de mélodies et la perception de la f_0.

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

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.

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*, **138**(3), EL347–EL351.

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

DUNIEC A., CROUZET O. & DELAIS-ROUSSARIE E. (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. In C. BENZITOUN, C. BRAUD, L. HUBER, D. LANGLOIS, S. OUNI, S. POGODALLA & S. SCHNEIDER, Eds., *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). Volume 1 : Journées d'Études sur la Parole*, p. 181–189, Nancy, France : ATALA.

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*, **1169**(1), 518–533.

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

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.

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

MOORE B. C. J. & GLASBERG B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, **74**, 750–753.

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*, **41**(3), 707–712.

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.

SMITH E. C. & LEWICKI M. S. (2006). Efficient auditory coding. *Nature*, **439**(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.