École des Hautes Études en Sciences Sociales

École doctorale de l'EHESS

Centre d'Analyse et de Mathématique sociales

Thèse de doctorat
Discipline : Sciences cognitives
(option Neurosciences computationnelles)

FRANÇOIS DELOCHE

Codage efficace de la parole à court terme

Short time-scale efficient coding of speech

Thèse dirigée par: Jean-Pierre Nadal

Date de soutenance : 22 octobre 2019

Rapporteurs	1	Frédéric BIMBOT, IRISA Rennes
	2	Bruno TORRÉSANI, Université Aix-Marseille
Examinateurs	3	Gilles CHARDON, CentraleSupélec
	4	Judit GERVAIN, Université Paris-Descartes
	5	Shihab SHAMMA, ENS Paris

Thèse rédigée en Markdown (cf Ref. 127).

Remerciements

J'adresse en premier lieu mes plus sincères remerciements à mon directeur de thèse, Jean-Pierre Nadal. Il m'a orienté, dès le début de mon stage de master M2 (Mathématiques, Vision, Apprentissage), sur ce sujet qui m'a passionné pendant ces 3 ans ½– et continue de me passionner ! – je lui en suis grandement reconnaissant. Je le remercie pour s'être toujours rendu disponible afin de répondre à mes interrogations et pour toute l'aide qu'il m'a apporté durant ces années. J'ai réellement apprécié la liberté qu'il m'a accordée dans le choix des directions de recherche, sans laquelle la thèse aurait pu prendre un tout autre chemin. Il a toujours accueilli mes diverses idées et propositions avec intérêt et bienveillance, tout en s'assurant que je sois bien encadré.
Je veux associer à ces remerciements Laurent Bonnasse-Gahot, qui, bien que son nom ne figure pas sur la couverture de cette thèse, a été pour moi comme un second encadrant. Je le remercie chaleureusement pour ses précieux conseils. Les nombreuses discussions, que j'ai eu le plaisir d'avoir avec lui, ont fortement contribué au bon déroulement de la thèse et m'ont personnellement beaucoup apporté.

Je tiens à remercier également les membres du projet ANR *SpeechCode* : merci à Judit Gervain, pour m'avoir indirectement amené à travailler sur ce sujet captivant, pour m'avoir aidé, avec Ramon Guevarra Erra, à plusieurs reprises, et pour avoir accepté de faire partie du jury de thèse. Merci à Christian Lorenzi pour son soutien et ses conseils d'une grande valeur ; j'ai été enthousiasmé par les discussions passionnantes autour de l'audition que j'ai eues avec lui.

Je remercie vivement Bruno Torrésani et Frédéric Bimbot d'avoir accepté d'être rapporteurs pour cette thèse, ainsi que Shihab Shamma et Gilles Chardon de me faire l'honneur de faire partie du jury. J'adresse des remerciements plus larges à tous les chercheurs avec qui j'ai été amené à interagir ces trois années. Je souhaite citer ici Monika Dörfler, que j'ai rencontrée à l'école d'été de Peyresq (2018), qui a fait le lien entre mon travail et les ondelettes de Gabor *flexibles*, et qui m'a ainsi initié à toute une bibliographie dont je n'avais pas beaucoup conscience. Merci enfin à Gabriel Peyré et Emmanuel Dupoux pour avoir fait partie du comité de suivi de thèse.

J'ai eu la chance d'être accueilli au Centre d'Analyse et de Mathématique sociales (CAMS), à l'EHESS, qui a été pour moi un lieu idéal afin de mener à terme cette thèse. Ces très bonnes conditions ont été rendues possibles par la présence et le travail des membres du CAMS, en particulier Sandrine Nadal, Nathalie Brusseaux, Thomas Tailpied et Francesca Aceto. Je salue amicalement tous les étudiants et post-docs que j'ai côtoyés au CAMS (ou à l'ENS), et qui ont considérablement amoindri l'âpreté de la thèse par les moments passés avec eux (par ordre approximatif d'apparition) : Kévin B., Quentin F., Romain D., Samuel N., Alessandro Z., Andrea T., Elisa A., Benedetta F., Antoine P., Noemi M., José M., Charles L., Beniada S., Gabrielle N., Federico B., Julien B., Imke M., Nicolas (Songshen) Y.

A titre plus personnel, je remercie famille & amis de longue date, à commencer par mes parents, pour leur soutien depuis de nombreuses années. C'est en pensant à ce *long-time scale* que je souhaite rendre hommage aux professeurs qui m'ont fait apprendre tant de choses durant mes études, et inspiré cet intérêt pour les sciences, la thèse marquant la fin d'un long processus de formation qui ne fait pas en vérité que trois ans ! Merci à mes colocs (Lawrence & Dhafer) qui m'ont supporté quotidiennement. Enfin, je ne pourrais pas terminer cette section de remerciements sans faire mention des soutiens IVL (Loup, Arthur, Constance & tous les autres :pandayay: :love:).

Résumé

L'analyse de données de parole a montré que la sélectivité fréquentielle de la cochlée est adaptée à la structure statistique de la parole. Ce résultat est conforme à l'hypothèse du codage efficace selon laquelle le traitement sensoriel adopte un schéma de codage qui est optimal pour les stimuli naturels. Cependant, le signal de la parole possède une structure riche, même sur des petites échelles de temps, du fait de la diversité des facteurs acoustiques à l'origine de la génération de la parole. Cette complexité de structure motive l'idée qu'une représentation non linéaire de la parole pourrait aboutir à un schéma de codage plus efficace qu'une simple représentation linéaire. La première étape dans la recherche de stratégies efficaces est la description de la structure statistique de la parole à un niveau fin. Dans cette thèse, j'explore la structure statistique au niveau des phonèmes en adoptant une approche paramétrique pour la représentation du signal. La décomposition la plus parcimonieuse est recherchée parmi une famille de dictionnaires de filtres de Gabor dont la sélectivité fréquentielle suit différentes lois de puissance dans la gamme des hautes fréquences 1-8kHz. L'utilisation de ces dictionnaires comme représentations temps-fréquence parcimonieuses est justifiée mathématiquement et empiriquement. Un lien formel avec les travaux précédents, fondés sur l'Analyse en Composantes indépendantes (ACI), est présenté. Les lois de puissance des représentations parcimonieuses offrent une interprétation riche de la structure statistique de la parole, et peuvent être reliées à des facteurs acoustiques clés déduits de l'analyse de données réelles et synthétiques. Les résultats montrent en outre qu'une stratégie de codage efficace, reflétant le comportement non linéaire de la cochlée, consiste à réduire la sélectivité fréquentielle avec le niveau d'intensité sonore.

Mots clé : Hypothèse du codage efficace, Codage parcimonieux, dictionnaires de Gabor, phonétique acoustique, codage auditif, statistiques de la parole, analyse temps-fréquence, Analyse en Composantes Indépendantes.

Abstract

Cochlear frequency selectivity is known to reflect the overall statistical structure of speech, in line with the hypothesis that low-level sensory processing provides efficient codes for information contained in natural stimuli. Speech signals, however, possess a complex structure, even on short-time scales, as a result of the diversity of acoustic factors involved in the generation of speech. This rich structure means that advanced coding schemes based on a nonlinear representation of speech sounds could provide more efficient codes. The first step in finding efficient strategies is to describe the statistical structure of speech at a fine level — at the level of phonemes or even finer at the level of acoustic events. In this thesis, I use a parametric approach to explore the fine-grained statistical structure of speech. The goal of this method is to find the sparsest representation of speech sounds among a family of dictionaries of Gabor filters whose frequency selectivity follows different power laws in the high frequency range 1-8kHz. I motivate the use of Gabor filters for the search of sparse time-frequency representations of speech signals, and I show that the dictionary method has a formal link with previous work based on Independent Component Analysis (ICA). The acoustic factors that affect the power law associated with the sparsest decomposition can be inferred from the analyses of synthetic and real data. The results suggest that an efficient speech coding strategy is to reduce frequency selectivity with sound intensity level, reflecting the nonlinear behavior of the cochlea.

Keywords : Independent Component Analysis, efficient coding hypothesis, sparse coding, Gabor dictionaries, acoustic phonetics, auditory coding, speech statistics, time-frequency analysis.

Table des matières
Table of contents

Codage efficace de la parole à court terme

Résumé substantiel en langue française

Introduction

Contexte : le projet *SpeechCode*

En 1944, Alvin Liberman tentait de réaliser au sein des laboratoires Haskins une machine à lire destinée aux aveugles. Ce système artificiel, inspiré de la parole, faisait correspondre chaque lettre à un son. A. Liberman pensait que ce système n'était pas fondamentalement différent de la parole classiquement considéré comme une succession de *phones* (unités acoustiques) associés de manière univoque à des *phonèmes* (unités linguistiques). Sa confiance en son mécanisme fut toutefois diminuée lorsque les performances du système se sont révélées être très en dessous de ses attentes, cela malgré des efforts répétés pour augmenter l'expressivité de la machine. Cette expérience témoigne de la difficulté de reproduire un code sonore aussi efficace que la parole en termes de vitesse et d'expressivité. Ce problème a amené A. Liberman à se poser la question suivante : en quoi le signal de la parole est-t-il spécial ? Quelles sont les caractéristiques qui en font un *code* adapté au système auditif ? Cette question l'a animé durant toute sa vie de chercheur [96]. A. Liberman a développé la théorie motrice de la perception de la parole selon laquelle le « code » de la parole est contenu dans la représentation mentale de la succession des mouvements de l'appareil vocal au moment de la prononciation [97]. Cette théorie est une proposition de représentation pertinente de la parole pour le système vocal et auditif. Pour A. Liberman, le code de la parole ne peut pas être réduit à une description simplifiée des traits acoustiques segment par segment. Il faut chercher une représentation plus abstraite qui prenne en compte la complexité du signal.

Les progrès scientifiques et technologiques permettent aujourd'hui d'aborder le problème du « code de la parole » d'une nouvelle manière. La complexité du signal de la parole n'est plus une barrière comme cela l'a été dans les années 1940 pour Alvin Liberman. Deux changements simultanés expliquent un changement de paradigme associé à des progrès remarquables pour la modélisation du signal de la parole : d'une part, un accès facilité à de grandes bases de données et ressources computationnelles conséquentes ; d'autre part, une utilisation plus systématique de méthodes d'apprentissage machine, avec des outils algorithmiques dédiés. Ces évolutions ont entraîné une transition progressive d'approches classiques de modèles conçus sur mesure grâce aux connaissances de l'ingénieur sur le signal à analyser, à des méthodes plus flexibles davantage axées sur les données. Les systèmes de reconnaissance de la parole ou de synthèse vocale entraînés par des méthodes d'apprentissage profond (*deep learning*), développés ces dernières années, affichent des performances bien meilleures que les systèmes les ayant précédés [170]. Les réseaux neuronaux profonds convertissent le signal acoustique brut en une représentation abstraite utile pour la reconnaissance ou la synthèse vocale, apportant une réponse pratique au problème de Liberman. Néanmoins, ces algorithmes se comportent un peu comme des « boîtes noires ». Ils ne permettent pas d'apporter des réponses sur la spécificité de la

parole : quelles en sont les propriétés qui en font un signal efficace pour la transmission d'information ? Existe t-il des principes computationnels qui lui sont spécifiques dans le système auditif, ou qui doivent être intégrés dans des systèmes artificiels de reconnaissance de la parole ? A quel niveau ? Ces questions motivent des approches multidisciplinaires pour *déchiffrer* le code de la parole. L'objectif est d'obtenir une description du code de la parole à tous les niveaux, ou, parce que la complexité du signal de la parole ne permet pas une compréhension complète sans outils computationnels, de trouver à défaut les principes clés qui en justifieraient l'efficacité. Une description du code de la parole pourrait nous aider à comprendre la façon dont nous analysons et acquérons la parole. Cela pourrait permettre à terme la conception d'appareils auditifs ou de systèmes de reconnaissance automatique de la parole optimisés pour leur tâche.

Cependant, le « déchiffrage » du code de la parole nécessite un large éventail de connaissances car la description doit englober toutes les étapes du processus, de la réalisation physique du signal acoustique jusqu'à sa représentation corticale abstraite, en passant par la représentation neuronale intermédiaire en sortie du système auditif périphérique. Participer à ce déchiffrage est l'ambition du projet *SpeechCode* faisant concourir trois domaines de compétence : psycholinguistique, psychoacoustique et modélisation computationnelle des systèmes sensoriels. Dans cette thèse, je présente la partie modélisation computationnelle de ce projet. Nous nous intéressons particulièrement aux premières étapes du traitement de la parole qui pourraient intervenir tôt dans l'acquisition de la parole. L'accent est ainsi mis en premier lieu sur le traitement périphérique. Si une comparaison est faite avec les réseaux de neurones profonds, cela correspond aux premières couches d'un réseau prenant le signal acoustique en entrée. Au niveau physiologique, les mécanismes associés se situent au niveau cochléaire et éventuellement dans le tronc cérébral. L'essentiel des travaux menés concerne donc la perception auditive de bas-niveau. Cependant, un volet porte également sur un aspect de plus haut niveau : celui de la perception du rythme de la parole. Bien qu'il s'agisse davantage d'une compétence de haut-niveau, la perception du rythme semble en effet intervenir à un stade précoce de l'acquisition de la langue.

L'une des principales préoccupations du traitement sensoriel de bas niveau est la façon dont le cerveau acquiert des représentations compactes des entrées sensorielles. Une réponse est que les neurones investissent un minimum de ressources dans le codage sensoriel, réalisant ainsi un code neuronal compact, de la même manière que nous préférons traiter des fichiers de données compacts sur nos ordinateurs. Le point de vue selon lequel les processus neuronaux cherchent à augmenter l'efficacité du codage lorsque des stimuli naturels sont présentés est appelé l'*hypothèse du codage efficace*, formulée pour la première fois par Horace Barlow en 1961 [16]. La théorie du codage efficace est associée à des méthodes de calcul pour étudier la structure statistique des signaux sensoriels et pour relier ces propriétés au codage sensoriel. Ces méthodes comprennent l'analyse en composantes indépendantes (ACI) et les méthodes de codage parcimonieux (*sparse coding*). Les analyses de données selon ces outils ont permis de comprendre en particulier certains aspects du codage des stimuli visuels. Des études menées dans les années 1990 ont montré par exemple que le profil des champs récepteurs des neurones du cortex visuel peut être retrouvé simplement à partir des statistiques des images naturelles [120, 164]. Des résultats comparables existent pour le système auditif. En 2002, Michael S. Lewicki a montré que l'analyse en composantes indépendantes appliquée à de petits fragments de parole – d'environ 10 ms – produit une famille de filtres dont la sélectivité fréquentielle est similaire aux filtres cochléaires [94]. Sur la base de ces travaux, il a formulé l'hypothèse suivante : sur des échelles de temps très courtes, la parole est adaptée de manière optimale au système auditif périphérique

des mammifères. Récemment, Christian E. Stilp et M. S. Lewicki ont approfondi ce sujet et ont montré que des variations significatives dans le code optimal sont trouvées lorsque différentes classes phonétiques sont présentées à l'entrée de l'ACI [153]. Ces variations sont caractérisées par un changement de comportement de la sélectivité fréquentielle dans les hautes fréquences 1-8kHz. En d'autres termes, le compromis temps-fréquence est différent selon la classe phonétique considérée : une décomposition fréquentielle convient aux fricatives, alors qu'une décomposition temporelle est plus adaptée aux plosives. Les voyelles sont mieux représentées par une décomposition intermédiaire. L'étude menée par C.E. Stilp et M.S. Lewicki soulève des nouvelles questions : 1) Quels signaux ou propriétés acoustiques expliquent les variations du code optimal (les différents comportements concernant la sélectivité fréquentielle) ? 2) Est-ce qu'un système de codage pourrait tirer profit de la diversité de structure révélée par l'ACI, en adaptant la représentation en fonction du signal d'entrée ? 3) Si la réponse à la dernière question est positive, cette stratégie est-elle mise en œuvre dans le système auditif ? L'objectif de cette thèse est, en décrivant la structure statistique « à grain fin » de la parole, d'envisager les manières d'exploiter cette structure pour coder efficacement le signal. Il s'agit également de voir si les stratégies ainsi révélées peuvent correspondre au codage auditif au niveau périphérique (cochlée). Les objectifs seront introduits de manière plus formelle à la fin de cette introduction, à la suite de la présentation de l'approche et du contexte scientifique.

Approche

Neurosciences computationnelles

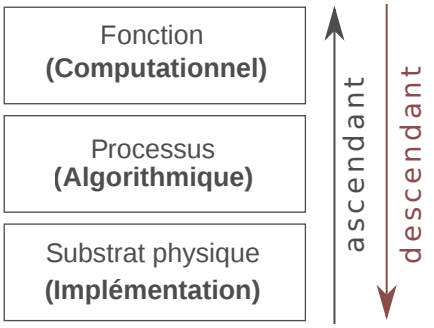
Les trois niveaux d'analyse (David Marr) . L'approche de cette thèse est celle des neurosciences computationnelles, telle que définie par David Marr et Tomaso Poggio. Dans leur article fondateur de 1976 [108], David Marr et Tomaso Poggio définissent trois niveaux d'analyse des systèmes complexes de traitement de l'information :

1. le niveau physique de l'implémentation (les récepteurs sensoriels, les connexions neuronales. . .) ;
2. le niveau des mécanismes et des algorithmes, ainsi que les représentations qui en émergent (ex : l'opération de filtrage réalisée par la cochlée et la réprésentation tonotopique [1] des sons) ;
3. le niveau fonctionnel ou « computationnel » : quelle est la finalité de ces opérations ? à quel problème théorique le système répond-il ? (ex : analyse temps-fréquence du signal) ;

Selon D. Marr et T. Poggio, la description du système à ces trois niveaux d'analyse est nécessaire pour le comprendre entièrement. Bien que le dernier niveau semble le plus évident, ils soutiennent au contraire qu'il s'agit d'un niveau « crucial mais négligé ». Parfois, même la description exhaustive d'un système élément par élément ne permet pas d'en discerner la finalité si le problème auquel il répond n'est pas connu. C'est l'enjeu du champ des neurosciences computationnelles de décrire les principes théoriques auxquels

1. La tonotopie est la propriété de faire correspondre un paramètre spatial (ex : position le long de la cochlée) à la fréquence.)

sont contraints les opérations neuronales, et de proposer des solutions algorithmiques qui peuvent correspondre à des implémentations biologiques.



En se référant à la vue des trois niveaux d'analyse, on peut définir schématiquement deux approches. L'approche *ascendante* consiste à faire des observations sur le système biophysique, puis en décrire les processus et expliquer leurs fonctions. L'approche *descendante* fait le chemin inverse. Cela consiste d'abord à poser une hypothèse sur la fonction du système, puis à trouver les propriétés théoriques qu'un système possédant cette fonction doit inclure. L'analyse peut conduire à faire une prédiction sur les mécanismes qui doivent être mis en place, et on peut ensuite vérifier si le système biophysique les implémente effectivement. Cette vérification peut se faire avec des données physiologiques préexistantes ou elle peut motiver la réalisation de nouvelles expériences. L'approche descendante est adoptée dans cette thèse. L'hypothèse initiale sur la fonction de l'oreille interne est que celle-ci code efficacement le signal de la parole. L'approche descendante m'a amené à travailler sur des propriétés du système auditif qui n'étaient pas l'objet d'étude au début de la thèse. L'analyse des signaux de parole a montré que les propriétés statistiques sont en accord avec le comportement non linéaire de la sélectivité fréquentielle cochléaire. La possibilité que les non-linéarités du filtrage cochléaire correspondent à une stratégie de codage efficace est discutée dans le chapitre 3 (chap. 5 dans la version anglaise). Cependant, le lien avec les non-linéarités cochléaires est venu au cours de la recherche de la thèse, qui n'avait pas initialement comme objectif de modéliser les non-linéarités des filtres auditifs.

Le cadre de la théorie de l'information. Les neurosciences computationnelles se prêtent bien à l'étude des systèmes sensoriels dont l'une des fonctions clés est de transmettre l'information sensorielle (visuelle, auditive, etc.). Ces fonctions rappellent les problèmes que l'on rencontre en traitement du signal et théorie de l'information et peuvent se traduire aisément en langage mathématique.

Dans cette thèse, je considère les principaux objets d'étude (la parole et le système auditif) comme faisant partie d'un système de communication (voir fig. 1 reprenant le diagramme originel de Shannon). La parole est l'élément central de ce système de communication, qui a la particularité que le système biologique se trouve des deux côtés du diagramme : le cerveau est à la fois l'*encodeur* et le *décodeur*. [*Codage* : le cerveau envoie des instructions pour produire une phrase. *Décodage* : l'information sémantique initiale est extraite du signal acoustique, après son analyse par l'oreille]. Cependant, même si certains aspects de la production vocale sont abordés dans cette thèse, je me concentre principalement sur la partie réceptive, c'est-à-dire le système auditif. Aussi quand j'utilise le terme d'« encodage », ou simplement de « codage », je ne parle pas de l'émission du signal mais bien de la partie réceptrice. De manière générale, la terminologie que j'emploie est largement inspirée de la théorie de l'information. La parole (ou le stimulus) est considérée comme la réalisation d'un processus stochastique : on l'appelle aussi la *source* ou le *signal d'entrée*. L'oreille interne décompose ensuite le signal acoustique en bandes de fréquences et convertit le signal mécanique en impulsions électriques qui sont transmises au cerveau par le nerf auditif. Le *code neuronal* émerge de cette activité d'un grand nombre de neurones. L'ensemble des activations correspond à une représentation sous-jacente,

parfois appelée *représentation neuronale*, qui est utile pour les tâches de niveau supérieur. Mathématiquement, elle correspond à un processus stochastique multivarié (le signal de sortie) obtenu par une transformation du signal d'entrée. Ce processus peut être considéré soit comme un canal unique codant pour des données multivariées (dont les processus marginaux, pris séparément, sont appellés *composantes de sortie*), soit comme plusieurs canaux de codage interconnectés. Dans les deux cas, il s'agit d'une abstraction de l'activité neuronale en réponse à un stimulus vocal. Je me limite volontairement à cette description quelque peu éloignée de la réalité physique du fonctionnement neuronal, l'objectif n'étant pas l'adéquation physiologique mais de comprendre les principes computationnels qui ont une incidence sur le codage sensoriel. Le paradigme de Shannon serait insuffisant si le but à atteindre était de décrire fidèlement l'activité cérébrale, en particulier la notion de « code neuronal » ne rend pas compte de la dynamique des neurones [27]. En fait, le « code neuronal » est implicite dans les analyses. C'est la caractérisation des propriétés statistiques des processus entrée/sortie – les densités de probabilité des lois jointes et marginales – qui est en réalité déterminante dans les algorithmes.

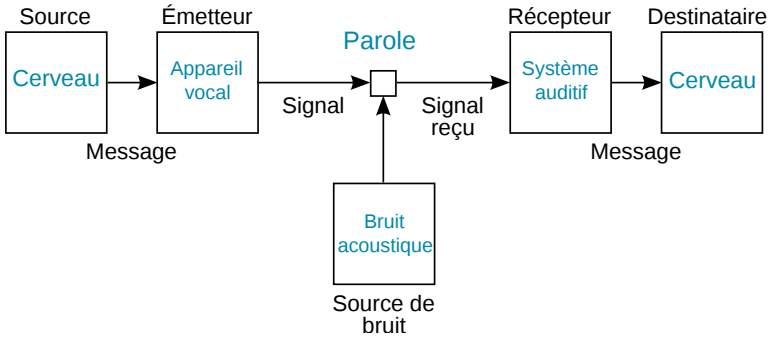


FIGURE 1 – La parole comme signal central d'un système de communication. Reprend le schéma original de Shannon [142].

La théorie de l'information de Shannon joue un rôle crucial dans l'étude des systèmes sensoriels reposant sur l'hypothèse de codage efficace (comme rendu explicite par le nom donné à l'hypothèse). En considérant les systèmes sensoriels comme des systèmes de communication, il est possible d'attribuer des critères de performance à l'activité neuronale qui ont un sens dans la théorie de Shannon. L'idée que les connexions cérébrales s'organisent au cours du développement, ou de l'évolution, pour optimiser un de ces critères caractérise l'hypothèse du codage efficace. Cette hypothèse est présentée dans les grandes lignes dans la section suivante et développée plus en détail dans le chapitre 1. C'est le point de départ des travaux que contient cette thèse. La formulation initiale de l'hypothèse de codage efficace [16] reprend la notion de redondance introduite par Shannon une décennie auparavant. Depuis lors, de nombreux autres critères d'efficacité des représentations, qui ont un sens dans la théorie de Shannon, ont été proposés [17, 10, 21]. Un autre critère, qui joue un rôle central dans cette thèse, est la *parcimonie* des activations neuronales [121]. Cette notion se démarque légèrement des autres parce qu'elle est déjà exprimée en termes probabilistes/statistiques, mais nous verrons qu'elle peut également se comprendre dans le cadre de la théorie de l'information.

Interdisciplinarité

La nécessité d'approches interdisciplinaires. Les algorithmes qui sont fondés sur la théorie du codage efficace sont les principaux outils d'analyse dans ce travail de recherche. Cependant, ces algorithmes ne suffisent pas toujours à décrire et à interpréter les propriétés statistiques qui sont ainsi révélées de manière compréhensible. Si les algorithmes d'apprentissage machine permettent de modéliser des données structurées complexes, la connaissance approfondie des données et de la manière dont elles ont été générées est parfois encore nécessaire. Un algorithme doit souvent être ajusté en fonction de la structure explicite des données en entrée. Sans cette familiarité avec les données, il est beaucoup plus difficile de concevoir des algorithmes de codage avancés spécifiques. C'est particulièrement vrai en ce qui concerne les données pour lesquelles on dispose d'une grande quantité d'informations et de connaissances. Ces connaissances pourraient être exploitées pour développer des stratégies de codage efficaces, mais il peut être difficile de les appréhender toutes à la fois. La conception d'algorithmes efficaces et la description des données, de manière compréhensible et cohérente, sont souvent séparées, parce qu'elles exigent des compétences et des connaissances dans des domaines différents. Le signal de la parole comporte un tel degré de complexité que son expertise englobe plusieurs domaines nécessitant des compétences différentes. En phonétique, on distingue la phonétique articulatoire (étude des organes de la parole et de la production des sons), la phonétique acoustique (étude du signal acoustique et de ses propriétés) et la phonétique auditive (étude de la réception et de la perception des sons de la parole), qui sont toutes étudiées selon des approches différentes. Sans cette connaissance approfondie de la parole, la conception de systèmes intelligents pourrait passer à côté d'aspects importants qui seraient nécessaires pour concevoir des systèmes entièrement adaptés. De plus, la connaissance des mécanismes de codage de l'audition est également requise dans les travaux où le système auditif est également l'objet d'étude. Ceci plaide en faveur d'approches à l'interface de plusieurs disciplines. L'interdisciplinarité est au cœur du projet *SpeechCode* et du Centre d'Analyse et de Mathématiques sociales (CAMS) où une partie de ce projet a été réalisée.

Disciplines impliquées dans le déchiffrage du « code de la parole » sur des courtes échelles de temps. Étant donné la complexité des données de parole à différentes échelles de temps et de fréquence, il est préférable de se concentrer sur un aspect en particulier. Ici, les propriétés pertinentes sont celles qui apparaissent sur des échelles de temps très courtes (~ 10 ms) et dans la gamme des hautes fréquences (1-8kHz). L'échelle de temps correspond à un cycle glottal pour les sons voisés. Il ne s'agit pas d'un choix courant pour l'analyse de la parole : la plupart des études font le choix de fenêtres plus longues et s'intéressent préférentiellement aux basses ou moyennes fréquences. Cela s'explique d'une part parce que considérer des fenêtres courtes nécessitent une grande puissance de calcul pour traiter la parole en temps réel, d'autre part parce qu'une grande partie de l'information phonétique est contenue dans des fréquences inférieures à 1 kHz (fréquence fondamentale, premiers formants…). Le choix d'échelles de temps courtes est motivé par les temps caractéristiques des réponses impulsionnelles des filtres cochléaires dans la gamme des hautes fréquences, qui ne sont que de quelques millisecondes, et par l'analyse en composantes indépendantes (ACI) qui produit des filtres localisés quand elle est appliquée à la parole [3]. Les disciplines impliquées dans la description du « code de la parole » sur de courtes échelles de temps sont représentées schématiquement sur la figure 2. Les outils

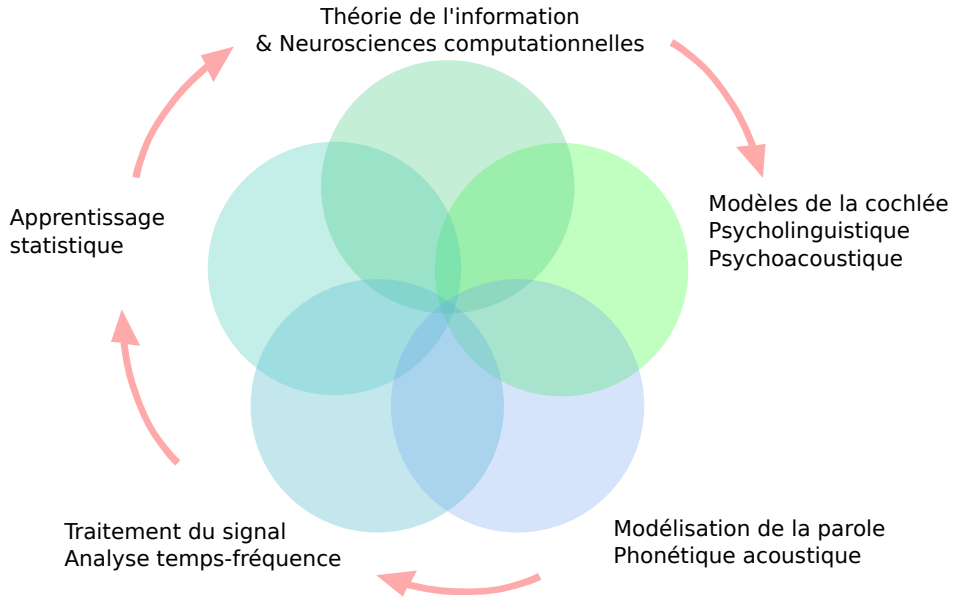


FIGURE 2 – Comprendre le « code de la parole » nécessite de rassembler les connaissances de différents domaines : connaissance des algorithmes et des méthodes quantitatives, ainsi que la connaissance des données (le signal de la parole). La figure est un diagramme schématique des disciplines impliquées. Les flèches indiquent le trajet de la parole dans le système de communication (voir figure précédente). A gauche se trouvent les domaines spécifiques au modèle (méthodes quantitatives transposables à d'autres objets d'étude), à droite les domaines spécifiques aux données (liés aux sciences de la parole).

mathématiques et informatiques sont des méthodes fécondes pour l'analyse de la parole. La combinaison de la théorie de l'information, de l'apprentissage statistique et du traitement du signal permet d'aborder le problème de la modélisation de la parole avec un haut niveau d'abstraction. Les algorithmes d'apprentissage machine sont à l'origine de plusieurs percées récentes en neurosciences computationnelles et dans la conception de systèmes artificiels, devenant un outil indispensable pour la modélisation de signaux complexes. Le contexte théorique spécifique à ce travail est présenté dans le chapitre 1 (*chap. 1 et 2* en version anglaise). Cependant, si l'on veut avoir une certaine intuition sur le « code de la parole », les domaines traditionnels, qui rassemblent une connaissance approfondie de la parole (par exemple, la phonétique acoustique) sont toujours pertinents. Cette connaissance inclut également la façon dont la parole est analysée par le système auditif puis le cerveau (codage auditif, psycholinguistique).

Structure statistique fine de la parole et phonétique acoustique. (chapitre 2, *chap. 3 et 4* en version anglaise) Le signal vocal possède une structure riche, même sur des échelles de temps courtes. Cette variabilité de structure reflète le nombre important de phones [2]

2. Les *phones* représentent les segments élémentaires de la parole. Ils permettent de catégoriser les sons de la parole en fonction de leurs propriétés acoustiques. Les *phonèmes* sont les catégories élémentaires de sons de parole qui partagent la même signification linguistique. Les phonèmes sont dépendants d'une langue. Les différentes réalisations acoustiques d'un phonème sont appelées les *allophones* : ces derniers peuvent différer par leurs propriétés acoustiques mais ils sont interprétés de la même façon par les locuteurs de la langue. Ex : [c] comme dans *qui* et [k] comme dans *cou* sont des allophones pour le phonème /k/ en français. Comme cette distinction prête souvent à confusion

associés à un langage, différents types d'articulation, différentes sources sonores (vibration des cordes vocales, sons turbulents...), et ainsi une diversité de facteurs acoustiques. La principale difficulté dans l'étude du codage efficace de la parole est d'obtenir une description cohérente et synthétique de la structure statistique qui englobe cette diversité des propriétés acoustiques. J'appelle une description de la structure statistique dont le niveau de détail va jusqu'au niveau des phonèmes (ou même au niveau des évènements acoustiques, ex : relâchement d'occlusion pour les plosives) la *structure statistique à grain fin de la parole*. Cette description est similaire à la tâche de classer les sons de la parole d'après leurs propriétés acoustiques. Par conséquent, un domaine qui s'est avéré pertinent pour l'analyse menée au cours de cette thèse est la phonétique acoustique,[152, 83] qui décrit les aspects acoustiques et les propriétés du signal des sons de la parole. Je montrerai que certaines variations dans la structure statistique de la parole peuvent s'expliquer par certaines propriétés acoustiques, et dans certains cas même à des propriétés physiques du système vocal lors de la production (par exemple, l'ouverture au niveau des lèvres a une incidence sur la bande passante des formants).

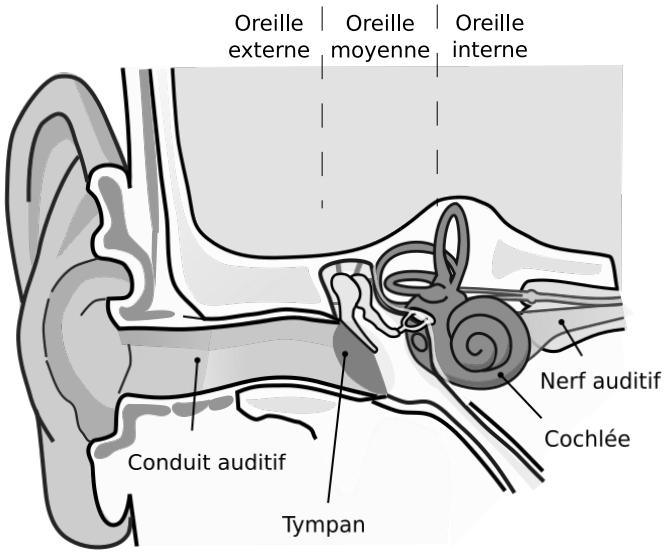


FIGURE 3 – Schéma de l'oreille humaine. Les sons (ondes acoustiques) parcourent le conduit auditif (oreille externe) et font vibrer le tympan. L'énergie mécanique est conduite à la cochlée par les os de l'oreille interne (malleus, enclume, étrier). L'oreille interne (cochlée + nerf auditif) décompose le signal en différents canaux, se comportant comme un analyseur fréquentiel du son. *Adapté de Wikimedia Commons.*

Représentations temps-fréquences non linéaires et modèles de codage auditif. (chapitre 3, *chap. 5 en version anglaise*) Par ailleurs, la description de la structure statistique de la parole peut être utilisée pour analyser certaines propriétés du codage auditif, pour voir si ces propriétés sont adaptées pour le codage de la parole. Au niveau physiologique, la décomposition temps-fréquence du signal est réalisée à la périphérie du système auditif par les cellules ciliées internes le long de la cochlée. Les mesures électrophysiologiques de l'activité des fibres nerveuses auditives chez des mammifères (par exemple chez les chats) permettent de connaître les formes des filtres auditifs et leur sélectivité de fréquence

et que les données phonétiques ne permettent pas toujours d'aller jusqu'au niveau des phones, les termes *phone* et *phonème* seront la plupart du temps interchangeables dans le reste du texte.

en fonction de la fréquence. Lewicki a démontré que les filtres obtenus avec l'analyse en composantes indépendantes (ACI), qui permet d'extraire des filtres optimaux selon certaines propriétés statistiques du signal, sont similaires aux filtres cochléaires, [94] mais nous ne savons pas si cette comparaison tient toujours avec une analyse statistique à un niveau plus fin, en relâchant la contrainte linéaire de la représentation du signal. Or, on sait que la sélectivité de la fréquence cochléaire diminue avec le niveau d'intensité sonore des filtres cochléaires [31], ce qui signifie que la décomposition cochléaire est en réalité non linéaire. La force de cette non-linéarité augmente avec la fréquence [171, 124, 166]. Les mesures électrophysiologiques de l'activité des fibres nerveuses auditives ont servi de base à des modèles linéaires ou non linéaires du filtrage cochléaire [31, 171]. Des mesures psychophysiques, plus faciles à réaliser chez l'homme, ont également été exploitées par d'autres modèles [100, 80, 141]. Dans le chapitre 3 (*chap. 5 en version anglaise*), je m'interroge sur l'adéquation possible entre le comportement non linéaire de la cochlée et la structure statistique à grain fin de la parole.

Bases théoriques

Cette section introduit le contexte théorique de cette thèse. Il est développé de façon plus formelle dans le chapitre 1 (*chap. 1 et 2 en version anglaise*).

Hypothèse du codage efficace

Une question centrale dans l'étude des systèmes sensoriels est de comprendre comment l'activité neuronale reflète l'information de l'environnement extérieur et comment les processus neuronaux sont organisés pour transmettre l'information. Les neurosciences computationnelles recherchent des principes mathématiques ou computationnels qui régissent l'organisation des processus neuronaux. L'idée que le cerveau cherche à maximiser la performance des processus neuronaux selon un ou des critères, qui ont un sens dans la théorie de l'information, est l'hypothèse du codage efficace ; c'est le point de départ de cette thèse. Cette hypothèse affirme que les systèmes sensoriels utilisent de manière optimale les ressources de codage, limitées, en s'adaptant au mieux aux statistiques des stimuli naturels. Les stimuli naturels sont les stimuli que l'individu ou l'animal rencontre dans son environnement naturel. L'introduction de l'hypothèse de codage efficace est attribuée à Horace Barlow, qui a présenté en 1961 la réduction de la redondance comme un principe plausible sous-tendant la transmission de l'information dans les systèmes sensoriels [16]. En vérité, l'hypothèse du codage efficace est à la base de travaux antérieurs, en particulier les travaux de Fred Attneave sur la perception visuelle dans les années 1950 [12]. L'hypothèse du codage efficace a donné lieu à de nombreuses études sur les propriétés des systèmes sensoriels, en lien avec les statistiques de signaux naturels, dans les années 1990 et au début des années 2000, avec l'émergence de nouveaux algorithmes dont l'Analyse en Composantes Indépendantes (ACI). Elle est notamment liée à des avancées significatives dans la compréhension du système visuel et de ses attributs : sensibilité aux contrastes, profils des champs récepteurs des neurones dans le cortex visuel primaire V1, le codage des couleurs, la sensibilité au mouvement, etc. [145, 154] Elle est aussi la base d'études comparables sur le système auditif, que ce soit sur le système périphérique [94, 91, 149, 112] – qui est l'objet de cette thèse – ou sur des aspect de plus haut niveau. Récemment, l'essentiel des efforts de recherches s'est concentré sur le traitement de haut niveau, en particulier l'étude

des filtres de modulation (détection de modulations d'amplitude et/ou de fréquence ou de motifs potentiellement plus complexes), en comparaison avec l'activité du colliculus inférieur ou du cortex auditif [92, 137, 30, 113].

Pour le système auditif, les stimuli naturels sont la parole et d'autres sons environnementaux (vocalisations animales ou sons provenant de sources non vivantes). Les sons naturels n'incluent pas seulement la parole, mais je considérerai essentiellement la parole comme l'entrée la plus pertinente pour le système auditif humain, bien que certains raisonnements sur les relations entre la structure statistique de la parole et les caractéristiques acoustiques pourraient être également pertinents pour d'autres sons naturels. Sauf lorsque cela est précisé, les sons de parole, phones et phonèmes, etc. qui sont le support des analyses dans cette thèse, proviennent tous de l'anglais américain.

Codes à entropie maximale et minimale. Dans la théorie du codage efficace, la performance du codage est liée à une mesure de la "taille" du code neuronal, qui à son tour est liée à la quantité d'information (ou *entropie*), concept clé de la théorie de Shannon. Il y a deux points de vue apparemment opposés sur l'efficacité de codage. Le premier point de vue, très proche de la proposition initiale de Barlow de réduction de la redondance [16], est que les opérations neuronales exploitent à fond la capacité de codage en maximisant la quantité d'information transmise par unité de temps. L'objectif des systèmes sensoriels serait alors de réaliser un code à entropie maximale des entrées sensorielles sous une contrainte de ressources (ce critère est appelé *maximisation de l'information* ou *infomax* [21, 99]). Le deuxième point de vue est que les opérations neuronales préfèrent s'appuyer sur des codes compacts afin d'économiser les ressources d'énergie dont les neurones disposent. En d'autres termes, le but des systèmes sensoriels serait d'atteindre des codes à entropie minimale [17]. En réalité, les deux points de vue sont similaires. De manière informelle, on peut dire que les codes à entropie maximale cherchent à *casser* toute structure dans les données afin que le processus de sortie soit aussi aléatoire que possible (par exemple, associé à une distribution uniforme). Les codes à entropie minimale cherchent à tirer parti de la régularité des données et à *conserver* cette structure tout au long du processus pour obtenir des codes compacts. L'idée commune est que la représentation neurale sous-jacente doit refléter la structure des données. Trouver des codes à entropie minimale peut être considéré comme la première étape d'un processus qui vise à maximiser le transfert d'information. La relation entre les deux critères est expliquée plus rigoureusement au chapitre 1. Les deux critères sont décrits dans cette thèse, mais je me référerai généralement à la notion de code à entropie minimale, car l'accent est mis sur la toute première étape du traitement sensoriel, dont l'objectif est de trouver une représentation qui capture la structure statistique de la parole.

La notion de structure statistique

La structure statistique est ce qui distingue un signal du bruit. Elle est définie par opposition à la notion d'entropie maximale. Un processus d'entropie maximale n'a pas de régularité quelconque : il est totalement aléatoire et ne peut être exploité spécifiquement d'aucune façon. Selon la contrainte exercée sur le processus de génération, les distributions associées à un processus d'entropie maximale peuvent être –dans les cas les plus courants : uniformes (valeurs bornées), gaussiennes (écart-type fixe) ou exponentielles (moyenne positive, moyenne fixe). Les processus gaussiens sont les exemples typiques de données générées qui n'ont pas de structure : les algorithmes recherchant une structure statistique

échouent systématiquement avec des données gaussiennes (bruit blanc gaussien). Toute régularité imposée aux données – en d'autres termes, toute structure ou *redondance* – diminue l'entropie du signal. Un exemple de redondance est lorsque certaines valeurs deviennent plus fréquentes que dans un processus totalement aléatoire (fig. 4, fig. 5).

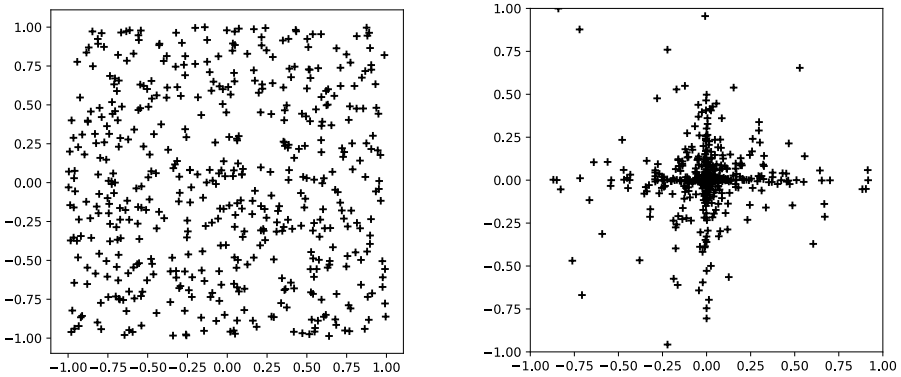


FIGURE 4 – Différence entre des points générés selon une distribution uniforme (pas de structure, à gauche) et des points générés selon une distribution présentant un pic (à droite). Pour la figure de droite, les valeurs proches de 0 sont plus typiques. A droite, les distributions marginales (projections sur les axes x et y) vérifient $\log p(x) \propto -1.8|x|$.

Indépendance et représentations parcimonieuses. Comment la notion de structure statistique se traduit-elle en propriétés statistiques plus concrètes ? Une première réponse apportée par la théorie du codage efficace est que les stimuli naturels devraient être représentés avec un ensemble de caractéristiques (*features*) aussi indépendantes que possible [10, 78]. La contrainte d'indépendance est étroitement liée aux critères théoriques proposés par Barlow (réduction de redondance [16] et codage à entropie minimale [17]). Cela correspond à l'idée que deux canaux ne devraient pas consommer inutilement des ressources en codant pour la même information. Le raisonnement est que si l'information du signal d'entrée est conservée tout au long de la transformation, mais que la quantité d'information devant être codée par chaque canal est minimale, alors la solution consiste à réduire l'information partagée en commun par différents canaux (information mutuelle). Des caractéristiques indépendantes sont obtenues en réduisant l'information mutuelle entre les canaux, ce qui peut être considéré comme un type de redondance associé au processus entier. Une décomposition du signal qui atteint la contrainte d'indépendance a parfois été appelée code factoriel [115], puisque alors les composantes représentent des caractéristiques qui ne sont pas liées les unes aux autres. On constate qu'un code factoriel, en plus d'être efficace, présente des avantages sur le plan de la représentation (les tâches de classification sont plus faciles étant donné des caractéristiques indépendantes). Cependant, bien que ce critère soit attrayant, la contrainte d'indépendance est très forte et difficile à exprimer par des mesures statistiques empiriques. Une représentation avec des caractéristiques indépendantes est également rarement réalisable dans la pratique : pour la parole en particulier, toutes les composantes de fréquence sont typiquement excitées au même moment (par exemple lors des excitations glottales), ce qui contredit toute propriété d'indépendance. Il est toujours judicieux de rechercher des caractéristiques indépendantes dans la pratique car la représentation obtenue conserve certains avantages de la représentation idéale. Les algorithmes s'appuient sur certaines des informations *a priori* sur les distributions de

probabilité marginales, pour obtenir des critères plus faibles d'indépendance, mais aussi plus faciles à mettre en oeuvre.

Une autre réponse pour trouver des représentations pertinentes en accord avec la structure statistique est donnée par la notion de *parcimonie* [121]. L'hypothèse selon laquelle les réseaux neuronaux réagissent aux stimuli naturels en minimisant l'activité des neurones (plus particulièrement en entraînant un faible nombre d'impulsions électriques), afin notamment d'économiser les ressources métaboliques, est appelée l'hypothèse du codage parcimonieux. La parcimonie (*sparsity* en anglais) est une notion intuitive et explicite pour le codage à entropie minimale : un code compact signifie ici simplement qu'un petit nombre d'activations des unités de codage est nécessaire pour décrire le signal. La commodité de cette notion ne doit pas conduire à confondre la quantité d'information avec le nombre d'unités activées : en fait, cette approximation ne peut être justifiée que lorsque le code est effectivement parcimonieux. Les données sensorielles satisfont généralement cette hypothèse de parcimonie. Le codage parcimonieux est semblable à la recherche de composantes indépendantes lorsque les composantes sont associées à des activations éparses. En fait, la parcimonie est souvent utilisée comme contrainte sur les composantes pour la recherche de composantes indépendantes [67], de sorte que les deux critères ne sont pas si clairement séparés en pratique. Une caractéristique commune aux deux méthodes est qu'elles capturent des régularités d'ordre supérieur (>2) dans les données – par exemple, certaines méthodes reviennent à trouver les directions de kurtosis maximum [78, 121]. Cela contraste avec de nombreuses autres méthodes classiques d'analyse statistique des signaux, qui s'appuient sur les moments d'ordre 2 (par exemple, la caractérisation du spectre de puissance spectrale de la parole en 1/f [168]).

La recherche de caractéristiques indépendantes ou intervenant de manière parcimonieuse dans la décomposition du signal est l'objectif de plusieurs méthodes et algorithmes, parmi lesquels l'Analyse en Composantes Indépendantes (ACI) [78] et les méthodes de codage parcimonieux avec des dictionnaires de caractéristiques [172]. Ces méthodes sont présentées dans le chapitre 1.

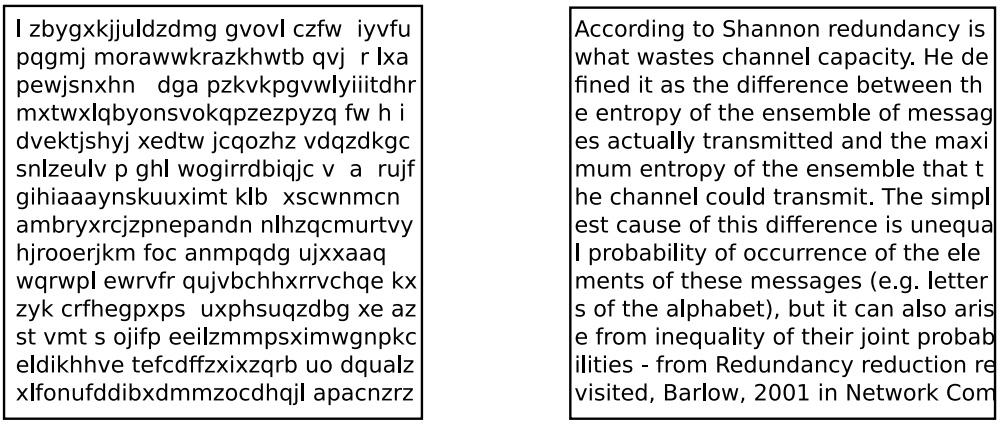


FIGURE 5 – Un autre exemple de données textuelles présentant ou non une structure. A gauche : texte généré aléatoirement (non structuré). A droite : texte en anglais (structuré). Le langage naturel a été utilisé par Shannon puis Barlow comme exemple de données présentant de nombreuses redondances : certains caractères (par exemple, 'e' ou'a') apparaissent plus fréquemment que d'autres (par exemple,'z').

Structure statistique et extractions de traits caractéristiques. Si l'on oublie le contexte de la théorie du codage efficace, les méthodes qui y sont rattachées (ACI et méthodes de codage parcimonieux) sont simplement des techniques d'extraction de caractéristiques, non supervisées. Ce sont ainsi des méthodes parmi d'autres pour la recherche d'une représentation pertinente du signal. Ce paragraphe décrit comment l'ACI et les méthodes de codage parcimonieux sont reliées à d'autres méthodes d'extraction de traits caractéristiques. On peut distinguer deux types de méthodes pour l'extraction de traits caractéristiques :

— les méthodes *a priori* qui se basent sur des modèles construits spécifiquement pour la parole (*model-based*) : les traits caractéristiques sont construits « sur mesure » en tirant parti de nos connaissances du signal de la parole ou du système auditif. Elles sont souvent le produit d'un travail d'ingénierie minutieux. Il en existe un grand nombre en fonction des applications (voir ref. [62] parties V&VI). On peut citer en exemple les traits caractéristiques souvent utilisés en reconnaissance de la parole qui sont les coefficients du cepstre mel-fréquence (MFCC). Ce sont déjà des traits d'assez haut niveau. Le calcul de ces coefficients est schématiquement celui-ci : la transformée de Fourier rapide est réalisée sur un segment du signal. Puis la puissance spectrale est sommée sur des bandes de fréquences calquées sur une échelle perceptive de la hauteur (échelle *mel*). Cette puissance est exprimée en décibel (échelle *log*) puis une transformation de Fourier inverse est appliquée. Comme exemples plus simples rentrent dans cette catégorie les techniques de spectrogrammes basées sur la transformée de Fourier à court-terme et les cochléogrammes basées sur des banques de filtres inspirés de la physiologie de l'oreille.

— les méthodes qui se basent sur les données (*data-based*) : les traits caractéristiques sont appris des données. Différentes représentations sont possibles en fonction de l'algorithme d'apprentissage. La méthode la plus commune est l'Analyse en Composantes principales (ACP), qui cherche à trouver les dimensions qui capturent la plus grande variabilité des données. L'Analyse en Composantes indépendantes (ACI) fait partie de cette catégorie et correspond à une contrainte plus forte sur les dimensions : l'indépendance statistique. Les réseaux de neurones profonds appartiennent aussi à cette catégorie, néanmoins ils se basent parfois sur des représentations transformées du signal de la parole (ex : spectrogramme). De plus, l'architecture imposée à ces modèles est équivalente à ajouter de l'information a priori sur les données [33]. Cependant les filtres sont parfois appris sur des données brutes (forme d'onde) pour la reconnaissance automatique [161].

Ces dernières années, nous avons assisté à un changement de paradigme avec un déplacement progressif des méthodes basées sur la modélisation de la parole vers des méthodes basées sur les données. La raison est que les techniques d'apprentissage statistique couplées avec des ressources computationnelles/données plus faciles d'accès offrent un pouvoir de modélisation plus important. Les méthodes construites « à la main » souffrent du biais de l'ingénieur qui n'a sans doute pas toutes les connaissances afin de déterminer la représentation optimale. Les méthodes d'apprentissage en grande dimension utilisent des algorithmes de descente de gradient stochastique qui affinent la représentation sans rigidité imposée par un modèle. L'illustration de cette tendance est l'apprentissage dit « de bout en bout » (*end-to-end*). Ce terme est apparu ces dernières années dans la communauté des chercheurs en reconnaissance automatique de la parole : ce terme indique que la tâche de reconnaissance automatique de la parole pourrait finalement être menée de bout en bout du signal brut au texte transcrit sans aucune étape de modélisation explicite intermédiaire.[65] Néanmoins, il s'agit sans doute là d'une position extrême, puisque ce serait affirmer qu'un

système efficace de codage de la parole ne nécessite pas de spécificités algorithmiques afin de s'adapter au mieux au signal de la parole. Cette thèse adopte une approche moins extrême, car nous recherchons des principes de calcul spécifiques à l'extraction des caractéristiques de la parole, qui ne sont pas implémentés dans un modèle générique, bien que l'étude soit dirigée par l'analyse statistique des données. L'accent est mis sur l'extraction de caractéristiques de bas niveau. L'Analyse en Composantes Indépendantes (ACI) cherche une décomposition linéaire associée à un code optimal. Elle correspond à la première étape d'un système hiérarchique. Ceci est à comparer avec la première couche d'un réseau de neurone profond et avec les techniques usuelles d'analyse temps-fréquence (transformée de Fourier à court terme, décomposition en ondelettes, etc.). Le traitement non linéaire est abordé (chap. 3, chap. 5 en version anglaise), mais le traitement hiérarchique hautement non linéaire des réseaux de neurones profonds n'est pas l'objet d'étude.

Analyse temps-fréquence

Concerant la parole, les informations importantes (par exemple la structure formantique) est plus visible sur des représentations temps-fréquence que sur la forme d'onde brute (fig. 6). L'analyse en Composantes Indépendantes (ACI) appliquée à la parole produit une famille de filtres, qui reproduit une analyse temps-fréquence analogue à la décomposition des signaux de parole par l'oreille interne. Le champ de l'analyse temps-fréquence des signaux permet d'apporter des éléments pour comprendre les principales propriétés de la représentation apprise.

Compromis temps-fréquence. L'information de fréquence est obtenue par intégration dans le temps, opération qui réduit nécessairement la résolution temporelle. Ce phénomène est à l'origine du principe d'incertitude selon lequel un filtre ne peut être précis en fréquence qu'au détriment de la précision en temps, et réciproquement [54, 134]. Le principe d'incertitude limite toutes les représentations temps-fréquence, et il faut en général choisir entre une bonne résolution temporelle et des fenêtres d'intégration courtes, ou une bonne résolution fréquentielle et des fenêtres d'intégration plus larges. Ce choix a un impact sur la structure révélée par la représentation. Pour la parole, la structure des formants et les instants précis des relâchements d'occlusion pour les plosives (ex : p) sont plus visibles sur les spectrogrammes à large bande (fig. 6), favorisant la précision temporelle par rapport à la précision fréquentielle. Au contraire, les spectrogrammes à bande étroite, qui préfèrent la précision fréquentielle, font apparaître la structure harmonique de manière plus visible, bien que l'harmonicité soit encore visible sur les spectrogrammes à large bande comme des répétitions du même motif (excitations glottales) à intervalles réguliers dans le temps. Pour déterminer quelle représentation temps-fréquence décrit le mieux la structure statistique de la parole, il faut identifier le compromis temps-fréquence le plus approprié pour les sons de la parole. Le principe d'incertitude révèle également l'importance des filtres de Gabor (ou ondelettes de Gabor). Les fltres de Gabor sont des sinusoïdes modulés par des fenêtres gaussiennes : ils permettent d'obtenir le meilleur compromis de résolution temps-fréquence. Le chapitre 2 (en version anglaise) rappelle ce résultat classique et mentionne une autre version du principe d'incertitude qui démontre également l'importance des filtres de Gabor dans le contexte du codage parcimonieux.

Sélectivité fréquentielle et facteur de qualité Q_{10} Une propriété fondamentale d'un filtre est donnée par sa sélectivité fréquentielle. La sélectivité fréquentielle décrit la largeur

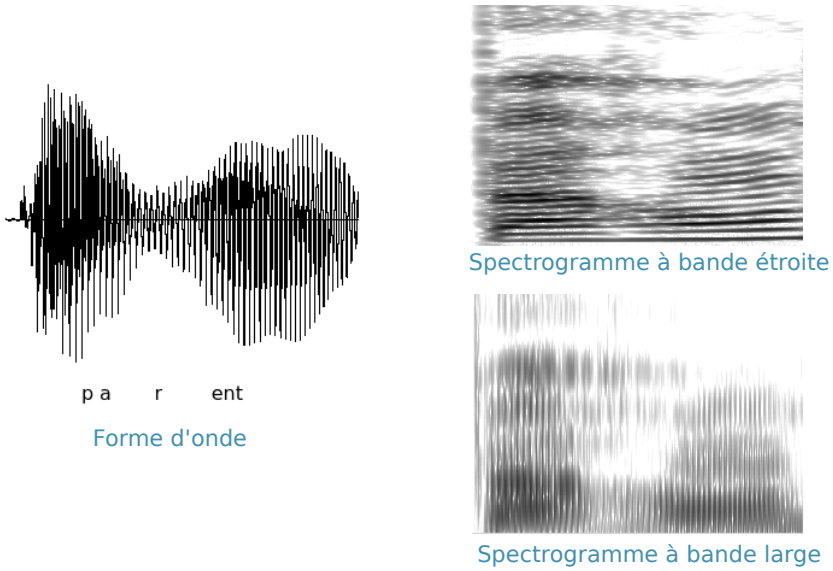


FIGURE 6 – Toutes les représentations du signal de la parole ne sont pas équivalentes. Cette illustration montre trois exemples de représentation pour la même occurrence du mot 'parent'. Les représentations temps-fréquence (spectrogrammes à bande étroite ou large) rendent directement visibles les informations fréquentielles, à la différence de la forme d'onde brute (à gauche), par exemple la structure formantique. L'harmonicité du signal apparaît de façon différente pour les deux spectrogrammes : on peut voir les harmoniques sur le spectrogramme à bande étroite, et la répétition des excitations glottales à intervalles réguliers pour le spectrogramme à large bande. Réalisé avec WAVESURFER [147]. Spectrogrammes : les zones sombres sont les zones d'intensité spectrale maximale (temps en abscisse, fréquence en ordonnée).

de l'intervalle des fréquences pour laquelle le filtre a une réponse élevée. Elle peut être quantifiée par le facteur qualité $Q_{10\ \mathrm{dB}}$ (en abrégé Q_{10} dans le reste de la thèse), défini par la fréquence centrale f_c divisée par la bande passante à 10 dB $\Delta f_{10\ \mathrm{dB}}$ (fig. 7). Les facteurs de qualité Q_{10} sont souvent employés en audition car les filtres auditifs ont une sélectivité fréquentielle relativement faible. Cette préférence est assez singulière, le facteur de qualité $Q_{3\ \mathrm{dB}}$ étant le plus courant pour la majorité des applications en dehors de l'audition. La bande passante définie à 3dB permet de caractériser des pics de fréquence plus étroits. Ceux-ci sont par exemple utilisés en acoustique (ex : caractérisation des résonances du conduit vocal). Pour une forme de filtre contrôlée par un seul paramètre, comme dans le cas des filtres de Gabor, la connaissance de Q_{10} est suffisante pour déterminer la largeur de la fonction d'onde (en temps) et la bande passante du filtre. Un ensemble de filtres de Gabor uniformément répartis en temps et fréquence peut alors être caractérisé par le comportement de Q_{10} en fonction de la fréquence : $Q_{10} = \mathrm{f}(f_c)$.

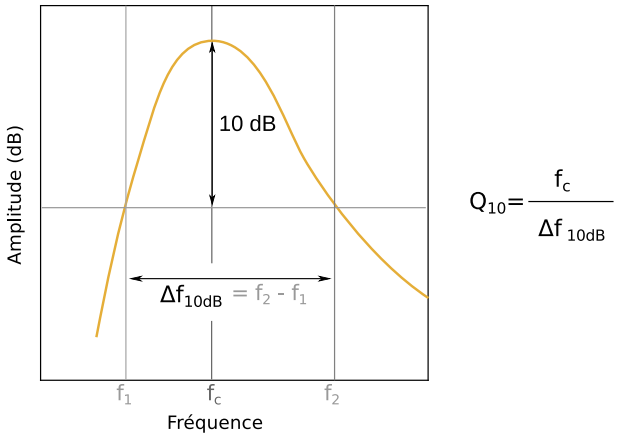


FIGURE 7 – La largeur de bande à 10 dB d'un filtre est la largeur de l'intervalle des fréquences dont la réponse en amplitude est supérieure à -10 dB par rapport à la réponse maximale. La fréquence centrale f_c est la fréquence pour laquelle le filtre a une réponse maximale. Le facteur qualité Q_{10} est défini par le rapport entre la fréquence centrale et la largeur de bande à 10 dB.

Analyse multi-résolution. Dans cette thèse, je considère les représentations temps-fréquence pour lesquelles le facteur de qualité suit une loi de puissance par rapport à la fréquence centrale :

$$Q_{10}(f) = Q_0 \left(\frac{f}{f_0} \right)^{\beta} . \tag{1}$$

Les valeurs des constantes que j'utilise sont $f_0 = 1.0kHz$ et $Q_0 = 2.0$. Le choix de cette famille de représentations est motivé empiriquement puisqu'elle se rapproche des représentations apprises par l'ACI appliquée aux données de parole (voir paragraphe suivant). Le **paramètre** β, l'exposant de la loi de puissance joue un rôle central dans cette thèse. β est également la pente de la droite définie par Q_{10} en fonction de f_c sur une échelle logarithmique (fig 8). Il y a deux interprétations sur le contrôle du paramètre β sur la représentation :

1. β contrôle le compromis temps-fréquence dans les hautes fréquences : les représentations sont toutes les mêmes à $f_c = 1kHz$ mais ensuite le facteur de qualité

augmente comme f_c^β. Dans les hautes fréquences, les filtres sont localisés en temps et peu sélectifs en fréquence pour des valeurs β faibles, et inversement pour des valeurs plus élevées.

2. β permet de séparer les décompositions à résolution unique des décompositions multi-résolution. Dans une décomposition à résolution constante, les filtres sont associés à une largeur de bande caractéristique unique et les réponses impulsionnelles ont le même temps caractéristique. Dans une décomposition multi-résolution, la largeur de bande des filtres est proportionnelle à la fréquence centrale et les réponses impulsionnelles ont des temps caractéristiques qui sont inversement proportionnels à la fréquence centrale. Des exemples de décomposition à résolution unique ($\beta = 1$) sont l'analyse classique de Gabor [68] et la transformée de Fourier à fenêtre glissante. Des exemples de décompositions multi-résolution ($\beta = 0$) sont la transformée à Q constant, ou la transformée en ondelettes classique [105].

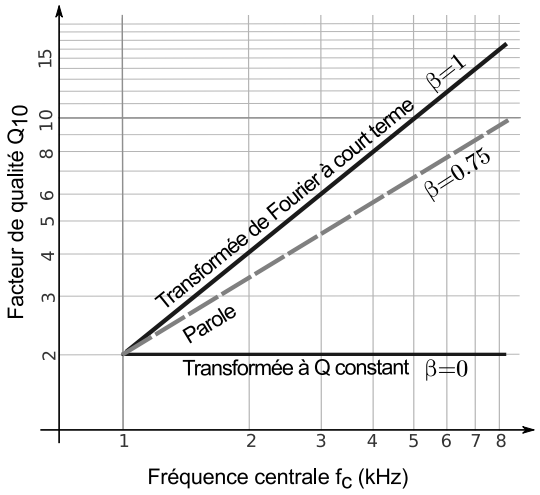


FIGURE 8 – Le paramètre β est la pente de regression du facteur de qualité Q_{10} en fonction de la fréquence centrale f_c (sur une échelle logarithmique). $\beta = 1$ caractérise les décompositions à résolution unique (transformée de Fourier à court terme, aussi appelée transformée de Fourier à fenêtre glissante), tandis que $\beta = 0$ caractérise les décompositions multi-résolution (transformée à Q constant, ou transformée en ondelettes). La décomposition la plus parcimonieuse de la parole est obtenue avec $\beta = 0.75$ (valeur basée sur les analyses antérieures, voir paragraphe suivant, et les analyses statistiques présentés dans la thèse).

Travaux antérieurs

Structure statistique de la parole dans son ensemble. Les premières recherches sur la structure statistique de la parole basée sur l'Analyse en Composantes Indépendantes (ACI) remonte à 2000-2001 [84, 3]. En 2002, Lewicki a montré que l'ACI appliquée à des fragments de signaux de parole sur 8 ms produit un ensemble de filtres similaires aux ondelettes de Gabor, reproduisant une décomposition temps-fréquence des signaux de la parole (fig. 9) [94]. De façon plus frappante, il a montré que la sélectivité fréquentielle de ces filtres suit la même loi de puissance, dans les hautes fréquences 1-8kHz, que la sélectivité

fréquentielle des filtres cochléaires chez les mammifères (bien que légèrement supérieure, $\beta = 0.7 - 0.8$ pour l'ACI comparé à $\beta = 0.6$ pour les profils de réponse des fibres auditives chez les chats [133, 111]). Ce résultat est une réplication dans le domaine de l'audition d'un résultat connu dans le domaine de la vision : l'ACI ou des algorithmes de codage parcimonieux produisent des filtres orientés en forme d'ondelettes de Gabor similaires aux profils réceptifs dans le cortex visuel primaire [120, 164]. La parole, cependant, a la particularité d'être un stimulus contrôlé par l'homme, même si elle est soumise à des contraintes physiques et acoustiques. La spécificité de la sélectivité fréquentielle de la cochlée humaine fait toujours débat [107], en particulier en réponse à des faibles niveaux d'intensité [124], mais il est généralement admis que la cochlée humaine n'est pas très différente d'autres mammifères non spécialisés (comme le chat) concernant la sélectivité fréquentielle. Comme la parole a émergé récemment relativement à l'évolution de la cochlée, Lewicki a proposé l'hypothèse que la parole a évolué pour être codée de façon optimale par le système auditif des mammifères. Il a également suggéré qu'une explication pour une valeur $\beta = 0.6$ est l'équilibre que l'on trouve dans les sons de parole entre les sons transients et des sons stationnaires. En utilisant à la place des données de parole un mélange de sons environnementaux et de vocalisations animales, on trouve ainsi le même accord avec les données physiologiques [94].

Ces premiers résultats sont conformes à l'hypothèse du codage efficace, en montrant que la structure statistique globale de la parole est adaptée aux propriétés physiologiques de l'oreille. Cependant, il est difficile d'interpréter véritablement les résultats en terme de structure du signal. La diversité des phones et phonèmes d'une langue ne permet pas d'avoir une seule interprétation de la décomposition révélée par l'ACI qui s'appliquerait à chacun des sons. De plus, il est possible que certaines régularités, qui n'apparaissent pas quand l'ACI est appliquée aux données de parole prises dans leur ensemble, existent à un niveau plus fin.

Structure statistique de la parole, divisée en catégories phonétiques. En 2013, C. Stilp et M. Lewicki ont étudié la structure statistique de la parole à un niveau phonétique plus fin [153]. Leur approche consiste à diviser les données de parole en sous-classes, rassemblant des sons partageant certaines caractéristiques acoustiques, afin d'obtenir une description qui soit basée sur des propriétés plus concrètes des signaux. Ils ont appliqué l'ACI à des catégories phonétiques assez larges (ex : fricatives, plosives, voyelles...) et ont constaté que le compromis temps-fréquence était différent selon la classe considérée. Ils ont utilisé le paramètre β, la pente de régression pour Q_{10} en fonction de f_c, pour comparer les représentations obtenues sur les différentes catégories phonétiques (fig. 10). Récemment, Ramon G. Erra et Judit Gervain ont également utilisé cette même méthode pour étudier les variations de la représentation lorsque l'ACI est appliquée à différentes langues [47].

D'après C. Stilp et M. Lewicki, les variations de la valeur du paramètre β lorsqu'on considère différentes catégories phonétiques s'expliquent par la nature transitoire des sons qui composent certaines classes. Des changements rapides dans le temps feraient passer les filtres optimaux d'une représentation fréquentielle à une représentation temporelle avec une sélectivité fréquentielle plus faible : par exemple, les plosives sont associées à la valeur la plus faible de β. Ce point de vue, cependant, n'explique pas entièrement, a priori, pourquoi les voyelles aboutissent à une représentation plus localisée en temps que les fricatives, par exemple. La dispersion des valeurs pour β lorsque l'ACI est appliquée à des sous-classes de parole pourrait signifier que le système auditif exploite cette structure à grain fin pour mieux s'adapter aux statistiques de la parole. Stilp et Lewicki ont ainsi

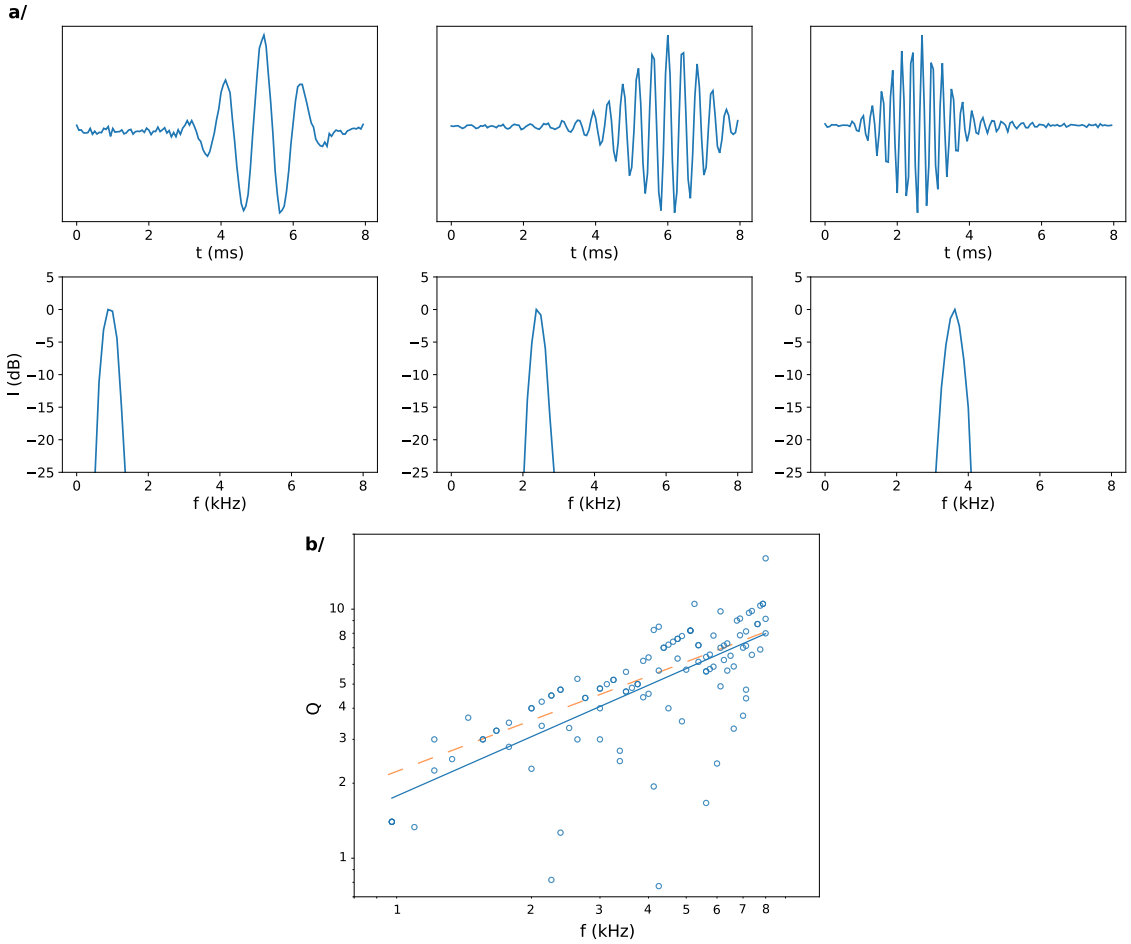


FIGURE 9 – L'ACI appliquée à des fragments de signaux de parole de 8 ms produit un ensemble de filtres ressemblant à des ondelettes de Gabor. *a/* Exemples de filtres appris avec ACI (anglais). En haut : réponses temporelles. En bas : réponses fréquentielles, en dB. *b/ :* Q_{10} en fonction de la fréquence (échelle log-log). Cercles : Q_{10} pour les filtres appris par l'ACI (droite en bleu : régression linéaire pour ces points). Ligne en pointillés : régression linéaire pour les données physiologiques. Voir aussi réf. [94] ou réf. [47] pour des figures similaires.

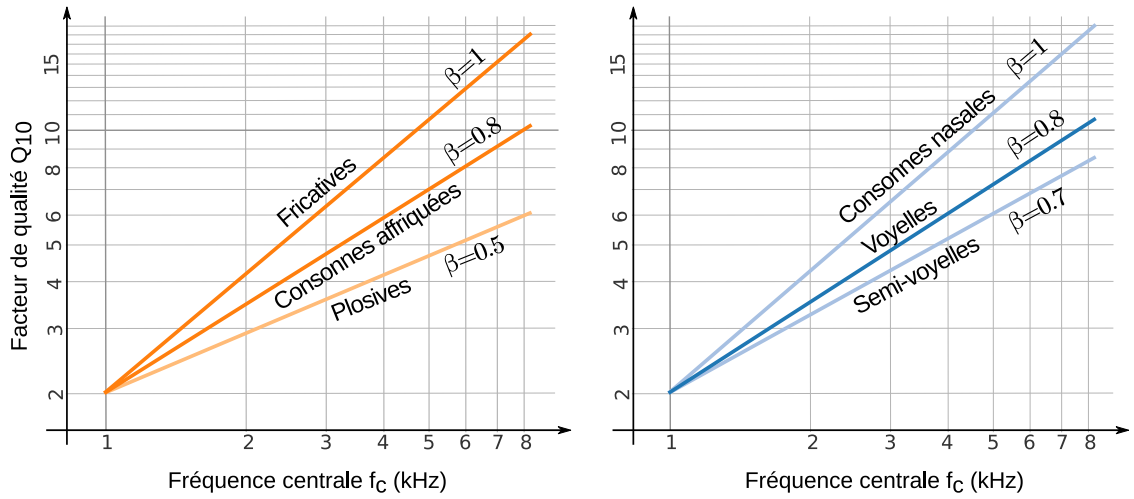


FIGURE 10 – L'ACI appliquée à différentes catégories phonétiques (anglais américain) montre que la variation de l'exposant β permet d'adapter la décomposition à différents types de sons de la parole. *A gauche :* Pentes de régression pour les consonnes (plosives, consonnes affriquées, fricatives). *A droite* : Pentes de régression pour les semi-voyelles (glides), voyelles et consonnes nasales. *Adapté de Stilp et Lewicki, 2013 [153].*

suggéré que la distribution des valeurs est congruente avec la diversité des propriétés de sélectivité fréquentielle que l'on trouve pour les réponses caractéristiques des neurones du noyau cochléaire.

Objectifs et structure de la thèse

L'analyse de la structure statistique de la parole, par C. Stilp et M. Lewicki, basée sur des catégories phonétiques, soulève plusieurs nouvelles questions :

— Pourquoi obtient-on des valeurs différentes pour l'exposant β, contrôlant la sélectivité fréquentielle dans les hautes fréquences ? Une question plus précise est la suivante : quelles caractéristiques du signal, ou caractéristiques acoustiques, permettent d'expliquer les variations et l'intervalle de valeurs prises par β ?

— Quelle est la division de la parole la plus significative pour la structure statistique de la parole ? Nous ne savons pas si les catégories phonétiques prédéfinies telles que employées par Stilp et Lewicki sont les plus pertinentes pour la structure du signal. Nous pourrions trouver une segmentation plus pertinente si des régularités étaient recherchées à un niveau plus fin, au niveau des phonèmes, ou même à un niveau plus fin, car des changements temporels de structure peuvent se produire même au sein d'unités phonémiques.

— Existe-t-il des régularités à un niveau plus fin qui peuvent être exploitées par un système de codage efficace ? Une tendance suffisamment significative qui apparaîtrait dans la description à grain fin de la structure statistique de la parole pourrait être exploitée par des stratégies de codage avancées, en permettant une représentation des sons de la parole qui s'ajuste à l'entrée. La stratégie doit être suffisamment simple et robuste : une représentation non-linéaire qui essaie de s'adapter à un niveau trop fin, à chaque son qui est présenté, résulterait en une stratégie trop complexe et irréalisable. Il y a un compromis à trouver entre adapter la représentation à la

structure statistique à un niveau fin et pouvoir proposer une représentation qui soit suffisamment générale.

— Si la structure statistique fine de la parole peut effectivement être exploitée par des systèmes de codage, de façon réaliste et efficace, ces stratégies sont-elles mises en œuvre dans le système auditif ?

Cette thèse vise à apporter des éléments de réponses aux questions ci-dessus. L'objectif principal est de décrire la *structure statistique à grain fin de la parole*, en caractérisant comment le compromis temps-fréquence optimal varie à un niveau de détail fin des sons de la parole, regroupés selon leur propriétés phonétiques. Cette description doit rendre explicites les caractéristiques acoustiques qui permettent de régler la valeur de β. Un objectif secondaire est de décrire comment la structure à grain fin de la parole pourrait être exploitée par des systèmes de codage efficaces. En particulier, je prétends qu'une stratégie efficace consiste à faire varier la sélectivité fréquentielle en fonction du niveau d'intensité sonore, d'une manière compatible avec le comportement non linéaire des filtres cochléaires. L'approche spécifique de la thèse repose sur un modèle de représentation paramétrique exploitant le paramètre β, en combinaison avec une méthode de codage parcimonieux à l'aide de dictionnaires de filtres. Plus spécifiquement, les fragments de signaux de parole sont décomposés dans une famille de dictionnaires dont les atomes sont des filtres de Gabor. Puis, un score (fonction de coût) reflétant la parcimonie des décompositions est calculé pour chaque dictionnaire. Le dictionnaire associé à la représentation la plus compacte des données, minimisant la fonction de coût, permet de réaliser une estimation de β. La distribution des valeurs prises par β a été analysée pour diverses données de parole, divisées au niveau des phonèmes, ou même à un niveau intra-phonémique. J'ai également réalisé des analyses pour des données synthétiques dont la structure s'apparente à celles de la parole. Je montre que la distribution des valeurs de β pour différentes configurations des données d'entrée offre une interprétation riche de la structure statistique à grain fin de la parole. Les variations du paramètre peuvent être ainsi reliées à des propriétés acoustiques spécifiques, qui sont déduites de l'analyse.

La partie en français de la thèse est un résumé substantiel de la version complète en anglais. La présente introduction a été reprise intégralement à partir de sa version anglaise. Les chapitres qui suivent reprennent succinctement le corps de la thèse. La structure du texte est conservée dans son ensemble, mais certains chapitres ont été regroupés pour permettre de résumer les éléments principaux. Lorsque des éléments ne sont pas développés, la version complète est mentionnée, aussi souvent que possible.

Le plan de la thèse est le suivant :

— Le premier chapitre développe les bases théoriques de la thèse. Il présente les principes de la théorie du codage efficace, et argumente le point de vue original que j'ai adopté pour le problème du codage efficace de la parole. Il motive les méthodes d'analyse et la fonction de coût qui sont utilisées dans les chapitres suivants. En particulier, il montre le lien théorique entre les méthodes employées pour cette thèse et les études précédentes sur la structure statistique de la parole, qui étaient basées sur l'Analyse en Composantes Indépendantes (ACI). Il est présenté rapidement un lien explicite entre la parcimonie de décompositions temps-fréquence et le principe d'incertitude, davantage détaillé dans la version anglaise (*chap. 2*). Ce résultat, connu dans le domaine des représentations quadratiques temps-fréquence, a été rarement mentionné dans le contexte du codage efficace.

— Les chapitres 2 et 3 représentent la principale contribution du travail de recherche menée durant la thèse.

- Le chapitre 2 résume l'analyse développée dans deux chapitres dans la version complète (*chap. 3 et chap. 4*). D'abord, la structure statistique est étudiée pour des signaux artificiels dont on s'attend à ce qu'ils partagent la même structure que les signaux de parole réels. Cette première analyse donne un premier aperçu de la structure statistique de signaux acoustiques et révèle les facteurs acoustiques les plus importants. Puis, la structure statistique est décrite pour des données de parole réelles, basée sur l'analyse d'un corpus d'anglais américain (base de données TIMIT). Le comportement de β est analysé à chaque fois à un niveau plus fin : d'abord en regroupant les sons par grandes catégories phonétiques, puis par phonèmes, puis par parties de phonème.
- Dans le chapitre 3 (*chap. 5 en version anglaise*), je m'interroge sur la possibilité d'exploiter la structure statistique à grain fin de la parole par une stratégie de représentation des signaux non linéaire. Je montre qu'une stratégie efficace consiste à diminuer la sélectivité fréquentielle en fonction du niveau d'intensité sonore, ce qui rappelle le filtrage cochléaire non linéaire. Je discute de la façon dont l'hypothèse d'une concordance pourrait être vérifiée par d'autres recherches, tant sur le plan théorique qu'expérimental.

- La version anglaise comporte un chapitre supplémentaire (*chap. 6 en langue anglaise*) portant sur la caractérisation du rythme de la parole à partir de statistiques résumant le signal (comme l'intensité). Ce travail représente une part importante de l'effort de recherche de cette thèse, initialement lié à la thématique principale mais qui est devenu par la suite largement indépendant du reste des travaux. Par conséquent, le contexte, les méthodes et les résultats de cette étude sont présentés séparément dans ce dernier chapitre.

La principale contribution de cette thèse sur le codage efficace de la parole fait l'objet d'un article, *Fine-grained statistical structure of speech* (soumis) `<hal-01931420>`. Ce travail a également été présenté au International Symposium on Auditory and Audiological Research 2019 (ISAAR 2019, Nyborg, Danemark) et sous forme de poster à la 177e conférence de l'Acoustical Society of America (Louisville, États-Unis, 2019) `<doi:10.1121/1.5101317>`.

Chapitre 1

Hypothèse du codage efficace

L'hypothèse du codage efficace a été introduite pour la première fois par Barlow en 1961[16], lorsque celui-ci a proposé la réduction de redondance comme principe plausible régissant la transmission de l'information dans les systèmes sensoriels. Sa proposition était que le code neuronal, émergeant de l'activité d'un large nombre de neurones, s'adapte aux statistiques des stimuli naturels en réduisant la redondance des signaux neuronaux. Ce chapitre présente le formalisme de la théorie du codage efficace et les différents critères théoriques de l'information qui ont été proposés par Barlow et d'autres. Il introduit également les méthodes et algorithmes statistiques qui sont liés à la théorie du codage efficace, à savoir l'Analyse en Composantes Indépendantes (ACI) et les méthodes de codage parcimonieux.

1.1 – Critères d'efficacité

Cette section introduit différentes mesures qui ont été proposées pour quantifier l'efficacité de codage.

Entropie et quantité d'information. Les critères d'optimalité dans la théorie du codage efficace sont des quantités d'information à minimiser ou maximiser. Ces fonctions de coût sont des variantes de la formule de l'*entropie*, définie pour un processus stochastique X associée à une distribution discrète $p(x)$ par :

$$H(X) = -\mathbb{E}(\log p(X)) \ = -\sum \log\left(p(x)\right)\, p(x)\ .$$

Cette quantité correspond à la fois à la quantité d'information contenue dans le processus X et aux *ressources* nécessaires pour encoder cette information. L'entropie quantifie le caractère aléatoire ou le manque de régularité d'une source stochastique. En particulier, si X prend une valeur constante (cas où le processus est déterministe), alors l'entropie est nulle. D'après la théorie de Shannon, l'entropie est la longueur moyenne minimale d'un code qui permet d'encoder toutes les occurrences de X. En base binaire, la longueur du code par occurrence (symbole) est le nombre moyen de bits (0 ou 1) nécessaires pour encoder un symbole en provenance de la source. Pour une discrétisation de plus en plus fine du processus X et des probabilités toujours discrètes $p(x)|\Delta x|$ ($|\Delta x| \to 0$), le terme d'entropie devient :

$$H(X) = -\sum \log(p(x)\Delta x)\, p(x)|\Delta x| = -\sum \log(p(x))\, p(x)|\Delta x| \ - \log(|\Delta x|)\,.$$

Le terme $-\log(|\Delta x|)$ lié à la discrétisation fait diverger l'entropie mais le terme à gauche est une somme de Riemann qui converge vers l'intégrale

$$H_d(X) = -\int \log(p(x))\, p(x) dx \ .$$

Cette quantité est appelée *entropie différentielle*, elle est une extension naturelle de l'entropie pour des processus continus X associés à des distributions continues $p(x)$. Dans le reste de la thèse, les termes d'entropie se réfèrent sans distinction à l'entropie différentielle. Ce paragraphe est une introduction extrêmement brève de la notion de quantité d'information, le lecteur qui souhaite se familiariser davantage avec la théorie de l'information trouvera une présentation plus complète dans d'autres ressources [142, 102, 35, 154] .

Modèle. Dans ce chapitre et les chapitres suivants, je considérerai le modèle simplifié suivant pour le traitement de bas niveau des systèmes sensoriels (fig. 1.1). Les stimuli sont modélisés comme des vecteurs multidimensionnels $X \in \mathbb{R}^n$ générés par une source stochastique (fragments de parole de l'ordre de 10 ms). A partir de ces entrées, les vecteurs de sortie $Y \in \mathbb{R}^m$ sont obtenus par l'application de la matrice W, qui dans le cas linéaire ne dépend pas de l'entrée :

$$Y = W^T X$$

où $W = (W_1, W_2, ..., W_m)$ est un ensemble de filtres (banque de filtres). La transformation linéaire est une abstraction de l'action des cellules sensorielles (les cellules ciliées internes pour le système auditif). Les composantes du vecteur de sortie modélisent les excitations des canaux neuronaux (nerf auditif).

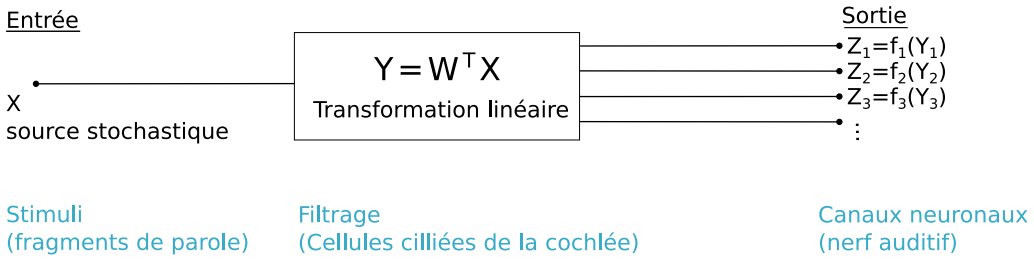


FIGURE 1.1 – Le traitement sensoriel est modélisé par un système simple à entrée/sortie. Les vecteurs de sortie Y sont des décompositions temps-fréquence de vecteurs d'entrée X, générés par une source stochastique. A partir d'un vecteur d'entrée (fragment de signal), le vecteur de sortie est obtenu par l'application d'une matrice W (fixée, dans le cas linéaire). Les composantes du vecteur de sortie Y modélisent les excitations des canaux neuronaux. L'opération linéaire peut être suivie d'une non linéarité, pour obtenir le vecteur d'activation Z.

Pour obtenir une mesure de l'activation des neurones, une opération non linéaire doit être appliquée au vecteur de sortie. Lorsque cette non linéarité est explicite, j'utilise aussi la variable Z :
$Z = f(Y)$

où f est une fonction non linéaire définie élément par élément ($z_i = f_i(y_i)$). Y est parfois appelé le vecteur d'excitation, et Z le vecteur d'activation (voir chap. 4 de ref. [170]). Une vue alternative consiste à décrire ce système comme un réseau neuronal artificiel multicouche (fig. 1.2), comme dans ref. [115].

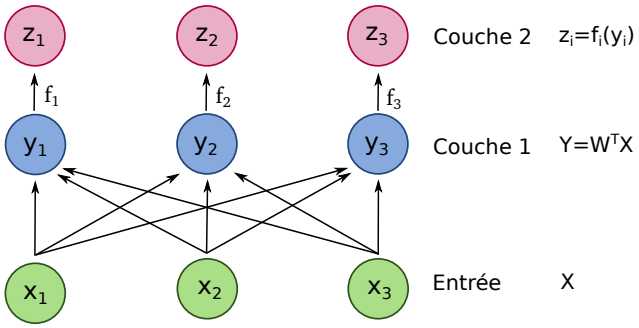


FIGURE 1.2 – Modèle de réseau de neurones à deux couches : la première couche est la sortie d'une opération linéaire, la seconde couche est la sortie d'une opération non linéaire.

Réduction de redondance. La réduction de la redondance a été le premier critère proposé pour l'hypothèse de codage efficace [16]. Il a été introduit par Barlow peu de temps après que Shannon ait défini la notion de redondance [142]. Selon Shannon, la redondance « quantifie le degré de contrainte imposé à [un processus] en raison de sa structure statistique » [143]. Shannon cite comme exemple le langage naturel, qui obéit à des règles statistiques (par exemple, la lettre 'e' est la plus fréquente, ou, en anglais, la lettre 'h' tend à suivre un't'). La redondance se définit par toute contrainte sur le processus qui rend un code naïf inefficace. Mathématique, la redondance s'exprime pour un processus Z par :

$$R(Z) = 1 - \frac{H(Z)}{C} \text{ [redondance dans le sens de Shannon et Barlow]} \tag{1.1}$$

où C est la capacité du canal. C est la valeur maximale possible de l'entropie pour un processus partageant les mêmes contraintes globales que le processus Z. La redondance, de ce point de vue, est une information inutile qui doit être filtrée, ou *compressée*, par un système de codage efficace. Lorsque la redondance est réduite, le canal utilise au mieux sa capacité de codage pour transmettre l'information de l'entrée.

Deux types de redondance. Atick a proposé de décomposer la redondance (eq. 1.1) du vecteur d'activation selon [10] :

$$R = 1 - \frac{H(Z)}{C} = \frac{1}{C}\overbrace{(C - \sum_i H(Z_i))}^{(a)} + \frac{1}{C}\overbrace{(\sum_i H(Z_i) - H(Z))}^{(b)} \tag{1.2}$$

Schématiquement, cette décomposition correspond à deux types de redondance :

— **a) :** $C - \sum_i H(Z_i)$: ce terme correspond à la somme des redondances pour chaque composante prise séparément, si l'on considère que la capacité totale est répartie sur plusieurs canaux $C = \sum_i C_i$. Ce type de redondance s'applique en particulier à un codage pour un signal univarié. La redondance augmente lorsque les distributions marginales (distributions unidimensionnelles) s'écartent de la distribution d'entropie maximale, par exemple si certaines valeurs deviennent plus typiques que d'autres.

Lorsque la somme des entropies marginales est égale à la capacité totale, le transfert d'information est maximisé : lorsqu'il n'y a pas de perte d'information au cours du traitement, cet objectif est appelé *maximisation de l'information* (critère *infomax*).

— **b)** : $\sum_i H(Z_i) - H(Z)$: ce terme est l'information mutuelle entre les composantes du vecteur du sortie. Ce type de redondance correspond à des informations qui sont codées plusieurs fois dans différents canaux de sortie. Cela est indésirable car cela revient à une mauvaise répartition des ressources de codage. Les codes qui minimisent ce terme ont été appelés *codes à entropie minimale* [10, 17]. Selon Barlow, ce type de redondance est le moins évident à déceler, et aussi le plus révélateur sur la structure statistique des données sensorielles [18]. La quantité de redondance dépend de la représentation sur laquelle repose le code neuronal, et est minimale si les composantes sont indépendantes. La minimisation de l'information mutuelle est ainsi le but de l'Analyse en Composantes Indépendantes (ACI).

Cette décomposition ne signifie pas cependant que les deux types de redondance sont des facteurs indépendants. Comme expliqué dans la version en anglais de la thèse, la minimisation de l'information mutuelle est largement redondante avec le critère *infomax*.

Critère *infomax*. Le critère *infomax* [99, 21] est un critère de maximisation de l'information mutuelle entre la sortie Z (l'activité neurale) et l'entrée X (entrée sensorielle). L'information mutuelle est la quantité d'information en commun entre les deux processus (fig 1.3). Elle se définie par

$$I(X,Z) = I(Z,X) = H(X) - H(Z) - H(X,Z)$$

où $H(X,Z)$ est l'entropie de la loi jointe. Autres définitions :

— avec l'*entropie conditionnelle* (ou *équivocation*) :
$I(X,Z) = H(Z) - H(Z|X)$.

— avec la *divergence de Kullback-Leibler* :
$I(X,Z) = D_{KL}\left(p(x,y)||p_x(x)p_y(y)\right)$,
où $p(x,z)$ est la distribution jointe de (X,Z), p_x et p_z sont resp. les distributions marginales de X et Z, et la divergence KL est définie par $D_{KL}(p||q) = \mathbb{E}_p(\log \frac{p}{q})$.

La formule avec l'entropie conditionnelle montre en particulier qu'en l'absence de bruit (f injective, W inversible), le critère *infomax* revient simplement à maximiser l'entropie de la sortie Z. En ce sens, ce critère est proche du critère de réduction de redondance de Shanon et Barlow (eq. 1.1), mais le critère *infomax* fait en plus la distinction entre l'information pertinente (liée à l'entrée) et l'information non pertinente, indésirable (bruit). Pour que le critère infomax soit contraignant sur le choix des paramètres, il faut naturellement qu'une contrainte de ressources soit imposée, autrement il suffirait de choisir des sorties avec des amplitudes toujours plus grandes. Un canal ne peut que coder pour un signal avec une dynamique déterminée, c'est-à-dire dont les valeurs sont délimitées. Une contrainte courante associée est de limiter la variance des canaux Z_i. Une forme avec un terme de pénalisation du problème infomax est ainsi [163] :

$$\max_{W,f} I(X,Z) + \rho \sum_{i=1}^{n} Var(Z_i) \ .$$

Une autre formulation est obtenue en reprenant la formule de l'information mutuelle avec l'entropie conditionnelle et en y introduisant un poids qui pénalise le terme $H(Z)$ ou amplifie l'effet du terme d'aversion au bruit $-H(Z|X)$. Atick propose par exemple [10] :

$$\max_{W,f}(1-\eta)H(Z) - H(Z|X)$$

avec $\eta > 0$.

► *Dans la version en anglais : il est expliqué que le critère infomax, lorsqu'il s'applique à une distribution unidimensionnelle, est équivalent à l'opération de « blanchiment » (transformation de la distribution de probabilité en la loi d'entropie maximale).*

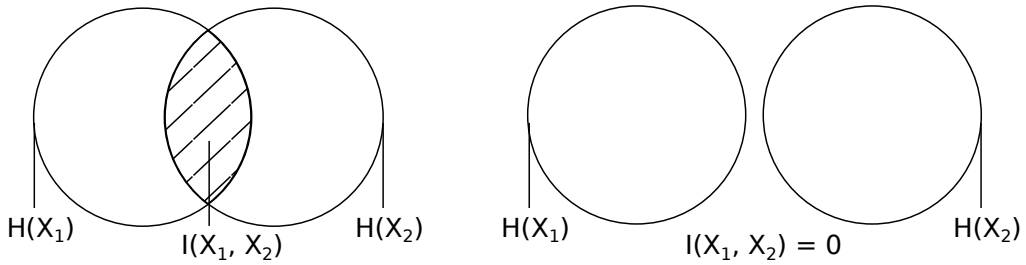


FIGURE 1.3 – L'information mutuelle $I(X_1, X_2)$ entre deux processus X_1 et X_2 est l'information que les processus possèdent en commun. Si l'information mutuelle est strictement positive, la connaissance d'un des processus donne une information sur l'autre ($H(X_1|X_2) < H(X_1)$ et inversement). Dans le cas où l'information mutuelle est nulle, les deux processus sont indépendants.

Code à entropie minimale Un *code à entropie minimale* [17] minimise l'information mutuelle entre les m composantes du vecteur de sortie :

$$I(Y_1, \cdots, Y_m) = \sum_{i=1}^{m} H(Y_i) - H(Y) \ . \tag{1.3}$$

Ce terme représente la partie de l'information redondante qui est due à la nature multivariée du processus (eq. 1.2, fig. 1.3). Ce terme est réécrit avec la divergence de Kullback-Leibler et les distributions de probabilité marginale $p_i(y_i)$:

$$I(Y_1, \cdots, Y_m) = D_{KL}(p(y)||p_1(y_1)p_2(y_2)\cdots p_m(y_m)) \ .$$

Ce terme est positif, et nul si et seulement si les composantes de sortie sont indépendantes. En cas d'indépendance, l'entrée est représentée par un ensemble de caractéristiques indépendantes, et la log-vraisemblance d'une configuration de sortie donnée est la somme des log-vraisemblances des probabilités marginales (*code factoriel*) :
$\log p(y) = \sum_i \log p_i(y_i)$.

Mesures d'indépendance. La principale difficulté pour trouver des composantes indépendantes est que l'indépendance stricte est une contrainte forte, qui ne peut être exprimée par une seule mesure empirique. Vérifier la contrainte d'indépendance $\log p(y) = \sum_i \log p_i(y_i)$ pour différentes décompositions nécessite l'estimation de nombreuses distributions de probabilités, ce qui est une tâche statistique exigeante lorsqu'il n'y a pas d'information *a priori* sur ces distributions. Les méthodes qui sont utilisées dans la pratique exploitent des définitions plus souples de l'indépendance, qui correspondent à différentes hypothèses concernant les distributions marginales. Une stratégie est étroitement liée à la définition

des processus indépendants : deux processus Y_1 et Y_2 sont indépendants si et seulement si pour chaque fonction $f_1, f_2 \in L_1$,

$$\mathbb{E}(f_1(Y_1)f_2(Y_2)) = \mathbb{E}(f_1(Y_1))\mathbb{E}(f_2(Y_2)))) \ .$$

Une définition moins stricte de l'indépendance consiste donc à choisir des fonctions spécifiques f_1, f_2, transformant la contrainte d'indépendance en une contrainte de décorrélation beaucoup plus faible [85]. D'autres stratégies sont basées sur des mesures de *negentropie* (défini par l'opposé de l'entropie) ou de non-gaussianité (voir paragraphe suivant). Si les coefficients en sortie sont supposés être distribués selon la distribution de Laplace, une mesure de *negentropie* correspondant à un terme d'entropie croisée est la norme l_1 :

$$||x||_1 = \mathbb{E}(|X|) \ .$$

Minimiser la norme l_1 pour une distribution (de variance fixée) revient à renforcer la parcimonie des décompositions. Cette stratégie a été utilisée plusieurs fois pour l'ACI appliquée à des données de parole [91, 153]. Voir la version anglaise pour plus de détails.

Structure vs diversité de la représentation. Un autre point de vue du *codage à entropie minimale* est que, pour obtenir un code maximalement compact, chaque canal de sortie doit coder pour une quantité minimale d'information. Cela conduit à minimiser la somme des termes marginaux d'entropie. Mais dans le même temps, toute la quantité d'information du processus d'entrée doit être bien représentée dans son ensemble. Par conséquent, le deuxième terme dans la fonction de coût
$h = \sum_i H(Y_i) - H(Y)$ est un terme de pénalité visant à garantir que l'information de l'entrée est conservée lors de la transformation des données. En particulier, ce terme empêche les filtres de se réduire à une seule dimension de l'espace. Ce point de vue montre qu'un code à entropie minimale doit trouver un équilibre entre *structure* (les composantes doivent correspondre à des distributions à entropie minimale) et richesse de la représentation ou *diversité* (les directions doivent représenter toutes les directions de l'espace).
Il a été montré que, empiriquement, les composantes sont rarement indépendantes, ou même décorrélées (par exemple, pour la parole, toutes les composantes sont activées au moment des excitations glottales). De fait, la structure peut devenir plus importante que la décorrélation des composantes dans l'ACI [79]. Une règle empirique pour trouver une représentation pour un code à entropie minimale est de trouver les directions maximisant la non-gaussianité (ou négentropie, c'est-à-dire l'opposé de l'entropie) des données, tout en s'assurant que les directions trouvées sont quasi-orthogonales [78, 79].

► *Dans la version en anglais : les lien entre negentropie, non-gaussianité et parcimonie est développé, en prenant les lois gaussiennes généralisées comme illustration. Il est aussi présenté un lien formel entre critère infomax et codage à entropie minimale.*

1.2 – Algorithmes et méthodes associées

Analyse en Composantes Indépendantes (ACI). La recherche de composantes indépendantes est l'objectif d'une classe d'algorithmes appelée Analyse en Composantes Indépendantes (ACI). L'ACI cherche une transformation – dans la plupart des cas linéaire – qui rend statistiquement indépendantes les composantes de données multivariées. Le terme a été introduit par C. Jutten et J. Herault à la fin des années 1980 [85] dans le contexte de la *séparation aveugle de sources* (SAS). La séparation aveugle de sources est bien connue dans la parole et les applications audio [167], car c'est la traduction mathématique du problème « cocktail party » : comment peut-on séparer un mélange de signaux de parole provenant de différents orateurs ? L'ACI permet d'apporter une solution à ce problème en tirant parti de l'indépendance des sources. Dans la théorie du codage efficace, la vue est inversée, les unités de codage neuronal se comportant comme si elles étaient les sources du signal entrant. L'algorithme initial de Jutten et Herault était motivé par le constat que la décorrélation des composantes était une contrainte trop faible pour retrouver des sources indépendantes. Ils ont dérivé une règle, semblable à la règle de Hebb, qui permet d'annuler des termes de corrélation croisée non linéaire (à l'aide de fonctions de transfert polynomiales). D'autres propositions qui ont suivi sont basées sur le même principe (voir Comon, 1994 [34] pour un historique plus détaillé des débuts de l'ACI). Il existe de multiples approches pour la modélisation de l'ACI [78], qui aboutissent toutes à des fonctions de coût similaires. Les paragraphes précédents ont introduit deux approches motivées par la théorie de l'information : la minimisation de l'information mutuelle, [34] d'une part, et le principe *infomax* [99], d'autre part. Deux autres approches de l'ACI souvent mentionnées sont le principe du maximum de vraisemblance (étroitement lié à la modélisation du problème de séparation de sources), et la maximisation de la non-gaussianité [34]. Cette dernière approche est une sorte de version inverse du théorème de la limite centrale : la gaussianité augmente lorsque des variables aléatoires sont mélangées, donc la gaussianité devrait diminuer lorsque ces variables sont au contraire séparées.

Fonction de coût de l'ACI. Le but de ACI est de trouver une matrice W telle que les composantes $Y_1, \cdots, Y_m$ de $Y = W^T X$ soient statistiquement indépendantes. Si l'on considère que la matrice W est une matrice carrée ($m = n$), on a $H(Y) = H(W^T X) = H(X) + \log|\det W|$. Le premier terme est la quantité d'information de l'entrée et ne dépend pas de W. Par conséquent, on peut dériver une fonction de coût pour l'ACI à partir de la contrainte de minimisation de l'information mutuelle (eq. 1.3) avec :

$$h(W) = \sum_i H(Y_i) - \log|\det W| = \sum_{i=1}^{n} H(W_i^T X) - \log|\det W| \ .$$

L'ACI cherche à minimiser la somme des termes d'entropie, avec la contrainte que W ne doit pas devenir dégénéré (d'où le terme de pénalité $\log|\det|$). Toutefois, les termes d'entropie marginale restent à estimer. Pour cela, on fait une hypothèse *a priori* sur les distributions des coefficients selon la décomposition de l'ACI. Les termes d'entropie peuvent être remplacés par les termes d'entropie croisée avec une distribution a priori q :

$$H(p_i) = H(p_i, q) + D_{KL}(p_i||q) \approx H(p_i, q) = -\mathbb{E}(\log q(y_i))$$

où p_i est la distribution marginale de la composante Y_i (le terme avec la divergence KL est

ignoré). $h(W)$ devient donc :

$$h(W) = -\sum_{i=1}^{n} \mathbb{E}\left(\log q(W_i^T X)\right) - \log|\det W| \ . \tag{1.4}$$

En particulier, si l'on utilise la distribution de Laplace comme distribution a priori, la fonction de coût est :

$$h(W) = \sum_{i=1}^{n} \left[\frac{1}{2}\log\left(2\sigma_i^2\right) + \sqrt{2}\ ||W_i^T x||_1/\sigma_i\right] - \log|\det W|$$

où σ_i est l'écart-type estimé de la i-ième composante.

► *Dans la version en anglais : il est mentionné un algorithme de descente de gradient pour minimiser $h(W)$ par rapport à W.*

Méthodes de codage parcimonieux. De nombreuses méthodes sont liées à la recherche de représentations parcimonieuses. L'analyse en Composantes Indépendantes elle-même, comme indiqué dans le paragraphe précédent, est souvent utilisée avec une distribution a priori encourageant des décompositions parcimonieuses (par exemple, distribution de Laplace) [91], et le terme d' *Analyse en Composantes Indépendantes parcimonieuses* a été inventé [67]. Les méthodes de codage parcimonieux viennent souvent avec la notion de *dictionnaire*. Un dictionnaire $\mathbf{D}$ est un ensemble de vecteurs $(d_1, ..., d_m) \in \mathbb{R}^{n\times m}$ dont les éléments sont appelés les *atomes*. Le but du codage parcimonieux est de représenter les signaux d'entrée dans ces dictionnaires avec un petit nombre d'atomes. Contrairement à l'ACI, qui fonctionne de préférence avec des matrices carrées ($m = n$), les méthodes de codage parcimonieux font le plus souvent appel à des familles de vecteurs redondantes ($m > n$). Les atomes peuvent être soit fixés à l'avance (dictionnaires de Gabor, ondelettes [105], etc.), soit appris des données. Lorsque les dictionnaires sont fixes, les atomes sont choisis de manière à ce que les dictionnaires aient une certaine « structure » (par exemple, les atomes d'un dictionnaire de Gabor sont des versions de la même fonction de base décalées dans le plan temps-fréquence).
Dans cette thèse, je fais une distinction entre deux paradigmes (voir aussi réf. 14 qui adopte un point de vue similaire) :

1. Le paradigme *analyse* du signal – les décompositions sont calculées en appliquant le produit scalaire entre le signal et les atomes ; ce qui signifie que le vecteur d'intérêt est
$$Y = W^T X = \mathbf{D}^T X,$$
comme cela a été le cas jusqu'ici.
2. Le paradigme de *reconstruction* du signal (ou *analyse-synthèse*) – seuls les atomes les plus significatifs sont sélectionnés, en faisant en sorte que le signal soit la somme des atomes sélectionnés. Le vecteur d'intérêt est un vecteur parcimonieux $Y = g(X)$ qui minimise l'erreur de reconstruction :
$$\epsilon = ||X - DY||_2 \ .$$

Le deuxième paradigme est mathématiquement attrayant, parce qu'il conduit à des problèmes bien posés, même dans le cas où les dictionnaires sont redondants. Pour cette raison, c'est le paradigme dominant dans les méthodes de codage parcimonieux. Une

limitation est qu'il existe une hypothèse implicite selon laquelle chaque signal d'entrée est une somme parcimonieuse d'atomes, mais cette condition n'est pas toujours vérifiée (par exemple, de nombreux sons de parole sont similaires à des bruits : fricatives, plosives...). Le paradigme de *reconstruction* s'accompagne de techniques de reconstruction du signal à partir des atomes (la version anglaise de la thèse fait un rapide inventaire de ces méthodes). La reconstruction du signal à partir d'un nombre réduit d'atomes permet une représentation plus parcimonieuse des données d'entrée, en échange cependant d'un coût algorithmique supplémentaire. Cette thèse se concentre principalement sur le point de vue *analyse du signal* : en d'autres termes, je n'utilise pas des méthodes pour reconstruire le signal à partir des activations neuronales. Savoir quel paradigme est le plus pertinent pour les systèmes sensoriels est un sujet de débat. Un point en faveur du paradigme *analyse du signal* est que cette première étape est nécessaire dans les deux cas, et qu'une reconstruction complète du signal est un problème complexe dans le cas général.
La façon la plus naturelle d'évaluer la parcimonie d'un vecteur est la "norme" l_0, qui est simplement le nombre de coefficients non nuls d'un vecteur. Cependant, la "norme" l_0 est une fonction non convexe et non continue, ce qui aboutit à des problèmes d'optimisation complexes lorsqu'elle intervient. Une solution est de remplacer la "norme" l_0 par une version régularisée, la norme l_1 [172] (faisant le lien à nouveau avec les méthodes décrites précédemment lorsque la distribution de Laplace est utilisée comme *a priori*).

Comment prendre en compte des représentations redondantes ? L'ACI fonctionne de préférence dans le cas déterminé (n la dimension est égale à m le nombre de sources). Dans ce cas, l'entropie du vecteur de sortie $H(Y)$ dans la définition de l'information mutuelle (eq. 1.3) :

$$h = \sum_{i=1}^{m} H(Y_i) - H(Y) \qquad \text{(infomin, rappel)}$$

peut être remplacée par le terme $\log|\det W|$, ce qui permet d'obtenir une fonction de coût pour l'ACI (eq. 1.4). Cependant, les bases non redondantes ont deux inconvénients : a) elles ne sont pas faciles à manipuler. Par exemple, il n'est pas simple de construire des bases d'ondelettes orthogonales [105]. b) elles ne correspondent pas à une situation de codage réel. Les modèles basés sur des représentations redondantes sont plus réalistes.
Dans le cas déterminé, la pénalité $\log|\det W|$ assure que les directions des composantes indépendantes ne sont pas corrélées en plus de capturer la structure statistique. La principale difficulté si l'on cherche à reproduire la dérivation de la fonction de coût lorsque la représentation est redondante est qu'il n'y a pas d'expression naturelle qui généraliserait ce terme de pénalité [79]. Une solution, évoquée dans le paragraphe précédent, est de forcer la parcimonie des vecteurs Y en essayant de reconstruire les entrées X à partir de Y [95, 67]. Le paradigme *reconstruction du signal* atténue le problème de la mesure de la corrélation des composantes en sélectionnant les atomes, ce qui conduit à des problèmes d'optimisation bien posés. L'autre solution, adoptée dans la suite de la thèse, n'est pas de contraindre le vecteur Y, mais plutôt de contraindre la représentation, c'est-à-dire que la matrice W (ou $\mathbf{D}$) est un ensemble d'un ensemble restreint $\mathbf{S} = \{\mathbf{D}_\theta\}_\theta$, paramétré par θ. Les dictionnaires de cet ensemble sont considérés comme équivalents du point de vue de la *diversité* des atomes (par exemple, il peut s'agir de dictionnaires de Gabor uniformément répartis dans l'espace temps-fréquence), et la fonction de coût se limite à caractériser la *structure* (minimisation de la somme des termes d'entropies des distributions marginales) :

$$h(\theta) = \sum_i H([D_\theta]_i^T X) \tag{1.5}$$

1.3 – Analyse temps-fréquence et décompositions parcimonieuses

Dans les sections précédentes, l'efficacité du codage a été reliée à des critères s'appliquant à des décompositions du signal d'entrée (en particulier, le critère de parcimonie a été évoqué). Je n'ai toutefois pas précisé la nature de ces décompositions. Lorsque l'ACI ou des méthodes de codage parcimonieux sont appliquées à des signaux sensoriels (images [120, 164] et sons naturels), la décomposition que l'on obtient reproduit une analyse temps-fréquence du signal. Le chapitre 2, dans la version anglaise, aborde le problème du codage parcimonieux, sous l'angle de l'analyse temps-fréquence, complémentaire de l'approche de la théorie de l'information. Le système entrée/sortie est relié aux représentations temps-fréquence quadratiques, en particulier la distribution de Wigner-Ville (croisée), définie par

$$\mathcal{W}(f,g)(x,\omega) = \frac{1}{2\pi}\int_t f(x+\frac{t}{2})\overline{g(x-\frac{t}{2})}e^{-i\omega t}\mathrm{d}t$$

pour $f, g \in L(\mathbb{R}^2)$. Le principe d'incertitude, qui nous dit qu'un filtre ne peut à la fois répondre spécifiquement à une fréquence et à un instant donné, limite la parcimonie des représentations temps-fréquences. Je présente dans ce chapitre un théorème, dû à Lieb [98], qui donne une borne inférieure de la parcimonie des décompositions temps-fréquence quadratiques (évalué selon la norme l_1) :

Théorème (Principe d'incertitude de Lieb [98])**.** *Soit $f, g \in L^2(\mathbb{R})$. La norme l_1 de la distribution de Wigner-Ville croisée entre f et g est minorée par le produit des normes euclidiennes :*

$$||\mathcal{W}(f,g)||_1 \geq ||f||_2||g||_2 \ . \tag{1.6}$$

L'égalité est atteinte si et seulement si $g = \lambda f$ ($\lambda \in \mathbb{C}$) (à modulations et translations près) et f est une ondelette de Gabor généralisée (c'est-à-dire avec potentiellement un 'chirp').

Ce théorème, outre expliciter une borne inférieure pour la parcimonie des décompositions temps-fréquence, permet de mettre en évidence l'importance des ondelettes de Gabor (sinusoides modulées par une gaussienne). Dans ce chapitre, je présente enfin quelques résultats fondamentaux de la théorie des *frames*, et introduis une famille de représentations dont le facteur de qualité suit la propriété de loi de puissance définie dans l'introduction (eq. 1).

1.4 – Limites

Le cerveau est-il vraiment analogue à un système d'encodage de l'information ? La théorie du codage efficace reprend des éléments de vocabulaire de la théorie de l'information (code, information mutuelle, capacité d'un canal, redondance...) et l'applique au code neuronal qui émerge de l'activité d'un grand nombre de neurones. Il n'est pas clair en vérité comment ces concepts reflètent l'activité neuronale [27] . Même en considérant un seul ou un petit nombre de neurones, comment mesurer la quantité d'information encodée par ces neurones (voir ref. 2 chap. 1 pour une analyse détaillée de cette question) ? La solution la plus simple est de compter le nombre de potentiels d'action (*spikes*) délivrés (ou de manière

équivalente estimer le taux de décharge). Cette solution a l'intérêt qu'elle s'approche du calcul mathématique de la quantité d'information pour un codage statique et parcimonieux. Mais cela ne prend pas en compte le codage temporel : les instants d'émission des potentiels d'action, les (auto)corrélations temporelles, etc. Considérer le codage neuronal selon le seul angle de la théorie de l'information, qui quantifie l'information par des distributions de probabilité d'une grandeur figée reflétant l'activité neuronale (ex : taux de décharge des potentiels d'action), ne prend pas en compte tous les aspects dynamiques et les implications de la connectivité neuronale. Une implication est que le fonctionnement physique des neurones introduit nécessairement du bruit dans la transmission de l'information qui reflète les mécanismes neuronaux, sans que ce bruit puisse être facilement modélisé [1]. Pour appliquer des résultats de la théorie de l'information, il est nécessaire d'avoir une modélisation stochastique claire de l'environnement et de la représentation qui en est faite par le système de codage. Or, construire une représentation de l'environnement n'est pas la finalité de tous les processus psychologiques qui s'inscrivent dans des boucles *perception-action*. Toutes ces raisons invitent à appliquer le vocabulaire des systèmes de communication aux processus cognitifs avec suffisamment de précaution, en ayant en tête qu'il s'agit avant tout d'une simplification permettant de faire intervenir des outils mathématiques.

Représentations génériques vs représentations spécialisées. La combinaison de critères d'efficacité de codage de l'information avec des modèles d'apprentissage (supervisés ou semi-supervisés) est nécessaire pour modéliser les phénomènes cognitifs car le cerveau n'encode pas l'information totale qui est perçue, mais supprime au cours des étapes du traitement sensoriel des informations non pertinentes pour les tâches de plus haut niveau (concernant le son et la parole, ces tâches peuvent être la reconnaissance de la parole, l'identification du locuteur ou la reconnaissance d'une scène acoustique). Ces tâches de haut niveau nécessitent le passage d'un code continu à des représentations discrètes, catégorielles (ex : identification des phonèmes) [141]. Elles cherchent des invariants de classes qui permettent de discriminer les catégories entre elles et, à l'inverse, elles se débarrassent de l'information qui correspond aux variations au sein d'une même classe. Ce filtrage de l'information – de plus en plus fin – a été appelé *information bottleneck* (goulot d'étranglement de l'information). Une modélisation du goulot d'étranglement de l'information, avec une formulation similaire au critère *infomax*, est de maximiser la quantité

$$\max_{p(z|x)} I(H,Z) - \frac{1}{B} I(X,Z)$$

où H désigne l'information réellement pertinente de l'information d'entrée (par exemple, la catégorie à laquelle X appartient). Un codage par catégories a des incidences sur l'encodage de bas niveau. Par exemple, si les catégories sont codées par une assemblée de neurones, la population de neurones va avoir des champs récepteurs concentrés au niveau des frontières de ces catégories dans l'espace des représentations [23]. Cependant, si le signal est exploité par diverses fonctions cognitives de haut niveau, chacune se basant sur des pans d'information différentes du signal, alors une première étape du traitement sensoriel est d'aboutir à une représentation générique prenant en compte le maximum d'information du signal. L'hypothèse du codage efficace est surtout pertinente dans l'étude

1. Un contre-exemple intéressant est donné par l'inférence de la sensibilité de contraste du système visuel selon la fréquence spatiale pour des stimuli de faible intensité [@VanHateren1992 ; @atick1992]. Dans ce cas, on sait que l'émission de potentiels d'actions introduit un bruit multiplicatif et cette connaissance intervient dans le calcul.

des processus sensoriels périphériques de bas niveau. Pour des processus de haut niveau, les représentations neuronales se font plus abstraites et complexes, cherchant des invariants de classe et par conséquent supprimant la connaissance de la variabilité intra-classe. De tels processus se prêtent plus difficilement à une modélisation selon le paradigme du codage efficace. Les algorithmes et méthodes liées à la théorie du codage efficace ne font pas d'hypothèse sur quelle est l'information pertinente à extraire de l'environnement : en général il s'agit d'algorithmes non supervisés. La perception de plus haut niveau (ex : reconnaissance d'objets) sont analogues à des tâches de classification, les méthodes associées sont des algorithmes supervisés (l'information de la connaissance de l'objet est donnée à l'algorithme) ou faiblement supervisés.

Le code neuronal est redondant Une autre critique de la théorie du codage efficace et du principe de réduction de la redondance est que le code neuronal est en vérité redondant. Un exemple souvent mentionné est le nombre de neurones dans le cortex visuel primaire V1 ($\sim 10^9$ neurones), beaucoup plus que le nombre de cellules ganglionnaires rétiniennes ($\sim 10^6$ cellules) [18]. La redondance est importante pour deux raisons : a) elle rend la représentation neurale plus robuste au bruit [99] ; b) le chevauchement entre les caractéristiques codées est nécessaire pour capturer les détails du signal (par exemple, les bords d'une image). En analyse temps-fréquence et dans la théorie des *frames* de Gabor, on sait que la redondance est une propriété inévitable pour de bonnes caractéristiques d'une analyse temps-fréquence [72, 134]. Le nombre élevé de neurones dans V1 pourrait être le résultat d'un équilibre entre représentation parcimonieuse à un bas niveau, et une représentation parcimonieuse à un niveau supérieur pour des caractéristiques plus abstraites [121]. En 2001, H. Barlow a écrit un article intitulé *Redundancy reduction revisited* [18], dans lequel il énumère les raisons pour lesquelles il est toujours pertinent de rechercher des représentations minimisant la redondance même si le code neuronal est redondant. De façon générale, un argument est que le codage parcimonieux est une façon de trouver des traits caractéristiques d'un signal, de façon non supervisée.

Chapitre 2

Structure statistique de la parole

L'objectif de ce chapitre est de décrire la structure statistique de la parole, et de lier cette structure aux propriétés acoustiques de la parole. L'analyse porte sur des signaux synthétiques dont on peut penser qu'ils ont une structure analogue à certains sons de parole (correspondant au chap. 3 de la version anglaise), et sur des données réelles de parole (correspondant au chap. 4 de la version anglaise). Le comportement de β, qui contrôle la sélectivité fréquentielle dans la gamme des hautes fréquences, est analysé pour des sous-classes des sons de la parole, chaque fois à un niveau plus fin (catégories phonétiques, puis phonèmes, voire parties de phonèmes). Les données réelles proviennent de la base de données TIMIT [56], qui est composé d'enregistrements de parole (en anglais américain) et qui donne également la nature des phonèmes segment par segment. L'étude de signaux artificiels permet en complément d'acquérir une meilleure intuition des caractéristiques acoustiques les plus pertinentes pour la structure statistique de la parole, et aide ainsi l'interprétation des données réelles. Deux types de signaux artificiels sont ainsi été analysés. Le premier type de signal est du bruit localisé (par la multiplication d'une fenêtre) en temps ou en fréquence. Ce premier type de sons synthétiques se rapporte aux consonnes (ex : fricatives). Le deuxième type de signaux est des sons émis par un guide d'ondes cylindrique uniforme, avec des cylindres de différents rayon. Ce type de signal est similaire aux voyelles.

2.1 – Méthodes

Grandes lignes. Le schéma général de la méthode est le suivant :.

1. Création d'un ensemble de dictionnaires redondants W_β d'ondelettes de Gabor de $\beta = 0.3$ à $\beta = 1.2$, dont les atomes sont uniformément distribués en temps-fréquence-phase.
2. Génération de fragments de parole de 16 ms X à partir de signaux de parole (ou de signaux artificiels).
3. Prétraitement des signaux : filtrage (filtre passe-haut à $f_c = 1.5kHz$) et normalisation (de telle sorte que la valeur quadratique moyenne quadratique soit constante).
4. Calcul de la fonction coût, reflétant (l'absence de) structure des décompositions et basé sur la mesure de parcimonie donnée par la norme l_1 (comme motivé dans le chapitre précédent)

$$h(\beta) = \mathrm{E}\left(||W(\beta)^T X||_1\right)$$

5. Sélection du meilleur paramètre minimisant la fonction de coût :

$$\beta^\star = \arg\min_\beta h(\beta) \ .$$

Dictionnaires. Les candidats pour les représentations les plus parcimonieuses sont un ensemble de 30 dictionnaires redondants dont les atomes sont des filtres de Gabor (fig 2.1). Chaque dictionnaire est composé de 600 filtres répartis uniformément en temps, fréquence et phase. Le choix des filtres de Gabor est motivé mathématiquement (voir chapitre 2 de la version anglaise) et est également cohérent avec les formes de filtres trouvées empiriquement avec l'ACI [84, 94, 91]. Lorsque l'ACI est appliqué à des sous-classes suffisamment larges de sons de la parole, le facteur Q_{10} des filtres appris par rapport à la fréquence centrale est bien modélisé par une droite (sur une échelle *log-log*). Les études antérieures ont montré que l'ordonnée à l'origine est redondante avec la pente de la régression, puisque la plupart des lignes se croisent autour du point ($f_0 = 1kHz, Q_0 = 2$) pour diverses données de parole à l'entrée de l'ACI [153, 47]. La pente de régression de Q_{10} sur f_c est le paramètre β qui a été introduit plus tôt dans cette thèse. Il peut être considéré comme un paramètre synthétique des représentations obtenues avec l'ACI. Chaque dictionnaire correspond à une valeur de β. La plage de valeurs β [0.3, 1.2] a été choisie de manière à comprendre toutes les valeurs prises par β dans l'analyse de Stilp et Lewicki [153]. Pour permettre une plus grande diversité des filtres, j'ai introduit un bruit multiplicatif dans le calcul du facteur Q_{10} (estimé à partir de graphiques d'ACI obtenus sur des données réelles), de sorte que Q_{10} vérifie :

$$Q_{10}(f) = \log Q_0 + \beta(\log f - \log f_0) + 0.04\eta \tag{2.1}$$

où le log est pris en base 10 et η est un bruit blanc gaussien. Quant aux autres paramètres qui sont distribués uniformément, les intervalles sont respectivement [1–6.5 kHz],[2–14 ms] et [0, π] pour la fréquence centrale, le décalage temporel et la phase.

Fonction de coût. On considère des vecteurs X de dimension n, qui peuvent être des fragments de signaux de parole. Le but est de sélectionner le meilleur ensemble de filtres $W_\beta = (W_1, ..., W_m)$ parmi les dictionnaires d'ondelettes de Gabor, indexés par β. Je désigne $Y_\beta = W_\beta^T X$ les vecteurs de sortie qui sont les décompositions des vecteurs d'entrée dans les chacun des dictionnaires. Une mesure brute de la parcimonie est exprimée par la somme des activations :

$$h_\beta(X) = ||Y_\beta||_1 = \sum_i |Y_{\beta,i}| \ . \tag{2.2}$$

Cette mesure est la norme l_1 du vecteur de sortie. Cette fonction de coût a été motivée dans le chapitre précédent. Sur des données réelles, la fonction de coût inclut également des pondérations comme facteurs de normalisation (+2,5dB/octave), qui correspondent à une normalisation spectrale partielle. h_β est ensuite moyennée en prenant en compte un groupe d'échantillons appartenant à la même catégorie, et la fonction de coût est normalisée en prenant $h_\beta = 1$ pour les signaux maximisant la fonction de coût. La meilleure valeur β minimise la fonction de coût moyen :

$$\beta^\star = \arg\min_\beta h_\beta$$

Dans le reste de la thèse, β fait référence à $\beta^\star$, le choix optimal du paramètre, quand il n'y a pas de confusion possible avec les autres valeurs.

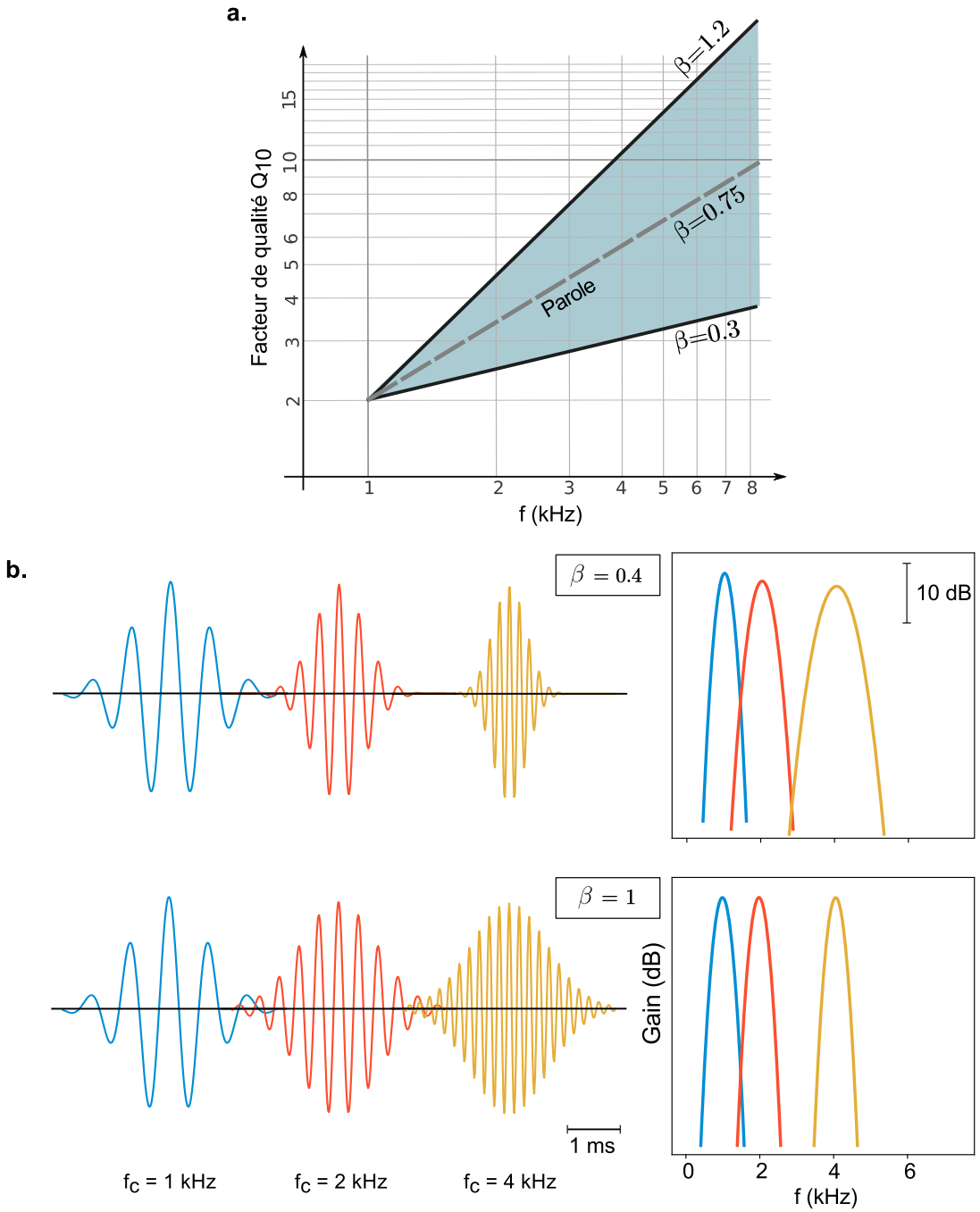


FIGURE 2.1 – **a.** Les dictionnaires, candidats pour la meilleure représentation, sont caractérisés par différentes lois de puissance pour Q_{10}, d'exposant β, avec β allant de 0.3 à 1.2.
b. Exemples de filtres Gabor utilisés dans les dictionnaires.
Gauche : Signaux pour deux valeurs de β à trois valeurs de f_c .
droite : Réponses en fréquence correspondantes

Le coût défini dans eq. 2.2 est une mesure du manque de structure. Les valeurs moyennes de h sur l'ensemble des dictionnaires de Gabor ont été considérées simultanément avec $\beta = \beta^\star$ dans les analyses. β décrit le comportement général de la représentation optimale pour la sélectivité fréquentielle, tandis que h quantifie le coût de calcul pour décomposer le signal. Des valeurs basses de h caractérisent les sons qui présentent une structure, typiquement des voyelles. Au contraire, des valeurs maximales de β caractérisent les sons qui sont similaires à du bruit (consonnes obstruantes : fricatives, stops...). Une autre interprétation de h qui s'applique à de nombreux sons de parole est qu'il s'agit d'une mesure de localisation. Un signal avec un seul pic est associé à un coût minimum h : si ce pic est sur l'axe des temps, β sera aussi minimal, si la localisation est en fréquence, alors β sera maximal. Cette interprétation est illustrée par l'analyse de signaux artificiels qui sont des bruits fenêtrés (chap. 3 en version anglaise).

Données. La génération des données synthétiques est décrite dans le chapitre 3. Les résultats ne sont pas détaillés dans ce résumé, mais j'évoque toutefois les principales conclusions dans la section suivante. Les données réelles de paroles ont été extraites de la base de données TIMIT [56]. TIMIT fournit des exemples audio de phrases en anglais américain ainsi que des informations sur leur contenu phonétique par segment. Des fragments de 16 ms sont sélectionnés, représentant 256 échantillons à $f_s = 16kHz$.

2.2 – Résultats

2.2.1 Données synthétiques

Cette section explique la motivation derrière l'analyse de signaux synthétiques, puis décrit les principaux enseignements de cette analyse.

Bruits localisés. La motivation de l'analyse de la structure statistique du signal pour des bruits localisés (fenêtrés en temps ou en fréquence par des gaussiennes de taille variable) est d'expliquer pourquoi on obtient des valeurs de β minimales ou maximales pour certains sons, notamment les consonnes obstruantes (fricatives, plosives, consonnes affriquées). Dans la production de la parole, les deux principales sources de sons sont : a) les vibrations des cordes vocales (qui intervient notamment dans la production des voyelles), b) du bruit turbulent, produit lors d'une constriction du conduit vocal et/ou lorsque le débit de l'air est augmenté (ex : aspirants). Cela génère du bruit qui est ensuite filtré par le conduit vocal. Souvent, même les premières résonances sont dans les hautes fréquences (ex : fricatives sibilantes), de sorte qu'un modèle assez grossier de ces sons est du bruit blanc gaussien passant par un filtre passe-haut. Autre type de son suivant ce modèle, cette fois dans le domaine temporel : le relâchement d'occlusion pour les plosives peut être modélisé par un bruit fenêtré (par une fenêtre rectangulaire) pour marquer le début de la plosive (implosion). On peut montrer alors que, suivant ce modèle de bruit fenêtré par une fenêtre rectangulaire (fig. 2.2), la décomposition sera plus parcimonieuse en temps (resp. en fréquence) si le fenêtrage se fait sur le signal dans le domaine temporel (resp. domaine fréquentiel). Les résultats de la simulation sur des bruits localisés sont cohérents avec cette explication. Se référer à la version anglaise pour plus de détails.

Pseudo-voyelles générées par un guide d'onde cylindrique. Le deuxième type de signaux

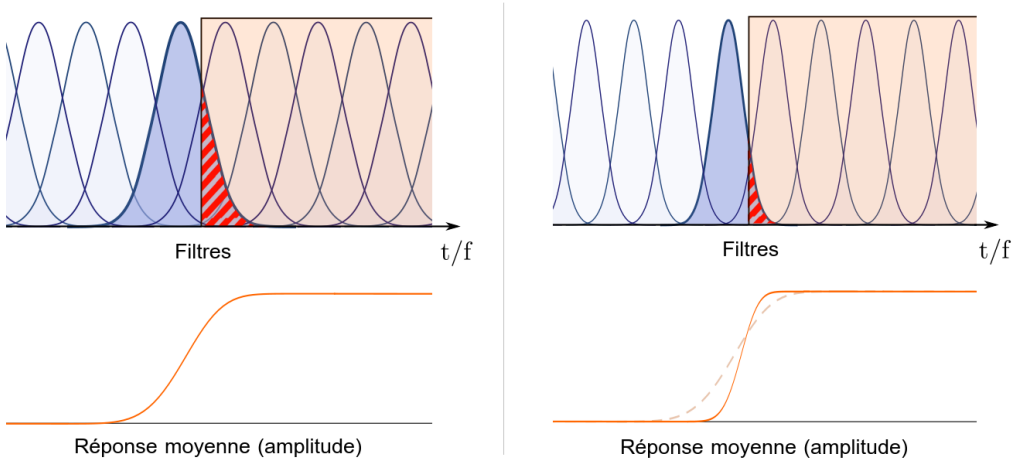


FIGURE 2.2 – Exemples de deux situations avec des filtres gaussiens disposés sur l'axe temporel (ou fréquentiel). Les filtres gaussiens diffèrent par leur sélectivité (à gauche : filtres à bande large, à droite : filtres à bande étroite). Les zones rouges représentent le bruit fenêtré. En bas : Réponse moyenne (amplitude) en fonction de la position du filtre. Des filtres sélectifs (en fréquence ou en temps selon l'axe) améliorent la parcimonie de la réponse. *En version couleur en ligne.*

générés pour l'analyse de la structure statistique de la parole est des sons émis par un guide d'onde cylindrique. Le but est d'analyser des sons similaires à des voyelles avec différentes largeurs de bande pour les formants. La réflexion des ondes acoustiques dans le conduit vocal résulte en des fréquences de résonance qui sont disposées uniformément sur l'axe des fréquences (modes). Ces pics sur le spectre, visibles pour toute voyelle, sont appelés les formants. Les formants, en particulier les premiers formants F1–F3, varient en fonction de la configuration du conduit vocal, et caractérisent les différentes voyelles. Les valeurs des formants ne sont pas les premières variables d'intérêt pour la structure statistique de parole. Un facteur plus déterminant est la largeur de la bande critique des formants : un élargissement de la bande critique correspondant a priori à un facteur de qualité inférieur pour la meilleure décomposition. Les largeurs des bandes critiques des formants sont déterminées par le niveau d'amortissement dû aux pertes acoustiques. Les facteurs acoustiques déterminants sont les pertes par vibration des parois pour les basses et moyennes fréquences et les pertes par rayonnement acoustique pour les plus hautes fréquences [48, 152]. Puisque les formants supérieurs jouent un plus grand rôle dans la détermination de β, un facteur clé pour la structure statistique des voyelles est probablement le degré de rayonnement acoustique au niveau des lèvres. Le but des exemples synthétiques était de simuler cet effet, dans un contexte simplifié d'un guide d'onde cylindrique, à rayon uniforme. Le degré de rayonnement acoustique augmente avec la fréquence et l'ouverture des lèvres, par conséquent l'objectif des simulations était de montrer que β diminue lorsque le rayon du cylindre augmente (voir fig. 2.3 et fig. 2.4) . L'estimation du degré du rayonnement acoustique en prenant seulement en compte le rayon d'ouverture n'est cependant qu'une approximation, en particulier cela ne tient pas compte des réflexions internes [48].

► *Dans la version en anglais : des détails sont donnés sur la génération des signaux simulant un guide d'onde cylindrique. Quelques notions d'acoustique sont introduites à cette occasion.*

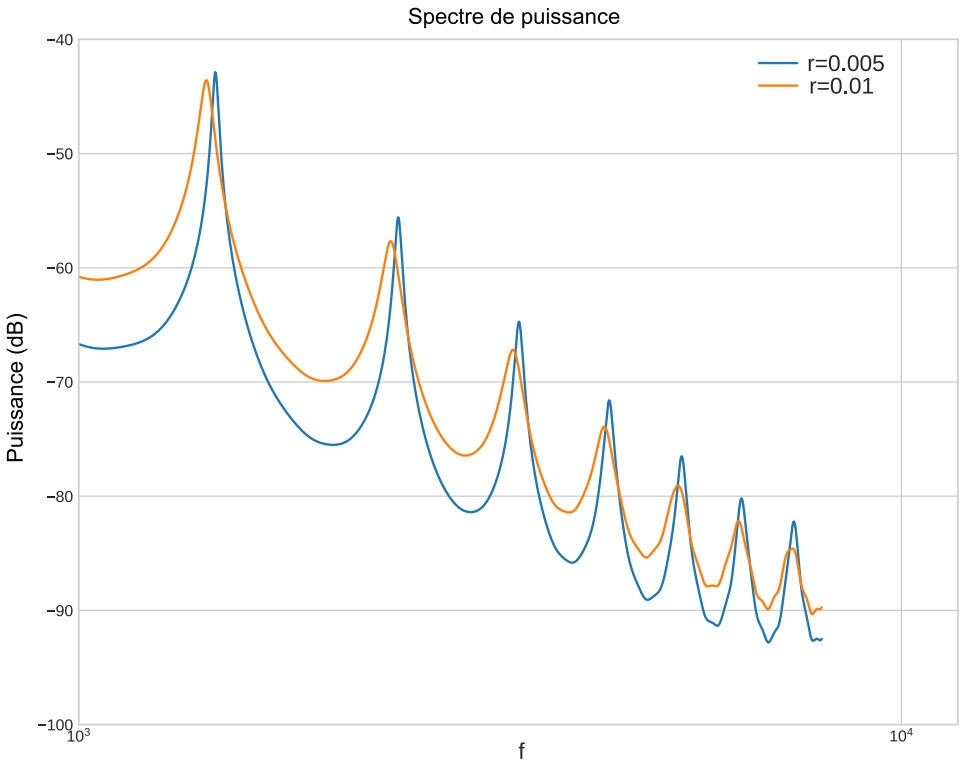


FIGURE 2.3 – Spectre de puissance de sons émis par un guide d'onde cylindrique uniforme, selon deux rayons différents. *Bleu :* r=0,5cm. *Orange :* r=1cm. *Version en couleur en ligne.*

Enseignements. L'analyse effectuée sur les signaux synthétiques donne quelques indications sur les facteurs acoustiques les plus importants pour la structure statistique de la parole :

- Pour les consonnes obstruantes (fricatives, consonnes affriquées, plosives), qui sont similaires à des bruits : il est attendu que, pour ces phonèmes, le comportement de β soit lié à la localisation du bruit sur l'axe temporel fréquentiel. Les plosives montrent une forte augmentation de l'intensité au moment du relâchement d'occlusion et devraient en principe être associées à de faibles valeurs de β. Au contraire, les fricatives présentent une forte augmentation du spectre de puissance dans les hautes fréquences, et on peut s'attendre à des valeurs élevées pour β. Les consonnes affriquées ont un comportement hybride, avec des caractéristiques empruntées des deux cas.
- Pour les voyelles, un facteur acoustique important est l'ouverture au niveau des lèvres, qui contrôle le degré de rayonnement acoustique. Une plus grande ouverture signifie des pertes acoustiques plus importantes et des bandes critiques formantiques plus importantes. Par conséquent, on s'attend à ce que les plus petites valeurs de β soient obtenues pour les voyelles avec la plus grande ouverture au niveau des lèvres. L'ouverture au niveau des lèvres est similaire à la notion de hauteur (distinction voyelle fermée/voyelle ouverte) mais pas tout à fait équivalente. La hauteur de la voyelle fait référence à la position de la langue et de la mâchoire, mais certaines voyelles ouvertes, telles que [O], sont prononcées avec des lèvres arrondies et ainsi une plus petite ouverture au niveau des lèvres.

L'analyse des signaux synthétiques montre également l'importance de la distinction entre les sons *structurés* et *non structurés*. Les sons structurés sont ceux qui peuvent être bien

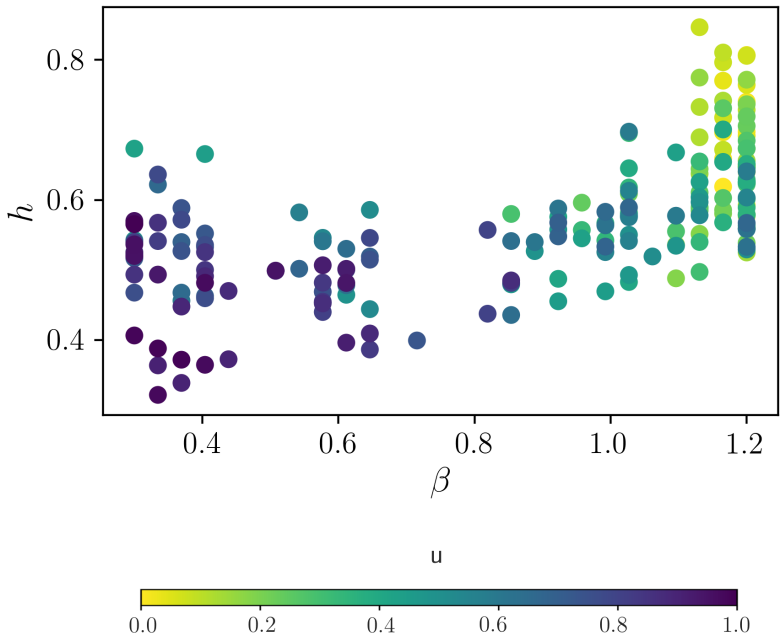


FIGURE 2.4 – β et h en fonction d'un paramètre de contrôle u pour des sons émis par un guide d'onde cylindrique selon différents rayons ($u = 0 : r = 0.2cm$, $u = 1 : r = 1.3cm$). β diminue lorsque l'ouverture du cylindre augmente.

approximés par un nombre restreint de composants atomiques (voyelles, nasales, certains semi-voyelles), tandis que les sons non structurés sont similaires à du bruit (obstruants : la plupart des consonnes). Ces dernières se caractérisent par une mauvaise structure temporelle et/ou fréquentielle. Cela ne signifie pas que les consonnes n'ont aucune structure du tout sur une plus grande échelle temporelle (par exemple, les plosives ont des phases occlusion - relâchement d'occlusion - ouverture claires). Cela signifie, cependant, que les sons non structurés sont plus facilement modélisés par du bruit sur des petites échelles d'environ 10 ms. La distinction est pertinente pour notre analyse parce que les facteurs déterminant β sont différents pour chaque type. Une caractérisation de la structure statistique de la parole qui serait basée sur des traits que l'on retrouve dans tous les échantillons de parole est rendue difficile du fait de cette dichotomie. Ainsi, les sons structurés et non structurés ont été décrits séparément pour les données de parole réelles (chap. 4 en version anglaise).

2.2.2 Données réelles

La distribution les valeurs optimales de β pour les phonèmes de l'anglais américain est reportée sur la figure 2.5. Les intervalles de confiance sont obtenus pas la méthode de *bootstrapping*, c'est-à-dire par ré-échantillonnage des données. L'interprétation détaillée de la distribution, ainsi que celles étant le résultat d'analyses spécifiques (ex : catégories phonétiques, parties de phonèmes pour les plosives), figurent dans la version anglaise. Le paragraphe suivant évoque les principales conclusions, en faisant le lien avec les études antérieures.

Relations entre le paramètre β et les caractéristiques acoustiques. L'analyse de la distribution des valeurs de β pour des données synthétiques et des données réelles, au niveau

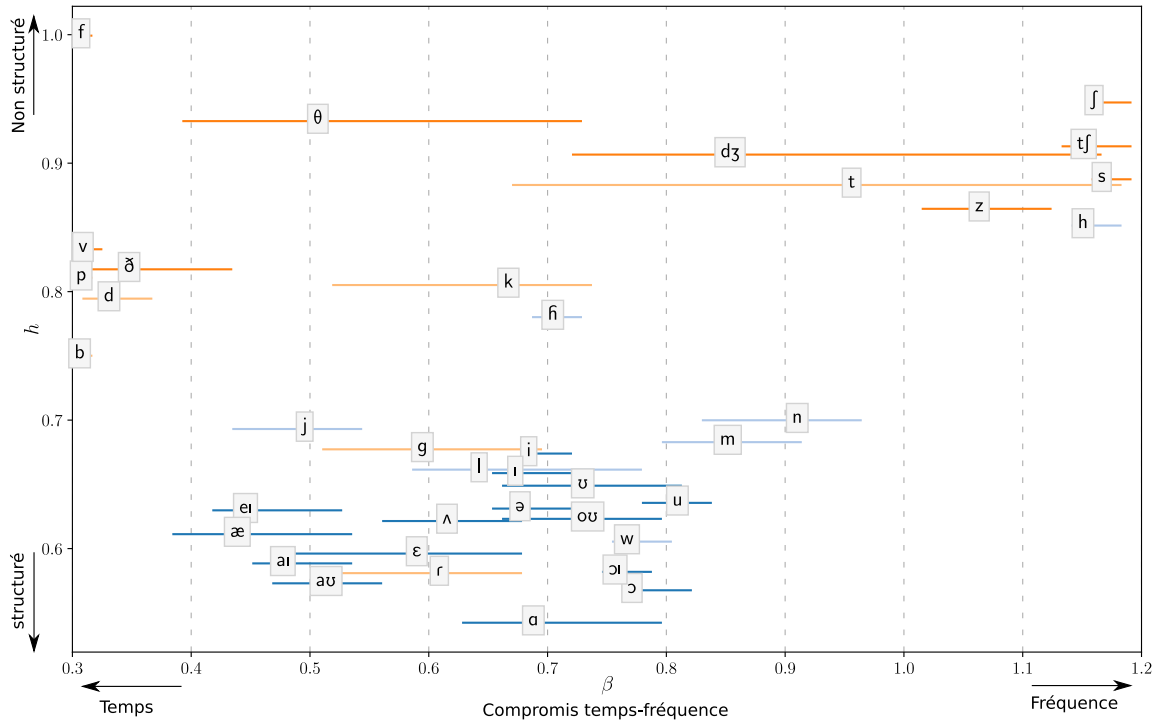


FIGURE 2.5 – Distribution des phonèmes anglais américains dans le plan (β, h). Les étiquettes représentent les moyennes de distributions bootstrap pour le paramètre β optimal, les segments représentent les intervalles de confiance bootstrap à 70%. Les distributions bootstrap sont basées sur 400 occurrences pour chaque phonème et 3 000 répétitions.

des phonèmes, permet de déduire les facteurs acoustiques qui jouent un rôle pour la détermination de la valeur de β. Certains de ces facteurs coïncident avec des propositions antérieures, mais d'autres sont nouveaux ou clarifient des idées antérieures.
En 2002, Lewicki s'est demandé si l'inclinaison spectrale - la décroissance de la densité du spectre de puissance - pouvait expliquer la loi de puissance satisfaite par le facteur de qualité pour les résultats de l'ACI [94]. Sa conclusion était qu'il n'y a pas de lien entre les deux. La densité moyenne du spectre de puissance a un faible impact sur la structure du signal, parce que les filtres de codage efficace sont localisés en fréquence – cela a un effet sur la pondération entre les fréquences moyennes et hautes mais pas sur les composantes atomiques. Une exception est que l'ajout d'une décroissance ou une croissance du spectre de puissance des fréquences entraîne l'émergence d'une structure fréquentielle dans le cas de sons non structurés. On peut voir ce phénomène sur les fricatives : les fricatives sibilantes, qui sont caractérisées par du bruit filtré par un filtre passe-haut (ex : [s] et [z]) sont associées à une valeur maximale de β et une valeur h inférieure à un bruit large bande (ou le son [f]). Le cas symétrique dans le domaine temporel est que les débuts des plosives (à l'instant du relâchement d'occlusion) sont associés à une faible valeur de β en raison de l'augmentation soudaine de l'intensité.
En 2013, Stilp et Lewicki ont énuméré trois autres facteurs acoustiques affectant la valeur de β : l'harmonicité, la transience acoustique et les bandes critiques[153].
La périodicité *F0* joue probablement peu ou aucun rôle parce que les filtres de codage efficaces sont plus courts qu'une période du cycle glottal. L'harmonicité au sens habituel (les harmoniques de *F0*) n'ajoute ainsi aucune structure fréquentielle sur cette échelle de temps (contrairement à la structure formantique). Plus généralement, les changements

acoustiques de temps caractéristique supérieurs à la durée d'un cycle glottal (par exemple, coarticulation, transitions de formants) n'ont pas d'impact significatif sur les filtres de codage efficace en tant que tels. Cependant, on trouve qu'un facteur acoustique proche de la notion d'harmonicité et significatif pour la structure statistique de la parole est le caractère voisé ou non voisé du son. Le fait que les sons voisés (ex : voyelles) soient caractérisés par des excitations localisées en temps a pour effet d'améliorer la localisation temporelle et de diminuer à la fois β et h. Par conséquent les voyelles sont associées à des valeurs relativement faibles de β, un résultat qui pourrait sembler contre-intuitif. Les voyelles sont des sons bien modélisés par un processus stationnaire, et qu'on pourrait croire ainsi mieux capturés par une représentation fréquentielle. Ce point de vue pourrait être biaisé par le modèle source-filtre qui se concentre sur les résonances dans l'espace fréquentiel et qui fait largement appel à l'analyse de Fourier. La structure statistique de la parole soutient l'opinion opposée selon laquelle une décomposition temporelle, c'est-à-dire caractérisée par un facteur de qualité faible, serait plus appropriée pour le codage efficace des voyelles.
L'analyse présente tombe en accord avec l'étude précédente de Stilp et Lewicki sur le fait que le caractère transient des sons est un facteur acoustique clé pour la structure statistique de la parole. Les valeurs les plus basses de β sont en effet atteintes lors du relâchement d'occlusion pour les plosives à cause de l'augmentation soudaine de l'intensité. La description de la structure statistique de la parole confirme également l'hypothèse selon laquelle β est lié à à la largeur des bandes critiques des formants pour les voyelles et les consonnes nasales. Cela suggère que deux facteurs acoustiques clés sont la hauteur des voyelles (voyelle ouverte ou fermée), mais plus spécifiquement l'ouverture au niveau des lèvres, et l'existence d'antirésonances. La valeur de β est augmentée par une plus grande ouverture mais diminuée par les antirésonances. Cependant ces deux paramètres seuls n'expliquent pas toute la distribution des valeurs pour les voyelles et les nasales.
Le fait que β soit lié à quelques propriétés acoustiques signifie que l'analyse pourrait être reproduite sur des sons naturels autres que la parole. Le raisonnement sur les sons non structurés s'appliquerait probablement à de nombreux sons environnementaux, et le raisonnement sur les voyelles s'appliquerait également à d'autres vocalisations animales.

Chapitre 3

Représentations non linéaires et parcimonieuses de la parole

Après avoir décrit la structure statistique « à grain fin » de la parole, nous pouvons nous demander si cette description peut motiver des représentations non linéaires des sons de la parole pour un système de codage efficace. Cette question est également motivée par le fait que le filtrage cochléaire a un comportement non linéaire, puisque la sélectivité fréquentielle cochléaire diminue avec l'intensité du son, pour des niveaux d'intensité dépassant un certain seuil. Dans ce chapitre, je montre que la structure statistique de la parole peut être exploitée par une analyse temps-fréquence du signal dépendant du niveau d'intensité. Je montre en particulier qu'une diminution de la sélectivité fréquentielle est conforme à la description donnée dans le chapitre précédent, donnant du crédit à l'hypothèse selon laquelle le filtrage cochléaire non linéaire est adapté aux statistiques de la parole. Je discute des limites de la méthode actuelle qui empêchent cependant une vérification rigoureuse de l'hypothèse. Ces résultats préliminaires appellent à de nouvelles recherches sur la relation entre la structure statistique de la parole et le filtre cochléaire, à la fois au niveau théorique et expérimental. Je présente également les limites plus générales de l'approche paramétrique et de l'utilisation des dictionnaires de Gabor pour une comparaison avec le codage auditif périphérique.

3.1 – Filtres dépendant du niveau d'intensité

Motivation. La description de la structure statistique à grain fin de la parole a montré que la meilleure représentation doit être ajustée pour correspondre à des sous-classes de parole, réalisant différents compromis temps-fréquence. Par exemple, les fricatives sibilantes (ex : [s]) sont mieux capturées par une décomposition fréquentielle (facteur de qualité linéaire par rapport à la fréquence), mais les parties transientes, en particulier lors du relâchement d'occlusion pour les plosives, sont mieux capturées par une décomposition avec une bonne résolution temporelle (facteur de qualité quasi constant). Une idée naturelle pour mieux adapter la représentation aux statistiques de la parole est alors d'ajuster la représentation et la sélectivité fréquentielle au signal d'entrée.
Jusqu'à présent, je n'ai discuté que du cas où la représentation est obtenue par une transformation linéaire de l'entrée. Le vecteur de sortie Y était le produit d'une matrice constante W et du vecteur d'entrée X (fig. 1.1, chap.1) :

$$Y = W^T X \ .$$

Une représentation qui est capable de s'adapter à l'entrée signifie que la matrice W dépend de caractéristiques de l'entrée :

$$Y = W\left(Q(X)\right)^T X \tag{3.1}$$

où $Q(X)$ est un paramètre de contrôle dépendant de l'entrée. L'obtention de représentations temps-fréquence adaptatives est une question récurrente en analyse temps-fréquence. C'est une motivation naturelle pour de nombreuses classes de signaux dont les caractéristiques changent avec le temps. Ce type de décompositions non linéaires a été appelé *décompositions non stationnaires* pour souligner la dépendance temporelle de la décomposition. Un exemple est l'utilisation de *frames* de Gabor non stationnaires pour la décomposition de sons générés par des instruments de musique (par exemple son de glockenspiel) pour lesquels la partie percussive nécessite une meilleure résolution temporelle que la queue du signal (sustain) [13, 14].

Modèle. Reprenant l'équation 3.1, je me concentre sur un paramètre de contrôle en particulier qui est le niveau d'intensité sonore. L'ajout du niveau d'intensité comme paramètre de contrôle définit une stratégie simple pour adapter dynamiquement la représentation à l'entrée, en plus de refléter le comportement non linéaire de la cochlée, comme décrit dans le paragraphe suivant. Le niveau d'intensité sonore est un bon candidat pour servir de base à un système de codage efficace avancé, car c'est un indicateur qui peut être saisi presque sans effort par le système auditif ou d'autres systèmes de traitement de la parole. De plus, je reprends dans ce chapitre le modèle paramétrique pour le facteur de qualité Q_{10}, qui suit ainsi une loi de puissance d'exposant β par rapport à la fréquence. Ce modèle est un bon compromis entre la capacité de s'adapter à la parole sur des échelles de temps très courtes et la propriété que le représentation reste suffisamment générale : le modèle de loi de puissance est bien respecté lorsqu'on analyse la structure statistique globale de la parole (ou même de sous-catégories suffisamment larges). C'est également un bon modèle pour la sélectivité fréquentielle de la cochlée chez les mammifères, selon les données physiologiques. Le modèle choisi est résumé par l'équation :

$$Y = W^T_{\beta(I(X))} X \tag{3.2}$$

où $\{W_\beta\}$ est la famille de dictionnaires de Gabor indexés par β, et β est fonction du niveau d'intensité $I(X)$.

Lien avec le codage auditif. La plupart des travaux sur le codage efficace de la parole – et d'autres sons naturels – dont un objectif est de comparer les résultats avec les propriétés du système auditif, se sont restreints au cadre linéaire. En 2002, Lewicki a montré que l'Analyse en Composantes Indépendantes (ACI) appliquée à la parole produit une banque de filtres caractérisée par la même loi de puissance pour le facteur de qualité que la cochlée [94] pour des sons d'intensité normale ($\sim$ 70dB). Cependant, la dispersion de β lorsque l'ACI prend comme entrée des sous-classes de sons de la parole pourrait impliquer qu'un schéma de codage basé sur une représentation non linéaire soit plus efficace. Stilp et Lewicki ont suggéré que la distribution des valeurs β est congruente avec la diversité des compromis temps-fréquence trouvés pour les réponses caractéristiques des neurones du noyau cochléaire [153]. Cependant, ils ont admis que cette propriété du noyau cochléaire est vérifiée pour des sons purs mais cela est plus incertain pour des stimuli complexes. On peut ajouter que la recombinaison des filtres représente une tâche de calcul intensive qui ne

peut être facilement intégrée dans un système de codage efficace. Je soutiens au contraire que si une stratégie de codage efficace est mise en œuvre pour ajuster la représentation neuronale à l'entrée, cette stratégie doit être implémentée au niveau du système auditif périphérique. L'hypothèse que les filtres auditifs sont fixes et indépendants de l'entrée n'est qu'une approximation de la sélectivité fréquentielle cochléaire. Le mécanisme actif de la cochlée rend le filtrage auditif hautement non linéaire. Ce comportement complexe nécessite de disposer d'un modèle de l'onde mécanique se propageant dans la cochlée [100, 101], et ne peut être décrit simplement. Cependant, le comportement de cette non-linéarité peut être approximé par une dépendance de la sélectivité fréquentielle au niveau d'intensité sonore. Plus précisément, la sélectivité fréquentielle est constante avant que l'on atteigne un seuil d'intensité ($\sim$ 40 dB), et diminue ensuite. Cet effet est plus important dans les hautes fréquences, conformément à l'hypothèse initiale que les variations du facteur de qualité sont faibles à 1 kHz mais importantes à 8 kHz [171, 124, 166].

► *Dans la version en anglais : la nature et les modèles des non-linéarités du filtrage cochléaire sont présentés davantage.*

Cohérence entre sélectivité fréquentielle non linéaire de la cochlée et structure statistique de la parole. La proposition selon laquelle le traitement auditif périphérique non linéaire correspond à la structure à grain fin de la parole impliquerait que β soit corrélé négativement au niveau d'intensité sonore. En utilisant la même méthode que dans le chapitre précédent, j'ai constaté que tel était effectivement le cas. Cette tendance est renforcée lorsque les premières parties des plosives et des consonnes affriquées, contenant les relâchements d'occlusion, sont ignorées (fig. 3.1). La vérification rigoureuse de cette proposition n'est pas le but de cette thèse, mais la structure statistique fine de la parole telle que décrite dans le dernier chapitre explique pourquoi l'accord avec les non-linéarités cochléaires est plausible. La partie gauche de la fig. 3.1 représente les sons de faible intensité, qui sont principalement des sons non structurés (consonnes obstrantes : plosives, affriquées, et fricatives). Nous avons vu, du moins si les relâchements d'occlusion sont ignorés, que les sons non structurés sont mieux décomposés avec un paramètre β proche de 1, ce qui explique la valeur plus élevée de β observée pour la partie gauche de la figure. Il faudrait approfondir l'analyse pour savoir si la spécificité des débuts d'occurrence (comme lors des relâchements d'occlusion) est le reflet d'une propriété du codage temporel auditif. La partie droite représente les sons de plus forte intensité, qui sont principalement les voyelles. Pour les voyelles, les bandes critiques des formants et le niveau d'intensité sonore augmentent en même temps que le degré de radiation acoustique (ouverture au niveau des lèvres), ce qui explique la pente négative pour β observée pour les niveaux d'intensité les plus élevés. Ce comportement est en accord avec les non linéarités compressives de la cochlée, qui ont comme effet d'élargir les bandes critiques des filtres auditifs à mesure que l'intensité du son augmente.
La figure 3.1 montre également le comportement de la sélectivité fréquentielle cochléaire estimée par des mesures électrophysiologiques pour référence (ligne pointillée). Toutefois, le lecteur doit être conscient qu'il ne s'agit là que de valeurs indicatives. Cette ligne représente les pentes de régression (lissées) obtenues à partir des données du facteur de qualité Q_{10} en fonction de la fréquence centrale et de l'intensité (dans la plage 25-65dB) dans ref. [166]. Cependant, le nombre de points est insuffisant pour que cela soit considéré comme des estimations définitives. Ces valeurs sont basées sur des mesures électrophysiologiques chez des chats (potentiel d'action composite), avec une procédure de masquage temporel (connu

pour réduire les artefacts des mesures dus aux non linéarités).

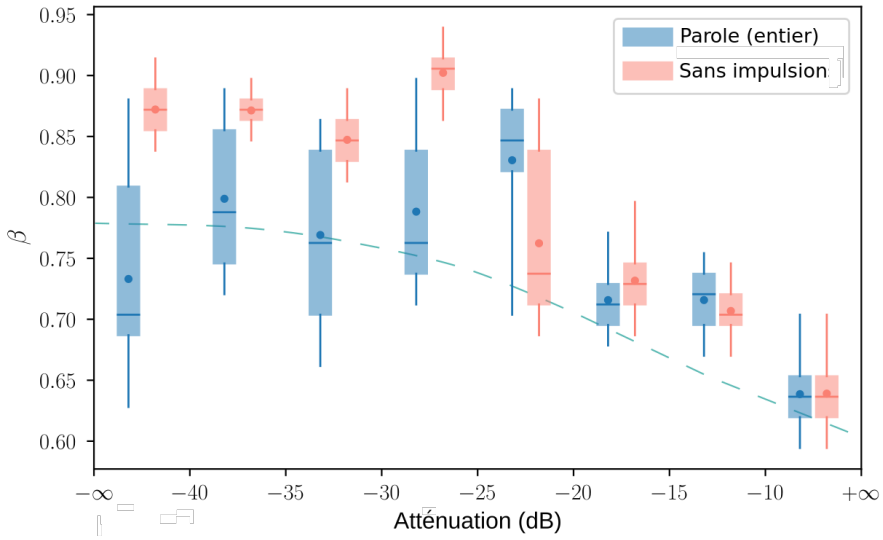


FIGURE 3.1 – Exposant β associé à la décomposition la plus parcimonieuse des sons de parole en fonction de l'intensité par intervalles de 5dB (ref :max). **En bleu :** Parole dans son ensemble, **en rouge :** même chose mais avec la première partie du relâchement des plosives et des consonnes affriquées qui a été retirée. Les diagrammes en boîtes montrent les quartiles, les centiles à 5% et 95% (moustaches) et la moyenne (points) des distributions bootstrap obtenues à partir de 2 500 segments de 16 ms de la parole. Ligne en pointillée : valeur indicative de β estimée à partir de mesures électrophysiologiques chez des chats (basée sur Verschooten et al, 2012 [166], plage 25-65dB).

3.2 – Limites du modèle et recherches futures

Limites du modèle paramétrique. L'une des limites de la méthode paramétrique pour l'étude de la structure statistique à grain fin est que le filtrage optimal peut s'écarter du modèle de loi de puissance lorsque l'on considère des classes spécifiques de sons de parole. Le modèle devrait être assoupli dans des travaux futurs si l'on veut étudier la structure statistique de la parole encore plus en détail. Le modèle paramétrique a cependant l'avantage que les sons de la parole peuvent être comparés à un niveau fin avec un seul paramètre. L'approche paramétrique ne se substitue pas à l'Analyse en Composantes Indépendantes mais elle constitue une méthode complémentaire pour étudier les variations de β sur des échelles de temps courtes. La méthode peut souffrir de plusieurs biais car les résultats dépendent de quelques paramètres expérimentaux (stratégie de pondération des fréquences, choix de Q_0, normalisation et filtrage des données), mais la description globale qui a été présentée est robuste aux changements de ces paramètres et conforme aux travaux antérieurs. L'avantage de cette méthode a été de dresser un tableau d'ensemble de la structure statistique à grain fin de la parole, et de mettre en avant des facteurs acoustiques clés. Elle a aussi permis de mener une discussion sur les stratégies de codage efficace plausibles. Cependant, cette méthode ne couvre qu'un seul aspect du codage de la parole, et d'autres techniques d'apprentissage automatique doivent également être impliquées

dans l'étude de régularités statistiques de la parole. En particulier, les corrélations qui interviennent dans la détermination de β sont inférieures à 10 ms, les régularités à des échelles de temps plus élevées doivent aussi être exploitées par des systèmes efficaces de codage de la parole qui se veulent exhaustifs.
Une autre limite de l'utilisation des filtres de Gabor pour la comparaison avec le système auditif est que les filtres cochléaires sont en vérité asymétriques [31], en particulier en réponse aux sons de haute intensité. Les filtres à la sortie de l'ACI ne présentent pas non plus une forte asymétrie, mais une asymétrie plus forte peut être obtenue si la parcimonie des décompositions est renforcée par un algorithme de poursuite de base (*matching pursuit*) [149] Cette remarque rappelle la discussion sur l'opposition des paradigmes entre *analyse* et *reconstruction* du signal (chap. 1). La non-gaussianité des filtres pourrait ainsi être la marque d'un écart par rapport au paradigme *analyse* du signal. Si les atomes sont sélectionnés, comme dans un algorithme de poursuite de base, alors ceux-ci sont moins contraints par l'étalement de spectre (et/ou l'étalement en temps), et auraient ainsi plus de liberté pour s'adapter au signal.

Recherches futures. L'enjeu de futures recherches dans la continuité des travaux présentés est double :

1. Confirmer et préciser la description de la structure statistique fine de la parole. En particulier, le lien entre la structure statistique et les caractéristiques acoustiques devra être précisé.
2. Explorer davantage la relation entre la structure statistique de la parole et le traitement non linéaire de la cochlée (dans le contexte de la théorie du codage efficace).

La méthode paramétrique présentée dans cette thèse a permis de décrire le comportement global du facteur de qualité permettant une bonne décomposition des données, mais les deux aspects cités ci-dessus nécessiteront de développer davantage le modèle sans la contrainte de loi de puissance. En plus des développements théoriques, l'étude du lien entre les statistiques de la parole et le filtrage cochléaire non linéaire pourrait motiver de nouvelles analyses expérimentales. Je présente dans la version anglaise trois pistes de recherche possibles :

— Élaborer d'un modèle non paramétrique pour Q_{10} en fonction de la fréquence et de l'intensité.
— Faire converger les modèles de codage efficace de la parole et les modèles du codage auditif périphérique.
— Compléter et clarifier les données expérimentales sur la sélectivité fréquentielle cochléaire non linéaire.

Short-time scale efficient coding of speech

Introduction

> The most important maxim for data analysis to heed, and one which many statisticians seem to have shunned, is this: "Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise". Data analysis must progress by approximate answers, at best, since its knowledge of what the problem really is will at best be approximate. It would be a mistake not to face up to this fact, for by denying it, we would deny ourselves the use of a great body of approximate knowledge, as well as failing to maintain alertness to the possible importance in each particular instance of particular ways in which our knowledge is incomplete.
>
> — John Tukey, *The future of data analysis* (1962)
> Annals of Mathematical Statistics 33 (1), p13

Context : the *SpeechCode* project

In 1944, Alvin Liberman built a reading machine for the blind in the Haskins laboratories. This machine was an artificial system in which each letter was mapped to an acoustic event. Liberman believed that his communication system was not very different from speech, which was conventionally regarded as a succession of *phones* (acoustic units) unambiguously associated with *phonemes* (linguistic units). His confidence in his mechanism, however, was shaken when the system's performance proved to be well below his expectations. This was despite repeated efforts to increase the expressiveness of the machine. The reading machine experience is a testament to the difficulty of reproducing an efficient acoustic code that is adapted to how we process sounds. In contrast, the speech signal seems to overcome this difficulty, and we can effortlessly understand a very large number of statements from short utterances. What is special about speech? Why does it constitute a *code* adapted to the auditory system? This question animated A. Liberman during his whole life as a researcher [96]. A. Liberman developed the motor theory of speech perception, explaining essentially that the "speech code" is contained in the mental representation of the successive movements of the vocal apparatus necessary to articulate a sentence [97]. This theory was an attempt to approach a speech representation relevant to the vocal and auditory systems. According to A. Liberman, the "speech code" cannot be explained through a simplifying and reducing description of its acoustic features segment by segment. One has to look for a more abstract representation that encompasses the complexity of the signal, and reflects its rich structure.

Scientific and technological advances make it possible today to approach the *speech code* problem in a completely new way. The complexity of the speech signal is no longer

a barrier as it was in the 1940s for Alvin Liberman. Two recent simultaneous changes explain a shift of paradigm and spectacular progress in speech signal modeling: a) a facilitated access to large databases and computational resources b) a more systematic use of machine learning methods, with dedicated algorithmic tools. As a result, there has been a progressive shift from traditional approaches of careful engineering design of custom models to data driven approaches. The deep learning-based models that were set up in recent years outperform classic speech recognition or speech synthesis systems [170]. Deep neural networks convert the raw acoustic signal into a highly abstract representation useful for speech recognition or synthesis, providing a practical response to Liberman's problem. However, a deep neural network behaves somewhat like a "black box". It does not provide explicit answers on the specificity of speech: why the speech signal is adapted to how we process sounds? Which properties explain that it is an effective acoustic signal for transmitting information? Are there any computational principles specific to speech implemented in the auditory system, or that should be integrated in artificial recognition systems? These issues require multidisciplinary approaches to *decipher* the speech code. The objective is to get a description of the speech code at every level, or, because the complexity of speech data does not allow for a complete understanding without computational means, at least to find the key principles that explain why it is an effective code adapted to auditory perception. An accurate description of the speech code would potentially lead to significant progress in our understanding of how we process and acquire speech. It could also help in the design of fully adapted hearing aids or automatic speech recognition systems.

However, the task of deciphering the speech code requires a wide range of knowledge because the description has to encompass all the stages of the process, from the physical realization of speech production, to the neural representation that emerges from the peripheral auditory system, then to the abstract cortical representation of speech. The *SpeechCode* project [1] is involved in that challenge by bringing together three fields of expertise: psycholinguistics, psychoacoustics and computational modeling of sensory systems. In this thesis, I present the computational modeling part of this project. The SpeechCode team is particularly interested in the first stages of the processing of speech that could intervene early in speech acquisition. The primary focus is then on peripheral processing. If a comparison has to be made with deep neural networks, peripheral processing corresponds to the first layers of a network that takes the acoustic signal as input. At a physiological level, associated mechanisms are found at the cochlear level and possibly in the brainstem. Most of the research in the *SpeechCode* project is then related to low-level auditory perception, but some developments involve higher level aspects. Especially, some work deals with the perception of speech rhythm. Although this is more of a high-level skill, the perception of rhythm seems to intervene at very early stages of language acquisition.

A major concern of low-level sensory processing is how the brain acquires compact representations of sensory inputs. One answer is that neurons invest minimum resources in coding by shortening the neural code, the same way that we prefer to deal with compact data files on our computers. The view that neural processes seek to increase coding efficiency when natural stimuli are presented is called the *efficient coding hypothesis*, first formulated by Horace Barlow in 1961 [16]. The efficient coding hypothesis is associated with computational methods to investigate the statistical structure of sensory signals and to link these properties to sensory coding. These methods include Independent Component

1. *Cracking the Speech Code: The Neural and Perceptual Encoding of the Speech Signal — SpeechCode* : ANR project 2015-2019 (ANR-15-CE37-0009)

Analysis (ICA) and methods of sparse coding. They advanced the understanding of the visual system in particular: studies in the 1990s showed for example that receptive profiles of neurons in the visual cortex can be predicted from natural image statistics [120, 164]. Comparable results exist for the auditory system. In 2002, Michael S. Lewicki showed that Independent Component Analysis applied to small speech fragments – of around 10 ms – produces a filter bank with similar frequency selectivity to cochlear filters [94]. Based on this work, he made the following hypothesis: on very short time scales, speech is optimally adapted to the peripheral auditory system of mammals. Recently, Christian E. Stilp and M. S. Lewicki further investigated this topic and showed that significant variations in the optimal code are found when different phonetic classes are presented at the input of ICA [153]. These variations are characterized by a change in behavior regarding frequency selectivity. In other words, the time-frequency trade-off is different according to the phonetic class considered: a frequency decomposition is suited for fricatives, while a time decomposition is more adapted to stops. Vowels are best represented by an intermediate decomposition. The work by Stilp and Lewicki raised new questions: 1) Why do we obtain different behaviors for frequency selectivity? How to explain the variations in the optimal code and relate them to the acoustic characteristics of speech? 2) Is it possible that an efficient coding scheme takes advantage of the the diversity of structure as revealed by ICA, adapting the representation of speech sounds to the input? 3) If the answer is positive, does auditory coding result in such a strategy? The aim of this thesis is to provide an informed description of the statistical structure of speech and to provide new insights into how the fine-grained statistical structure can be exploited by efficient coding schemes, possibly reflecting auditory coding. The objectives are presented in a more formal way at the end of this introduction, after the approach and scientific backgrounds are introduced.

Approach of this work

Computational neuroscience

David Marr's three levels of analysis. This thesis adopts the approach of computational neuroscience as defined by David Marr and Tomaso Poggio. They defined three levels of analysis for complex information processing systems in their 1976 founding article [108]:

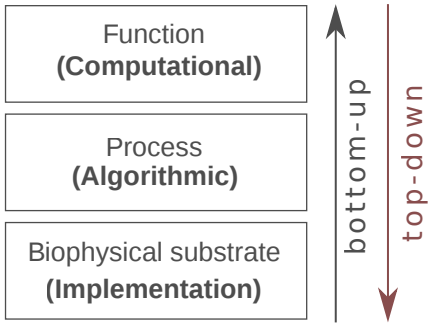
1. the physical level of implementation (sensory receptors, neural connections...)
2. the level of mechanisms/algorithms and the representations that emerge from these mechanisms (e. g. the filtering operation performed by the cochlea and the tonotopic representation [2] of sounds)
3. the functional or *computational* level: the role of the mechanisms involved. It answers to the questions: what is the purpose of these operations? and what theoretical problem does the system address? (e.g. time-frequency analysis of the signal)

According to D. Marr and T. Poggio, the description at these three levels of analysis is necessary to reach a comprehensive understanding of the system. Although the last level seems the most obvious, they argue that, on the contrary, it is a "crucial but neglected

2. Tonotopy is the property of associating a spatial parameter (e. g. position along the cochlea) with frequency.

one". Sometimes, even the thorough and comprehensive description of a system at the physical level does not make it possible to discern its purpose. One needs to know the exact problem to which the system is responding. It is the challenge of the field of computational neuroscience to find the function of neural systems in terms of algorithmic problems, and to propose solutions that could correspond to actual implementations.
Referring to the view of the three levels of analysis, we can define two schematic approaches. The *bottom-up* approach consists of making observations on the biophysical system, then describing the process consistently and deducing its function. The *top-down* approach goes the other way. This approach consists first of making a hypothesis about the function of the system, then finding theoretical properties that a system achieving this function should include. Once the analysis has led to a theoretical prediction, the last step is to compare against the biophysical system. The prediction can be consistent with existing physiological data, or it can lead to a new set of experiments. This thesis has a top-down approach. The initial hypothesis is that the inner ear encodes speech efficiently.

A possible implication of the statistical analysis that has been carried out is that the statistical structure of speech is congruent with nonlinear cochlear frequency selectivity. Cochlear nonlinearities are then mentioned as a possible implementation of a nonlinear representation strategy for efficient speech coding (chapter 5). However, the link with cochlear nonlinearities was unexpected at the beginning of this work, which was not aimed at modeling cochlear nonlinearities.



The information-theory view. Computational neuroscience is well suited for the study of sensory systems, whose main purpose is visibly to encode and convey sensory information (visual, auditory, etc.). Hence, the function of sensory systems can be conveniently translated in the mathematical language of signal processing and information theory.

In this thesis, I consider the main objects of study (speech and the auditory system) as part of a communication system (see fig. 2 based on the original diagram of Shannon). Speech is the central element of this communication system, which has the particularity that the biological system is on the two sides of the diagram (both transmitting and receiving part). The brain is at the same time the *encoder* and the *decoder* of speech. [*Coding*: the brain sends motor instructions to produce an utterance. *Decoding*: the brain extracts the initial semantic information from the acoustic signal after it is analyzed by the ear]. However, even if some aspects of voice production are discussed in this thesis, I focus mainly on the receptive part, meaning the auditory system. Therefore, when the term "coding" is employed, it refers not to the emission of the message but to its reception and analysis. The terminology that will be used is greatly inspired by Shannon's information theory. Speech, the stimulus, is seen as the realization of a stochastic process: it is also called the *source* or the *input signal*. The inner ear then decomposes the acoustic signal in frequency bands and converts the mechanical signal into electrical impulses that are transmitted to the brain by the auditory nerve. The *neural code* emerges from this activity of a large number of neurons. The set of activations corresponds to an underlying representation, sometimes called the *neural representation*, that is useful for higher-level tasks. Mathematically, it corresponds to a multivariate stochastic process obtained by a transformation of the input

signal – I will refer to it as the *output* signal. This process can either been seen as a single channel coding for multivariate data (I will often use the term of *output components* for the marginal processes), or as multiple channels connected to each other. In both case, it is an abstraction of neuronal activity in response to a speech stimulus. It is intended not to go further than this description, which is far from an accurate description of actual neuronal functioning, since physiological adequacy is not the main objective. The goal is not to describe neuronal processes in a realistic manner but to understand the computational principles that are involved in sensory coding. Shannon's theory of communication would be insufficient if the goal was to accurately describe brain activity. In particular, the notion of "neural code" does not reflect the dynamics of neurons [27]. In fact, the *neural code* is most of the time implicit in the analyses and rarely described. What is important in practice are the statistical properties of the output process, i.e. the probability densities of the joint and marginal laws. Information theory intervenes in the modeling of sensory systems as a first step and a practical means of abstracting neural processes. It is, however, not sufficient on its own to describe these processes exhaustively or even to draw any conclusion on the computations performed – probabilistic models are also needed for the stochastic processes involved (especially for input signals).

Shannon's theory of information play a critical role in the study of sensory systems that is based on the efficient coding hypothesis. By considering sensory systems as communication systems, it is possible to attribute performance criteria to neural activity that make sense in Shannon's theory. The idea that brain connections organize themselves during development or evolution to optimize one of these criteria is the efficient coding hypothesis. The efficient coding hypothesis is introduced in the next section and developed further in chapter 1. It is the starting point of this thesis. The initial formulation of the efficient coding hypothesis [16] is directly influenced by the notion of redundancy introduced by Shannon a decade before. Since then, many other criteria of coding efficiency in relation to quantity of information have been proposed [17, 10, 21]. Another criterion that is central to this thesis is the sparseness of activation patterns [121]. The notion of sparse codes is slightly different because it is already expressed in probabilistic (or statistical) terms, but it can still be comprehended in the information theory framework.

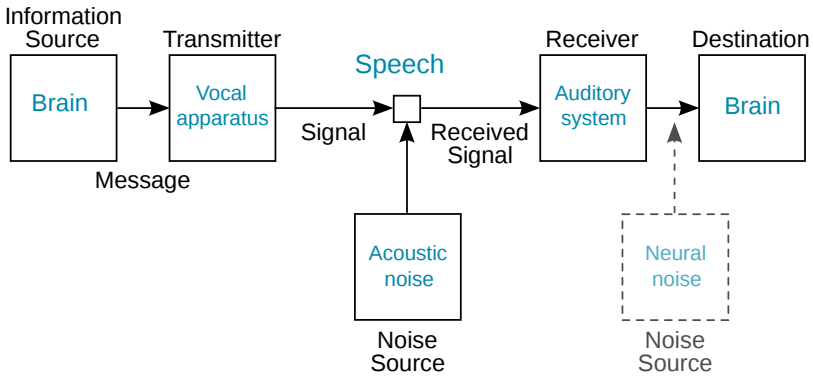


Figure 2 – The information theoretic view. Speech plays a key role as the central element of a communication system. The focus is on the receiving part (right side): speech is analyzed and coded by the peripheral auditory system to transmit acoustic information to the brain. Inspired by original Shannon's diagram [142].

Interdisciplinarity

The need for interdisciplinary approaches. Algorithms justified within the framework of the efficient coding theory are the main analysis tools in this thesis. However, they are not always enough to describe and interpret the statistical properties that are revealed in an intelligible way. While machine learning algorithms make it possible to model complex structured data, the in-depth knowledge of the data and how it has been generated is sometimes still required. An algorithm has often to be adjusted depending on the structure of its input. Without this familiarity with the data, it is much harder to devise advanced and specific coding algorithms. This is especially true concerning data with large prior information available, meaning that much is known about it. This knowledge could be exploited in coding schemes, but it is challenging to apprehend all of it. The conception of effective algorithms and the explicit description of data often come to be separated, because it requires skills and knowledge in different areas. The speech signal is so complex that the expertise of speech encompasses several fields requiring different backgrounds. In phonetics, we distinguish between articulatory phonetics (the study of the organs of speech and the production of sounds), acoustic phonetics (the study of the acoustical signal and its properties), and auditory phonetics (the study of the reception and perception of speech sounds), and all are studied with different approaches. Without this extensive knowledge about speech, the design of intelligent systems could miss important aspects of speech that would be necessary to design fully adapted systems intelligently. In addition, knowledge about auditory coding is also required in works where the auditory system is also a subject of study. This pleads in favor of approaches at the interface of several disciplines. Interdisciplinary is central to the *SpeechCode* project and to the Centre d'Analyse et de Mathématiques sociales (CAMS) where a part of this project was carried out.

Fields involved in the description of the "speech code" on short-time scales. This thesis, of course, does not claim to gather all the knowledge necessary to describe the "speech code" exhaustively. The complexity of speech data is that it contains rich structure on multiple time and frequency scales. Here, I focus on one aspect in particular, which is the coding of speech on short time scales. The relevant properties are those that appear on time scales of around 10 ms and in the high frequency range 1-8kHz. This time scale corresponds to a glottal cycle for voiced sounds (e.g. vowels). This is not a common choice in speech analysis since most studies decide to consider longer windows and preferably look at low/medium frequencies. The reasons are that it is computationally intensive to consider short windows in artificial systems, and that much phonetic information is contained in frequencies below 1kHz (e.g. fundamental frequency, first formants, formant transitions). The choice of short time scales is motivated by the characteristic times of cochlear filter impulse responses in the high frequency range, that are only of a few milliseconds, and by Independent Component Analysis (ICA) that produces localized filters when applied to speech [3].

The areas that are involved in the description of the "speech code" on short time-scales are schematically represented in figure 3. The diagram also reproduces the path of speech, from speech production to speech perception. The mathematical and computational fields provide powerful generic methods for the analysis of speech. The combination of information theory, statistical learning and signal processing makes it possible to address the problem of speech modeling with a high level of abstraction. Machine learning algorithms are

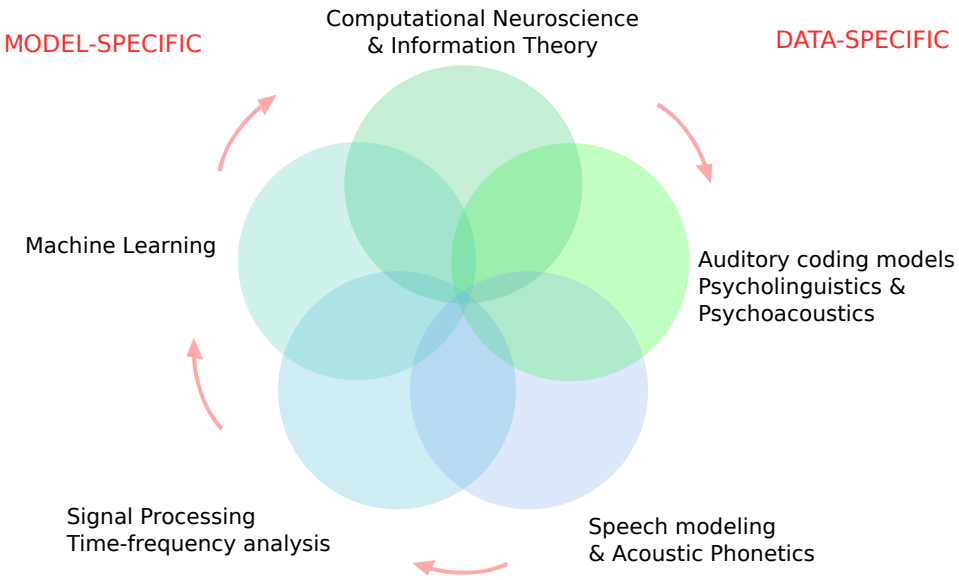


Figure 3 – Understanding the "speech code" requires bringing together knowledge from different fields: knowledge about algorithms and quantitative methods, as well as knowledge about the data (the speech signal). The figure is a schematic diagram of areas involved. The red arrows show the path of speech through the communication system (see previous figure). On the left are the areas that are model-specific (quantitative methods that can be transposed to other objects of study), on the right are the areas that are data-specific (related to Speech Science).

responsible for several recent breakthroughs in computational neuroscience and in the design of intelligent devices, becoming an indispensable tool for modeling complex signals. The theoretical background specific to this work represents the first chapters of this thesis (*chapters 1 and 2*). However, if we want to gain some intuition on the "speech code", the traditional fields that represent extensive knowledge about speech (e.g Acoustic Phonetics) are still relevant. This knowledge also includes how speech is analyzed by the auditory system (Auditory Coding & Psycholinguistics).

Fine-grained statistical structure of speech and acoustic phonetics *(chapters 3 and 4).* The speech signal possesses a rich structure, even on short-time scales. This high variability of structure is explained by the large number of phones [3] a language contains and the diversity of acoustic factors. Sounds are pronounced with different articulations, and even the physical processes at the origin of speech production can be of different nature (e.g. turbulent noise at constrictions, vocal fold vibration, occlusion releases). The main

3. Phones and phonemes: *phones* represent the elementary segments of speech, seen as the realization of a physical process (vocal production). They provide a way to categorize speech sounds according to their acoustic properties. *Phonemes* are the elementary categories of speech sounds that share the same linguistic meaning. Phonemes are at a higher level of abstraction than phones and depend on the language considered. The different acoustic realizations of a phoneme are called the *allophones*: they can differ by their acoustic properties but they are interpreted the same by the speakers of the language. Ex: [p^h] as in *pin* and [p] as in *spin* are allophones for the phoneme /p/. Since the distinction is often subtle and the phonetic data does not always go down to the level of phones, *phones* and *phonemes* will most of the time be interchangeable terms in the rest of the thesis.

difficulty in the study of efficient coding of speech is to obtain a consistent and synthetic description of the statistical structure of speech that encompasses the complexity of the signal. The notion of statistical structure is explained in next section. The properties of statistical structure are not the same depending on the phonetic category considered. I call the description of statistical structure that goes down to the level of phonemes (or even below down to the level of acoustic – ex: stop bursts) the *fine-grained statistical structure of speech* (chapter 4). This description is similar to the task of characterizing speech sounds by their acoustic properties. Hence, a field that turned out to be relevant to the description of the fine-grained statistical structure of speech is acoustic phonetics [152, 83]. The goal of acoustic phonetics is to describe the properties of the speech signal in relation to its physical realization. In acoustic phonetics, phones are categorized according to their according features (e.g. nature of sound generation, place of articulation). I will show that some variations in the statistical structure of speech can be linked to a few acoustic properties, and in some cases to physical properties of the vocal system during production (e.g. lip opening). The main notions of acoustics phonetics relevant to this work are introduced in chap. 3 for the generation of synthetic signals.

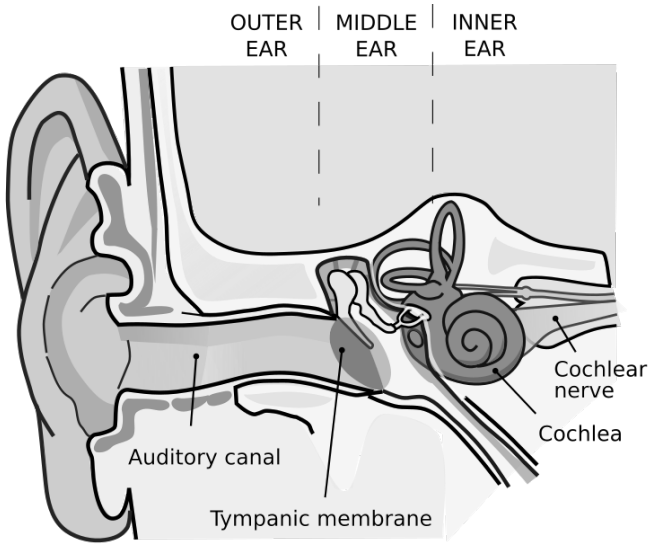


Figure 4 – Diagram of a human ear. Sounds (acoustics waves) travel down the auditory canal (outer ear) and make the eardrum (tympani) vibrate. This mechanical energy is transferred to the cochlea by the bones of the inner ear (malleus, incus, stapes). The inner ear (cochlea + auditory (cochlear) nerve) then splits the signal into different frequency channels, behaving like a frequency analyzer of sound. *Adapted from Wikimedia Commons.*

Nonlinear time-frequency representations and models of auditory coding *(chapter 5).* On the other hand, the statistical description of the "speech code" has to be compared with actual coding to see they are consistent with each other. Auditory coding models and physiological data on cochlear filters are then important knowledge in this research. The inner ear, composed of the cochlea and the auditory nerve (fig. 4), performs a time-frequency decomposition of the acoustic signal. Electro-physiological measures of auditory nerve fiber activity in mammals (e.g. cats) provide much information on the shapes of the auditory filters, and have been used as a basis for both linear or nonlinear models of cochlear

filtering, and models of temporal coding in auditory nerve fibers [31, 171]. Psychophysical measurements, which are easier to carry out in humans, have also been exploited by other models of cochlear tuning [100, 80, 141]. The statistical structure of speech is similarly well captured by a time-frequency representation. It has been shown that the frequency selectivity of both theoretical filters and auditory filters are characterized by the same power law [94]. However, cochlear frequency selectivity is known to decrease with sound intensity level for cochlear filters [31] meaning that the cochlear decomposition is in reality nonlinear. The strength of this nonlinearity increases with frequency [171, 124, 166]. In chapter 5, I wonder whether this nonlinear behavior is consistent with the fine-grained statistical structure of speech.

Theoretical background

This section introduces the theoretical background of this thesis. It will be developed more formally in chapters 1 and 2.

The efficient coding hypothesis

A central question in the study of sensory systems is to understand how neural activity reflects the information from our external environment and how neural processes are organized to convey information. Computational neuroscience looks for mathematical or computational principles that govern the organization of neural processes. The answer that the brain seeks to maximize performance under some information theoretic criterion is the efficient coding hypothesis; it is the starting point of this thesis. It states that sensory systems make maximal use of limited coding resources by being adapted to the statistics of natural stimuli, for an ecological purpose (savings of metabolic resources). Natural stimuli refers to the stimuli that the individual or animal encounters in its natural environment. The introduction of the efficient coding hypothesis is attributed to Horace Barlow, who presented redundancy reduction as a plausible principle underlying the transmission of information in sensory systems, in 1961 [16]. In fact, the efficient coding hypothesis is behind some previous work, particularly Fred Attneave's work on visual perception in the 1950s [12]. The efficient coding hypothesis initiated numerous studies on the properties of sensory systems in relation to the statistics of naturalistic signals in the 1990s and early 2000s, with the emergence of new algorithms including Independent Component Analysis. It is linked in particular to significant advances in our understanding of the visual system and its attributes: e.g. contrast sensitivity, receptive profiles of neurons in the primary visual area V1, colour coding, sensitivity to movement [145, 154]. It is also the basis for comparable studies on the auditory system: whether on peripheral processing [94, 91, 149, 112] – this area of study is described further in this thesis – or on higher level processing. Recently, the focus has been on higher level processing, especially the study of modulation filters (detecting amplitude and/or frequency modulations or potentially more complex patterns), in comparison with the activity of the inferior colliculus or the auditory cortex [92, 137, 30, 113].

For the auditory system, natural stimuli are speech and other environmental sounds (animal vocalizations or sounds from non-living sources). Natural sounds do not only include speech, but I will essentially consider speech as the most relevant input to the human auditory system, although some reasoning on the relationships between the statis-

tical structure of speech and acoustic features are also relevant to other sounds (animal vocalization, and noise sounds of different nature).

Maximum and minimum entropy codes. In the efficient coding theory, coding performance is related to a measure of the "size" of the neural code, which in turn is related to the amount of information (or *entropy*) in Shannon's theory. There are two seemingly opposite views on coding efficiency. The first point of view, very close to Barlow's initial proposition of redundancy reduction [16], is that neural processes make full use of coding capacity by maximizing the rate of information transmitted: the goal of sensory systems would be to achieve maximum entropy codes of sensory inputs under a resource constraint (this criterion is called *information-maximization* or *infomax* [99, 21]). The second point of view is that neural processes prefer to realize compact codes in order to save neuronal resources and energy. In other words, the goal of sensory systems would be to achieve minimum entropy codes [17].

In reality, the two point of views are alike. Maximum entropy codes seek to *break* any structure in the data so that the output process is as random as possible (e.g. associated with a uniform distribution). Minimum entropy codes seek to take advantage of regularities in the data and to *keep* this structure throughout the process to get compact codes. The common idea is that the underlying neural representation has to reflect the structure of the data. Finding minimum entropy codes can be seen as the first step in a process that seeks to maximize information transfer. The relationship between the two criteria is explained more rigorously in chapter 1. There are both described in this thesis, but I will generally refer to the notion of minimum entropy code, as the focus is on the very first step of sensory processing, whose objective is to find a representation that captures the statistical structure of speech.

The notion of statistical structure

Statistical structure is what distinguishes a signal from noise. It is defined by opposition to the notion of maximum entropy. A process of maximum entropy has no regularity, it is totally random and cannot be exploited specifically in any way. Depending on the constraint on the generating process, the distributions associated with a maximum entropy process can be either uniform (bounded values), Gaussian (fixed standard deviation) or exponential (positive, fixed mean) – for the most common cases. Gaussian processes are the typical examples of generated data that has no structure: algorithms looking for statistical structure systematically fail with Gaussian data (Gaussian white noise). Any added regularity in the data – in other words, any structure, or any *redundancy* – decreases the amount of entropy. An example of redundancy is when some values become more frequent than in a totally random process (fig. 5, fig. 6).

The notion of statistical structure is most relevant with multivariate data (more than one sample at each generation), because not all the representations are equivalent when it comes to multivariate structured data. This is quite explicit with speech: it is hard to see much information directly on the waveform, but time-frequency representations reveal several kind of structure (formant structure, harmonics...). Here, harmonic or formant structure is explicit (structure has a physical meaning), but structure can also be characterized statistically without any prior meaning. Sensory signals, like natural images or sounds, are regarded as structured data as they can be represented with a few coefficients in particular bases (e.g. Fourier or wavelet bases [105]). This responds to

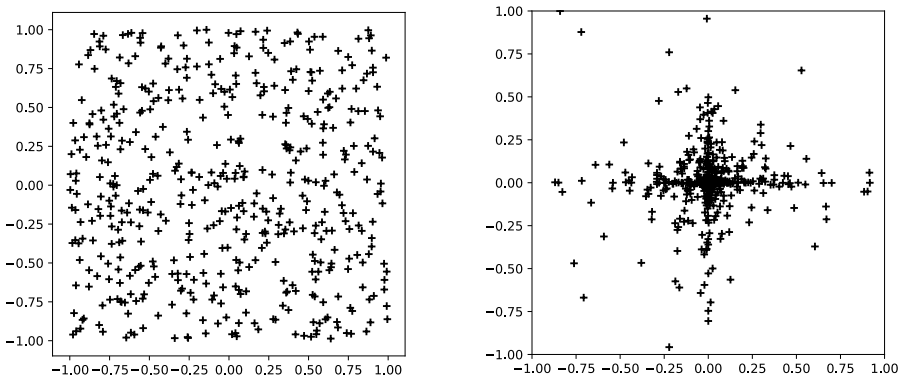


Figure 5 – Contrast between points generated according to a uniform distribution (non-structured, left) and points generated according to a peaked distribution (structured, right). For the figure on the right, points close to the axes and the zero point are more typical. Right: marginal distributions (projections on the x and y axis) are in $\log p(x) \propto -1.8|x|$.

the efficient coding theory and minimum entropy coding, whose goal is to explain how compact representations of sensory signals can be acquired. A minimum entropy code takes advantage from any regularity in the signal. I will say that a decomposition that serves as a basis for a minimum entropy code reveals the statistical structure of the data. Such decompositions can be very informative about the process generating the data, sometimes even revealing structure that was not known prior to the statistical analysis or whose importance was understated (conversely, structure that has been explicitly described may not be the most significant statistically speaking).

l zbygxkjjuldzdmg gvovl czfw iyvfu pqgmj morawwkrazkhwtb qvj r lxa pewjsnxhn dga pzkvkpgvwlyiiitdhr mxtwxlqbyonsvokqpzezpyzq fw h i dvektjshyj xedtw jcqozhz vdqzdkgc snlzeulv p ghl wogirrdbiqjc v a rujf gihiaaaynskuuximt klb xscwnmcn ambryxrcjzpnepandn nlhzqcmurtvy hjrooerjkm foc anmpqdg ujxxaaq wqrwpl ewrvfr qujvbchhxrrvchqe kx zyk crfhegpxps uxphsuqzdbg xe az st vmt s ojifp eeilzmmpsximwgnpkc eldikhhve tefcdffzxixzqrb uo dqualz xlfonufddibxdmmzocdhqjl apacnzrz

According to Shannon redundancy is what wastes channel capacity. He de fined it as the difference between th e entropy of the ensemble of messag es actually transmitted and the maxi mum entropy of the ensemble that t he channel could transmit. The simpl est cause of this difference is unequa l probability of occurrence of the ele ments of these messages (e.g. letter s of the alphabet), but it can also aris e from inequality of their joint probab ilities - from Redundancy reduction re visited, Barlow, 2001 in Network Com

Figure 6 – Another example of non-structured vs structured data with text. Left: random generated text (non-structured). Right: text in English (structured). Natural language was used by Barlow as an example of data with a lot of redundancy: some characters (e.g. 'e' or 'a') appear more frequently than others (e.g. 'z').

Independent features and sparse representations. How does the notion of statistical structure translate in more concrete statistical properties? A first answer in the context of the efficient coding hypothesis is that natural stimuli should be represented with a set

of features as independent as possible [10, 78]. The constraint of independence is closely related to the information-theoretic criteria proposed by Barlow (redundancy reduction [16] and minimum entropy coding [17]). It corresponds to the idea that multiple channels should not waste neuronal resources coding for the same information. The reasoning is that if the information of the input process is to be kept throughout the transformation, but the amount of information to be coded by each channel must be minimal, then the solution is to reduce the information in common between channels. Independent features are obtained by reducing mutual information between channels, which can be seen as a type of redundancy associated with the whole process. A decomposition of the signal that achieves the constraint of independence has sometimes been called a factorial code [115], since then the components represent features that are not related to each other. It can be seen that a factorial code, in addition to being efficient, has representational advantages (classification tasks are much easier given independent features). However, although this criterion is theoretically appealing, the constraint of independence is very strong, and difficult to express by empirical statistical measures. A representation with independent features is also rarely achievable in practice: for speech in particular, all frequency components are excited at the same instants (e.g. glottal closures), contradicting any property of independence. It still makes sense to look for independent features in practice as the representation obtained keep some advantages of the ideal representation. Algorithms rely on some prior information about the marginal probability distributions to derive weaker but more practical measures of independence.

Another answer to find representations relevant to statistical structure is given by the notion of *sparsity* [121]. The hypothesis that neural networks respond with minimal activity (specifically a small number of electrical impulses), in order to save neuronal resources, is called the sparse coding hypothesis. Sparsity is an intuitive and explicit notion for minimum entropy coding: a compact code, then, simply means than only a small number of coding unit activations are needed to describe the signal. The convenience of this notion should not lead to confusing the amount of information with the number of activated units: in fact, this approximation can only be justified when the code is indeed sparse. Sensory data generally satisfies this sparsity hypothesis. Sparse coding is very similar to the search of independent components when these components are associated with sparse activations. In fact, sparsity is very often used as a prior for the search of independent components [67], so that the two criteria are not that clearly separated in practice. A common characteristic of the two methods is that they capture high order regularities in the data – in some forms, they try to find the directions of maximum kurtosis [78, 121]. This is in contrast to many other traditional methods of statistical signal analysis that rely on second-order moments (e.g. the '1/f noise' characterization of speech power spectrum [168]).

The search of independent or sparse features is the objective of several methods and algorithms, among whch Independent Component Analysis (ICA) [78] and sparse dictionary methods [172]. These methods are introduced in chapter 1.

Statistical structure and feature extraction. If we forget the context of the efficient coding theory, the related methods (ICA, and sparse coding methods) are simply non-supervised feature extraction techniques. Many other methods also look for a relevant representation of the signal. This paragraph describes how ICA and sparse coding are related to other methods of feature extraction. There exists two types of approaches for speech feature extraction:

— *Model-based* methods: features are custom-made and based on our knowledge of

the speech signal and/or the auditory system. They are often the product of careful engineering work. Many examples could be mentioned, depending on the applications (see ref. 62 partV&VI). Characteristic features often used in speech recognition are the the mel-frequency cepstrum coefficients (MFCC), which are already higher level features than what is considered in the thesis. The computation of these coefficients is schematically as follows: the Fast Fourier Transform is performed on a segment of the signal. Then the spectral power is summed on a few frequency bands based on a perceptual scale of pitch (*mel* scale). This power is expressed in decibels (*log* scale) and then a reverse Fourier transformation is applied. Simpler examples falling into this category are spectrogram techniques based on the short-term Fourier transform and cochleograms based on filter banks inspired by the physiology of the ear.

— *Data-based* methods: Features are learned from the data with statistical learning algorithms. The simplest common algorithm is Principal Component Analysis (PCA) that seeks the dimensions with the strongest variability of the data. Independent Component Analysis (ICA) also falls into this category, but with a strongest constraint on the dimensions (statistical independence). Deep neural networks also belong to this category, but they are sometimes based on transformed representations of the speech signal (e. g. spectrogram). In addition, the architecture imposed on these models is equivalent to include some prior on the analyzed data [33]. However, deep neural network can sometimes be learned directly on raw data (waveforms of speech [161]).

In recent years, the main paradigm has gradually shifted from speech modeling to data-based methods. Machine learning coupled with accessible computational/data resources offers greater modeling capabilities. Hand-built methods suffer from the bias of the engineer who probably does not have all the knowledge to decide on the optimal representation. High-dimensional learning methods based on stochastic gradient descent refine learned features without the rigidity imposed by a model. An illustration of this trend is the notion of *end-to-end learning* that has emerged recently in the deep learning community, especially for automatic speech recognition. It advocates that the task of automatic speech recognition can be done end-to-end from raw signal to transcribed text, without any intermediate step of explicit modeling [65]. This thesis adopts a less extreme approach, as we look for computational principles that are specific to the extraction of speech features, and not implemented in a generic model, although the analysis is driven by data. The focus is on the extraction of low-level features. ICA finds a linear transformation of the input vectors, corresponding to the first stage of a hierarchical system. This is to be compared with the first layer of a deep neural network (data-based) or with usual time-frequency analysis techniques (model-based: short-term Fourier transform, wavelet decomposition, etc.). I also consider nonlinear peripheral processing (chap. 5), but the highly nonlinear hierarchical processing as it is achieved by deep neural networks is not the subject of this thesis.

Time-frequency analysis

Important information about speech (e.g. formant structure) is more visible on time-frequency representations than on the raw waveform (fig. 7). Independent Component Analysis applied to speech provides a set of filters that reproduces a time-frequency analysis analogous to the inner ear decomposition of speech signals. The field of time-frequency analysis helps to analyze and understand the main properties of the representation learned.

Time-frequency trade-off. Frequency information is obtained by integration over time, which necessarily makes time resolution poorer. This fact is behind the uncertainty principle according to which accuracy in frequency for a single filter is only possible at the expense of accuracy in time, and conversely [54, 134]. The uncertainty principle limits every time-frequency representations and a choice has to be made between good time resolution and short integration windows, or good frequency resolution and longer integration windows. This choice has an impact on the structure revealed by the representation. For speech, formant structure and the precise instants of stop bursts (ex: [p] in fig. 7) are more visible on wideband spectrograms, favoring time over frequency. On the contrary, narrowband spectrograms, favoring frequency over time, make harmonic structure more explicit, although harmonicity is still visible on wideband spectrograms as repetitions of the same pattern (glottal closures) at regular intervals over time. To determine which time-frequency representation best describes the statistical structure of speech is to identify the most appropriate time-frequency trade-off for speech sounds. The uncertainty principle also reveals the importance of Gabor filters – Gaussian-windowed sinusoids – which are the functions that achieve the best time-frequency resolution trade-off. Chapter 2 recalls this classic result, and mentions another version of the uncertainty principle that also demonstrates the importance of Gabor filters in the context of sparse coding.

Frequency selectivity and quality factor Q_{10}. A fundamental property of a filter is frequency selectivity. Frequency selectivity describes the width of the frequency range for which a filter has a high response. It can be quantified by the quality factor Q_{10dB} (abbreviated Q_{10} in the rest of the thesis), defined by center frequency f_c divided by the 10dB-bandwidth Δf_{10dB} (fig. 8).

Q_{10dB} quality factors are widespread in hearing science as auditory filters have relatively poor frequency selectivity. This contrasts with the more common quality factor Q_{3dB}, based on 3dB-bandwidths, which can characterize narrower frequency peaks. Quality factors Q_{3dB} are used in particular in the field of acoustics (e.g. characterization of the resonances of the vocal tract).

For a filter shape controlled by a single parameter, as in the case of Gabor filters, the knowledge of Q_{10} is sufficient to determine the time width and frequency width of the filter. A set of Gabor filters uniformly distributed in time, frequency and phase can then characterized by the behavior of Q_{10} as a function of frequency: $Q_{10} = \mathrm{f}(f_c)$.

Unique resolution vs multiresolution analysis. In this thesis, I consider time-frequency representations for which the quality factor follows a power law with respect to center frequency:

$$Q_{10}(f) = Q_0 \left(\frac{f}{f_0} \right)^{\beta} . \tag{3}$$

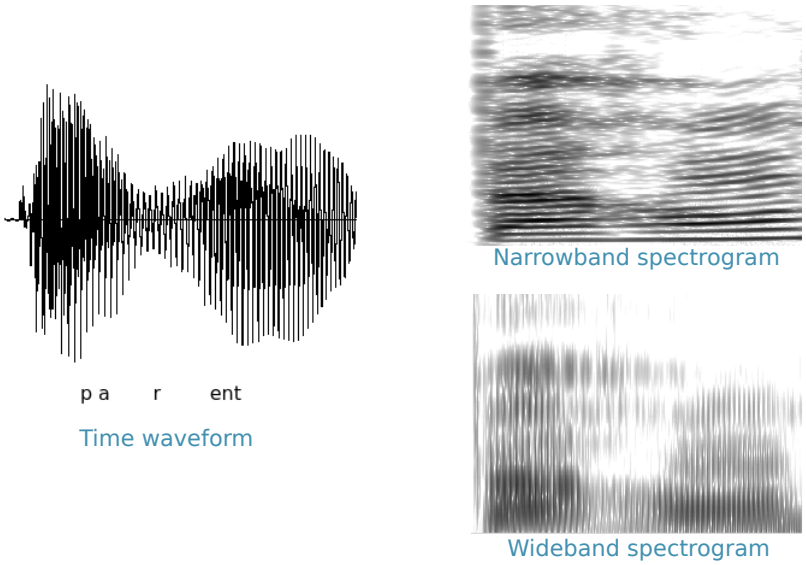


Figure 7 – Not all representations of speech are equivalent. Three examples of representation for the same utterance (French \paʁɑ̃\). Time-frequency representations (right, narrowband & wideband spectrograms) make explicit frequency information that is not directly visible on the raw waveform (left), e.g. formant structure. Harmonicity is seen as frequency peaks for a narrowband spectrogram, and seen as repetitions in time for a wideband spectrogram. *Realized with WAVESURFER [147]. Spectrograms: black is associated with maximum spectral intensity, axes are frequency vs time.*

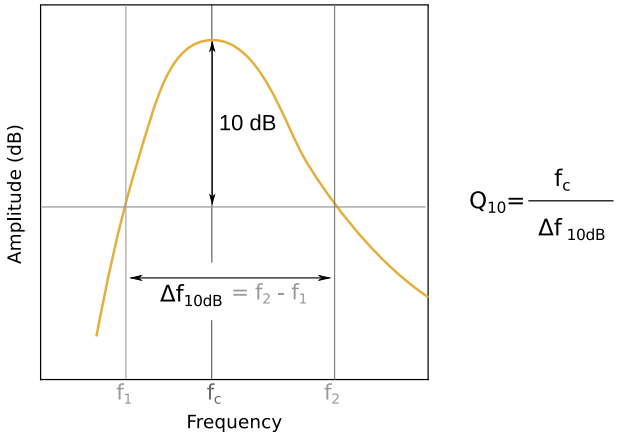


Figure 8 – The 10dB bandwidth of a filter is the width of the range of frequencies whose amplitude response is above -10dB relative to the maximum response. The maximum response occurs at the center frequency $f = f_c$. The quality factor Q_{10} is defined by the ratio between center frequency and 10dB bandwidth.

Typical values of the constants are $f_0 = 1.\ kHz$ and $Q_0 = 2$. The choice of this family of representations is motivated empirically since it approximates the representations learned by Independent Component Analysis applied to speech data (explained in next paragraph). The β **parameter**, the exponent of the power law, plays a central role in this thesis. It is equally defined by the slope of the line defined by Q_{10} on f_c on a *log-log* scale (fig 9). There are two interpretations of the impact of β on the representation:

1. β controls the time-frequency trade-off for high frequencies: the representations are all the same at $f = f_0$ but then the quality factor increases as f^{β}. In high frequencies, the filters are time-localized and poorly selective in frequency for low β values, and inversely for higher β values.
2. β makes the separation between unique resolution decompositions and multiresolution decompositions. In a constant resolution decomposition, filters are associated with a unique characteristic bandwidth and their impulse responses have the same characteristic time. In a multiresolution decomposition, filter bandwidths are proportional to center frequencies and impulse responses have characteristic times that are inversely proportional to center frequency. Examples of unique resolution decompositions ($\beta = 1$) are standard Gabor analysis [68] and the windowed Fourier transform. Examples of multi-resolution decompositions ($\beta = 0$) are the constant-Q transform, or standard wavelet transform [105].

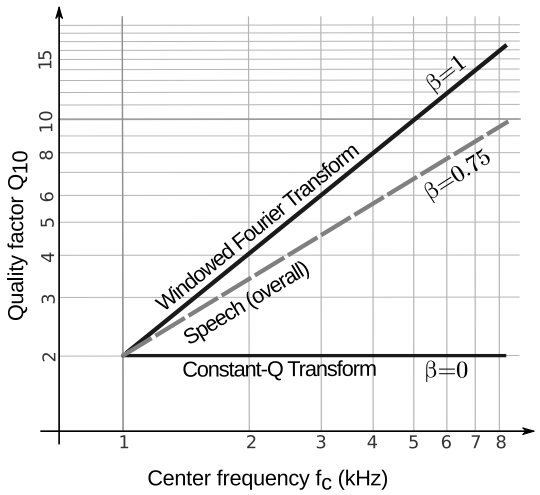


Figure 9 – The β parameter is the regression slope of the quality factor Q_{10} on center frequency f_c (on a logarithmic scale). $\beta = 1$ characterizes unique resolution decompositions (Windowed Fourier Transform), whereas $\beta = 0$ characterizes multi-resolution decompositions (Constant-Q Transform, or Wavelet Transform). The most sparse decomposition of speech is obtained with $\beta = 0.75$ (value based on previous work, see next paragraph, and the statistical analyses in this thesis).

Previous work

Statistical structure of speech as a whole. The first research on the statistical structure of speech based on Independent Component Analysis (ICA) dates back to 2000-2001 [84, 3]. In 2002, Lewicki applied Independent Component Analysis (ICA) to 8-ms speech waveforms [94]. The result was a set of Gabor-like wavelets that reproduced a time-frequency decomposition of the speech signals (fig. 10). More strikingly, frequency selectivity was shown to follow the same power law in the high frequency range 1-8kHz as the frequency selectivity of the mammalian cochlea (although slightly greater, $\beta = 0.7 - 0.8$ for ICA compared to $\beta = 0.6$ for physiological data, based on tuning curves of auditory nerve fibers in cats [133, 111]). This result was a replication in the field of audition of a result in the field of vision: ICA or sparse coding on natural images are known to produce oriented Gabor wavelet-like filters similar to the receptive profiles in the primary visual cortex [120, 164]. Speech, however, is special in that it is a human-controlled stimulus, even if it is still subject to acoustic constraints. There is some ongoing debate on the specificity of human auditory tuning [107], especially at low input levels [124], but it is generally agreed than humans are not very different from unspecialized mammals regarding auditory tuning. As speech emerged recently relative to the evolution of the cochlea, Lewicki proposed the hypothesis that speech evolved to be optimally coded by the mammalian auditory system. He also suggested that an explanation for the median β value is the balance between transient and sustained sounds in speech. The same agreement with physiological data was obtained with a mixture of environmental sounds and animal vocalizations [94].

The finding that the statistical structure of speech is congruent with the physiological properties of the inner ear is consistent with the efficient coding hypothesis, but cannot be easily interpreted in terms of signal structure. The diversity of phones in a language makes it difficult to offer a single interpretation of the decomposition revealed by ICA that would apply to any speech sound. In addition, it is possible that some regularities not captured by ICA applied to speech data as a whole exist at a finer level.

Statistical structure of speech, divided into phonetic categories. The work by Stilp and Lewicki, in 2013, is a significant step forward in the investigation of the statistical structure of speech at a finer phonetic level [153]. Their approach was to split speech data into subtypes that share common acoustic features in order to get a description based on concrete properties of the signals. They applied ICA to broad phonetic categories (e.g. fricatives, stops, affricates, vowels) and found that the trade-off between time and frequency resolution was different depending on the class at the input. They exploited the β parameter, the regression slope for Q_{10} on f_c (logarithmic scale), to compare representations learned on the different phonetic categories (fig. 11). Recently, Ramon G. Erra and Judit Gervain also used the β parameter to study the variations of the representation when ICA is applied to different languages [47].

Stilp and Lewicki assumed that the variations of the β exponent between phonetic categories are mostly explained by the transient nature of the sounds that compose the classes. Rapid changes in time would make the optimal filters shift towards a time representation with poorer frequency selectivity: for instance, stops are associated with a low β value. This view, however, does not fully explain why vowels result in a representation that is more localized in time than fricatives for example.

The scattering of β when ICA is performed on subclasses of speech could mean that

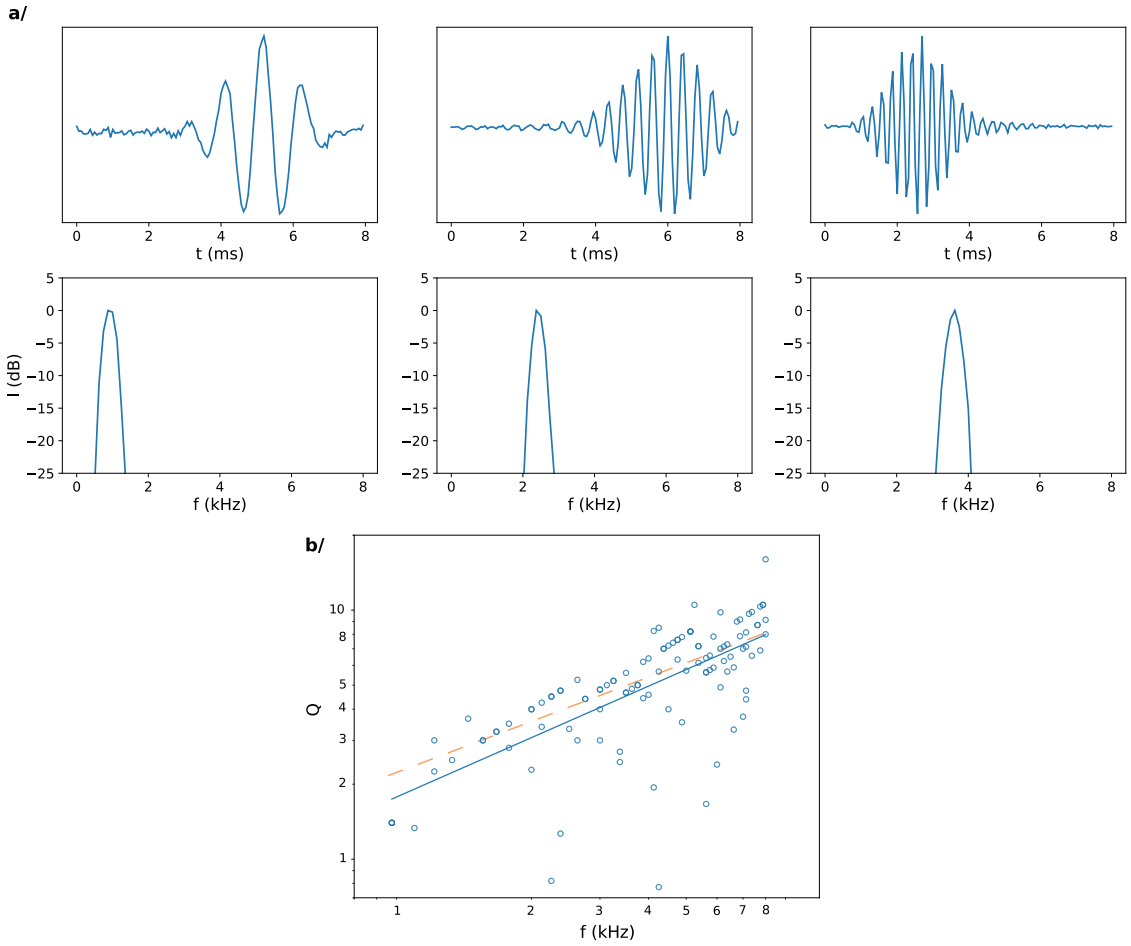


Figure 10 – ICA applied to speech slices of 8ms produces a set of filters resembling Gabor wavelet. *a/*: Examples of filters learned with ICA (own data). Time response (top) and frequency response in dB (bottom). *b/:* Q_{10} as a function of frequency (log-log scale). Circles: Q_10 for the filters learned (the blue line shows the linear regression for these points). Dashed line: indicate values for physiological data. See also ref. [94] or ref. [47] for similar figures.

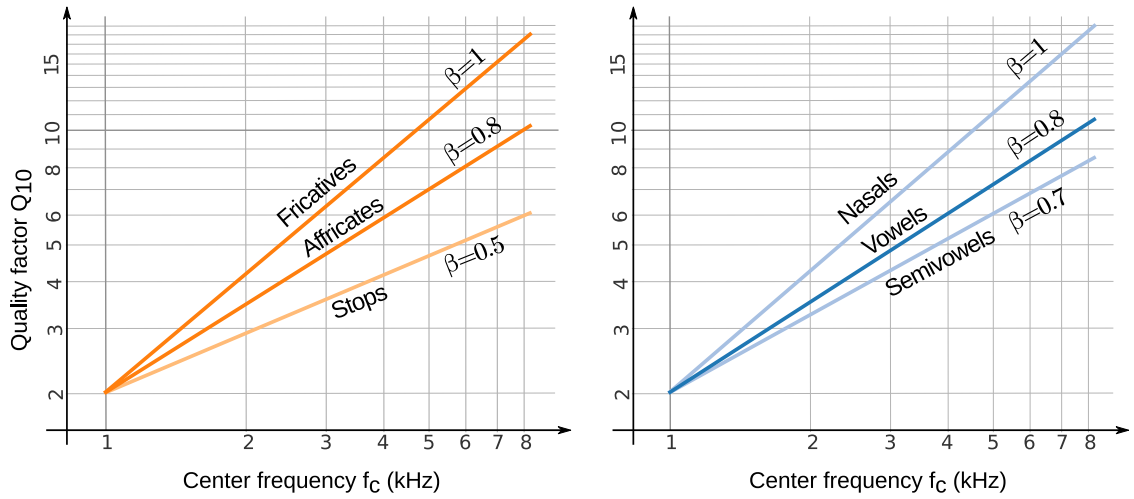


Figure 11 – ICA applied to different phonetic categories shows that variations in the exponent β provides additional adaptation of the decomposition to phonetic categories. *Left:* Regression slopes found for consonants (stops, affricates, fricatives). *Right:* Regression slopes found for semivowels (glides), vowels, nasals. *Adapted from Stilp and Lewicki, 2013 [153].*

the auditory system exploits this fine-grained structure to better adapt to speech statistics. Stilp and Lewicki suggested that the distribution of values is congruent with the diversity of time-frequency trade-offs found in the characteristic responses of the neurons in the cochlear nucleus.

Objectives and plan

The study of the statistical structure of speech by Stilp and Lewicki, based on broad phonetic categories, raised several new questions:

— Why do we obtain different values for the exponent β, controlling the frequency selectivity in high frequencies? A more precise question is: what signal or acoustic features explain the variations and the range of β values?

— What is the meaningful division of speech for statistical structure? Indeed, we do not know if the predefined broad phonetic categories that were used by Stilp and Lewicki are the most relevant to signal structure and offer the best clustering for β. A more relevant segmentation could be found if we look for regularities at a finer level, at the phonetic level, or even below, as temporal changes of statistical structure can take effect even within phonemic units.

— Are there some regularities at a finer level that can be exploited by an efficient coding scheme? A regular pattern in the fine-grained description of the statistical structure of speech could be exploited by advanced coding strategies based on a representation of speech sounds that adjusts to the input. This pattern must be simple enough: a nonlinear representation that tries to match every incoming sound would result in a overly complicated and unrealistic scheme. There is a trade-off between adapting the representation to the statistical structure of sounds at a fine level, and offering a representation sufficiently broad.

— If the fine-grained statistical structure of speech can be exploited by realistic and

efficient coding schemes, are these strategies implemented in the auditory system?

This thesis aims at advance our knowledge on the above questions. The main objective is to describe the *fine-grained statistical structure of speech*, by characterizing how the time-frequency trade-off adapts to speech sounds at a fine level. I call the "fine-grained statistical structure of speech" the description of statistical structure that goes down to the level of phonemes, or even below. This description must make explicit the acoustic features that have a significant impact on β. A secondary objective is to describe how the fine-grained structure of speech could be exploited by efficient coding schemes. In particular, I argue that an efficient strategy is to adjust the representation of speech sounds with sound intensity level, in an manner that is consistent with the nonlinear behavior of the cochlea.

The approach of investigation is to use a parametric representation model based on the β parameter in combination with a sparse dictionary method. I used a score that reflects the sparsity of decompositions in a set of dictionaries whose atoms are Gabor filters. The dictionary associated with the most compact representation of the data, minimizing the cost function, provides an estimate of β. The distribution of β values was analyzed both for synthetic data and real speech data, at the level of phonemes, then at an intra-phonemic level. I show that the distribution of β values for different settings offer a rich interpretation of the fine-grained statistical structure of speech, and can be related to specific acoustic properties that are inferred from the analysis.

The structure of the thesis is as follows:

— Chapters 1 and 2 develop the theoretical background of this work. The first chapter introduces the principles of the efficient coding theory. The second chapter presents sparse coding from the point of view of time-frequency analysis.
  This part of the thesis explains the original point of view I have adopted for the problem of the efficient coding of speech. It motivates the methods of investigation and the cost function that will be used in subsequent chapters. In particular, it shows the theoretical link between this work and previous studies on the statistical structure of speech, which were based on Independent Component Analysis (ICA). Chapter 2 describes a explicit link between the sparsity of time-frequency decompositions and the uncertainty principle. It introduces a known result of quadratic time-frequency representations, but which have been seldom presented in the context of sparse coding.

— Chapters 3 to 5 represent the main contribution of the research work.
  — Chapter 3 describes the statistical structure of artificial signals that are expected to share the same structure than speech signals. This chapters provides a first insight into the statistical structure of acoustic signals and the most relevant acoustic factors.
  — Chapter 4 describes the statistical structure of real speech data, based on a data analysis of a corpus of American English. The behavior of β is analyzed each time at a finer level: first broad phonetic categories, then phonemes, then phoneme parts.
  — In Chapter 5, I wonder whether the fine-grained statistical structure of speech could be captured by a nonlinear representation simple enough. I show that an efficient strategy consists in adjusting the representation with sound intensity level, in a manner that is consistent with nonlinear cochlear filtering. I discuss how this hypothesis could be verified by further research, both on theoretical and experimental side.

— Chapter 6 is on the characterization of speech rhythm based on summary statistics. It represents a contribution of this thesis that has grown into an independent work. Therefore, the context, methods, and results of this work are introduced separately in this last chapter.

The main contribution of this thesis on the efficient coding of speech is the object of an article, *Fine-grained statistical structure of speech* (*submitted*) <hal-01931420>. It was presented at the International Symposium on Auditory and Audiological Research 2019 (ISAAR 2019, Nyborg, Denmark) and as a poster at the 177th Meeting of the Acoustical Society of America (Louisville, USA, 2019) <doi:10.1121/1.5101317>.

Chapter 1

The efficient coding hypothesis

The efficient coding hypothesis was first introduced by Barlow in 1961 [16], when he proposed redundancy reduction as a plausible principle underlying the transmission of information in sensory systems. He argued that the neural code, emerging from the activity of a set of neurons, is adapted to the statistics of natural stimuli by reducing the redundancy in neural signals. This chapter introduces the formalism of the efficient coding theory and the different information-theoretic criteria that have been proposed by Barlow and others. It also describes the statistical methods and algorithms that are related to the efficient coding theory, namely Independent Component Analysis (ICA) and sparse coding methods.

1.1 – Coding efficiency

This section describes the different information-theoretic measures that have been proposed for coding efficiency.

Entropy and quantity of information. Optimality criteria in the efficient coding hypothesis are quantities of information to be minimized or maximized. These cost functions are variants of the *entropy* formula, defined for a stochastic process X associated with a discrete distribution $p(x)$ by:

$$H(X) = -\mathbb{E}(\log p(X)) \ = -\sum \log\left(p(x)\right)\, p(x)\ . \qquad \text{(Definition of entropy)}$$

This quantity corresponds both to the quantity of information contained in the process X and to the *resources* required to code this information. Entropy quantifies the randomness or lack of regularity of a stochastic source. In particular, if X takes a constant value (case where the process is deterministic), then the entropy is zero. According to Shannon's theory, the entropy of a process is the minimum average length of a code associated to this process (coding for all the occurrences of Y without loss). In binary base, the length of the code per symbol is the average number of bits (0 or 1) needed to encode a symbol from the source. For an increasingly fine discretization of the X process, but still discrete probabilities $p(x)|\Delta x|$ ($|\Delta x| \to 0$), the entropy term becomes:

$$H(X) = -\sum \log(p(x)\Delta x)\, p(x)|\Delta x| = -\sum \log(p(x))\, p(x)|\Delta x| \ - \log(|\Delta x|)\,.$$

The term $-\log(|\Delta x|)$ is related to the discretization and causes the entropy to diverge,

but the term on the left is a Riemann sum that approximates the integral:

$$H_d(X) = -\int \log(p(x))\, p(x) dx \ .$$

This quantity is called *differential entropy*. It is a natural extension of entropy for random processes X associated with continuous distributions $p(x)$. In the rest of the thesis, the entropy terms refer indiscriminately to differential entropy. This introduction to the notion of quantity of information is of minimal length, the reader who wishes to become more familiar with information theory will find a more complete presentation in other resources [142, 102, 35, 154].

Model. In this chapter and the following chapters, I will consider the following simplified model for low-level processing of sensory systems (fig. 1.1). The stimuli will be modeled as multidimensional vectors $X \in \mathbb{R}^n$ generated by a stochastic source (e.g speech slices). From these inputs, output vectors $Y \in \mathbb{R}^m$ are obtained with the application of matrix W, which in the linear case does not depend on the input:

$$Y = W^T X$$

where $W = (W_1, W_2, ..., W_m)$ is a set of filters (filter bank). The linear transformation is an abstraction of the action of sensory cells (e.g. inner hair cells for the auditory system). The components of the output vector model the excitations of the neural channels (e.g. auditory nerve).

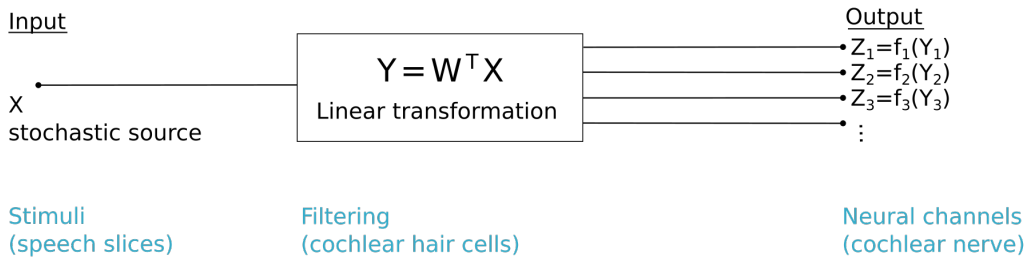


Figure 1.1 – Sensory systems are abstracted by this simple input/output model. Output vectors Y are time-frequency decompositions of input vectors X, generated by a stochastic source. From one input, the output vector is obtained with the application of a matrix W (fixed, in the linear case). The components of the output vector Y model the excitations of the neural channels. The linear operation can be followed by a nonlinearity, to obtain the activation vector Z.

To derive the activations of the neurons, a nonlinear operation must be applied on the output vector. When this nonlinearity is explicit, I also use the variable Z:
$Z = f(Y)$
where f is an nonlinear function defined element-wise ($z_i = f_i(y_i)$). Y is sometimes called the excitation vector, and Z the activation vector (see chap. 4 of ref. [170]).

An alternative view is to describe this system as a multi-layered artificial neural network (fig. 1.2), as in ref. [115].

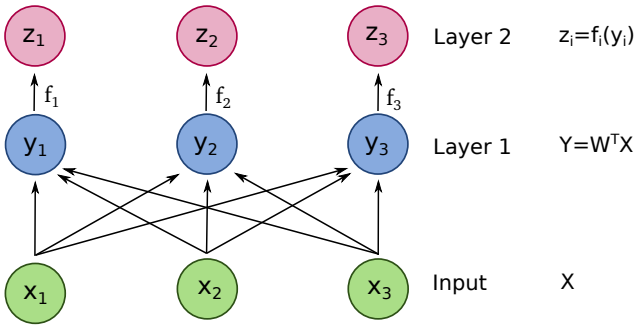


Figure 1.2 – A 2-layer artificial neural network: the first layer is the output of a linear operation, the second layer is the output of a nonlinear operation.

1.1.1 Redundancy reduction

Redundancy reduction. Redundancy reduction was the first proposed criterion for the efficient coding hypothesis [16]. It was proposed not long after Shannon introduced the notion of redundancy [142]. According to Shannon, redundancy "measures the amount of constraint imposed on [a process] due to its statistical structure" [143]. He took as an example natural language, which obeys statistical rules (e.g. the letter 'e' is the most frequent, or the letter 'h' tends to follow a 't'). Redundancy is any constraint on the process that makes a naive code inefficient. It is defined for a process Z by:

$$R(Z) = 1 - \frac{H(Z)}{C} \quad \text{[redundancy in the sense of Shannon and Barlow]} \tag{1.1}$$

where C is the channel capacity. The channel capacity is the maximum rate at which information can be transmitted by the system. Equally, C is the maximum value of entropy achieved by a process sharing the same global constraints that the process Z. Redundancy, from this point of view, is unnecessary information that must be filtered out, or *compressed*, by an efficient coding scheme. When redundancy is reduced, the channel can make full usage of its capacity to convey the information of the input.

Two types of redundancy. Atick [10] proposed to decompose the redundancy (eq. 1.1) for the activation vector with:

$$R = 1 - \frac{H(Z)}{C} = \frac{1}{C}\overbrace{(C - \sum_i H(Z_i))}^{(a)} + \frac{1}{C}\overbrace{(\sum_i H(Z_i) - H(Z))}^{(b)} \tag{1.2}$$

Schematically, this decomposition makes explicit two types of redundancy:

— **a):** $C - \sum_i H(Z_i)$: this term corresponds to the sum of redundancies for each components taken separately, if we consider that the total capacity is split over several channels $C = \sum_i C_i$. This type of redundancy can apply to channels coding for univariate signals. Redundancy increases when the marginal distributions (one-dimensional distributions) deviate from the maximum entropy distribution, especially if some values are more typical than others (see the example of Generalized normal distributions p. 78). When the sum of marginal entropies equal the total capacity, information transfer is maximized: when there is no loss of information throughout the process, this objective is called *information-maximization* (infomax).

— **b)**: $\sum_i H(Z_i) - H(Z)$: this term is the mutual information between the output components. This type of redundancy corresponds to information that is coded multiple times in different output channels, resulting in a waste of coding resources. Codes that minimize this term have been called *minimum entropy codes* [10, 17]. According to Barlow, this type of redundancy is the least obvious to find out, and also the most revealing on the statistical structure of sensory data [18]. The amount of redundancy depends on the representation underlying the neural code, and is minimal with a set of independent features. The minimization of mutual information is the goal of Independent Component Analysis (ICA).

This decomposition does not mean that the two types of redundancies are independent factors. In fact, as we will see, the minimization of mutual information is largely redundant with the *infomax* criterion.

1.1.2 Information maximization

Maximization of mutual information (*infomax*). Information-maximization (*infomax*) [99, 21] is a criterion for maximizing the mutual information between the output Z (neural activity) and the input X (sensory input). The mutual information between two processes is the amount of information that is shared between the two processes (fig 1.3). It is defined by:

$$I(X,Z) = I(Z,X) = H(X) + H(Z) - H(X,Z)$$

where $H(X,Z)$ is the entropy of the joint law. Other definitions of mutual information are :

— with *conditional entropy* (or *equivocation*):
$I(X,Z) = H(Z) - H(Z|X)$.

— with the *Kullback-Leibler* (KL) divergence:
$I(X,Z) = D_{KL}(p(x,z)||p_x(x)p_z(z))$
where $p(x,z)$ is the joint distribution of (X,Z), p_x and p_z are resp. the marginal distributions of X and Z, and the KL divergence is defined by $D_{KL}(p||q) = \mathbb{E}_p(\log \frac{p}{q})$.

The formula with conditional entropy shows in particular that, in the absence of noise and in the case of an invertible transformation (f injective, W invertible), the *infomax* criterion is simply to maximize the entropy of the output Z. In this sense, the *infomax* criterion is very close to the reduction of redundancy as seen by Shanon and Barlow (eq. 1.1) : in both cases, the goal is to maximize information transfer, but the view with mutual information in addition makes the distinction between relevant information (related to the input) and irrelevant information (noise: environmental noise and neural noise).

Resource constraints: For the infomax criterion to be restrictive on the output Y and the matrix W, a resource constraint must be applied, otherwise one solution would be to choose outputs with ever greater amplitudes. A channel typically can only code for a signal with a specific dynamic range. A common constraint is to limit the variance of the Z_i channels. A formulation of *infomax* with a penalty term on the variance is as follows [163] :

$$\max_{W,f} I(X,Z) + \rho \sum_{i=1}^{n} Var(Z_i) \ .$$

Another formulation is obtained by using the formula of mutual information with conditional entropy and introducing a weight that penalizes the term $H(Z)$ or amplifies the effect of the noise aversion term $-H(Z|X)$. Atick proposed for example [10] :

$$\max_{W,f} (1-\eta)H(Z) - H(Z|X)$$

with $\eta > 0$.

Output normalization/whitening. Considering a univariate channel Y (associated with one dimensional distribution p_y), information-maximization is equivalent to the process of output normalization, or *whitening*. The goal of whitening is to map the distribution p_y into a distribution of maximum entropy p_z. This process is called *whitening*, since under the hypothesis of identically distributed and independent samples, the output process Z is no different from white noise.
Example: we assume here that the constraint on the output process Z is to produce values between 0 (no activation) and 1 (activated output). This constraint is used in the infomax algorithm proposed by Bell & Sejnowski (1995) [21] in which Z is the output of a sigmoid function:

$$z = f(y) = \sigma(a\, y) = \frac{1}{1+\exp(-ay)} \ .$$

The distribution of maximum entropy associated with this constraint is the uniform distribution $p_z(z) = 1$. The mapping between p_z and p_y with f sets the relation $dz = p_z(z)dz = p_z\left(f(y)\right) f'(y)dy = p_y(y)dy$, that implies $p_y(y) = f'(y)$. In other words, to whiten the data, f should be as close as possible to the cumulative distribution of p_y.

1.1.3 Minimum entropy codes

Minimization of mutual information. A *minimum entropy code* [17] minimizes the mutual information between the m output components:

$$I(Y_1, \cdots, Y_m) = \sum_{i=1}^{m} H(Y_i) - H(Y) \ . \tag{1.3}$$

This term represents the part of redundant information that is due to the multivariate nature of the process (eq. 1.2, fig. 1.3). This term is rewritten with the Kullback-Leibler divergence and marginal probability distributions $p_i(y_i)$:

$$I(Y_1, \cdots, Y_m) = D_{KL}(p(y)||p_1(y_1)p_2(y_2)\cdots p_m(y_m)) \ .$$

This term is non-negative and equal 0 when the output components are independent. In case of independence, the input is represented with a set of independent features, and the log-likelihood of a given output is the sum of the log-likelihoods of the marginal probabilities (*factorial code*):

$$\log p(y) = \sum_i \log p_i(y_i) \ .$$

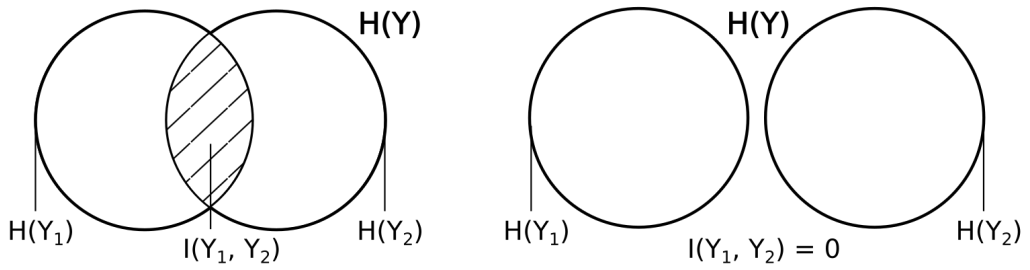


Figure 1.3 – The mutual information $I(Y_1, Y_2)$ between two output channels Y_1 and Y_2 is the common information that is encoded by the two channels. This redundant information corresponds to a waste of total capacity. If the mutual information is strictly positive, knowledge of one process gives information about the other ($H(Y_1|Y_2) < H(Y_1)$). In the case where mutual information is zero, the two processes are independent.

Measures of independence. The major difficulty of finding independent components is that independence is a strong statement, and cannot be expressed with a single empirical measure. Verifying the independence constraint $\log p(y) = \sum_i \log p_i(y_i)$ for different decompositions require to estimate many probability distributions, which is a demanding statistical task when there is no prior information about these distributions.

Practical methods use instead loosened definitions of independence, corresponding to different assumptions about the marginal distributions. One strategy is closely related to the definition of independent processes. Two processes Y_1 and Y_2 are independent i.f.f. for every function $f_1, f_2 \in L_1$,

$$\mathbb{E}(f_1(Y_1)f_2(Y_2)) = \mathbb{E}(f_1(Y_1))\mathbb{E}(f_2(Y_2)) \ .$$

A loosened definition of independence, then, is to chose specific functions f_1, f_2, transforming the constraint of independence into a much weaker decorrelation constraint [85]. Other strategies are based on measures of *negentropy* or non-gaussianity (see next paragraphs).

Structure vs diversity. A concurrent view of *minimum entropy coding* is that, in order to achieve a maximum compressible code, each output channel should code for a minimum quantity of information. This leads to the minimization of the sum of the marginal entropy terms, but at the same time, the quantity of the information of the entire process should not be squeezed. Hence, the second term in the cost function

$h = \sum_i H(Y_i) - H(Y)$

is a penalty term that ensures that the information of the input is kept throughout the transformation. In particular, it prevents the filters from collapsing in one direction. This view shows that a minimum entropy code must find a balance between *structure* (the components should be of minimal entropy) and richness of representation or *diversity* (the directions should represent all directions in space).

It has been argued that components are rarely independent or even uncorrelated for real data (e.g. for speech, all the components are activated at glottal excitations) [37], and that structure can become more important that 'uncorrelatedness' for ICA [79]. A rule of thumb for finding minimum entropy codes in practice is to "*find the most nongaussian projections, and use some (soft) decorrelation*" (quote from Hyvarinen, 1999 [78]).

Measures of *negentropy*. In the same way the empirical verification of independence requires assumptions about the components, prior information about the marginal distributions is

needed to estimate the entropy terms. Many practical methods rely on empirical measures of *negentropy* (opposite of entropy), that have to be maximized. More formally, the negentropy of a probability distribution is defined as the difference between the entropy of the distribution and the entropy $H_{\max}$ of a Gaussian distribution with same variance :

$$\text{Negentropie}(Y) = J(Y) = H(Y) - H_{\max}.$$

A notion close to negentropy is non-gaussianity. Many measures of negentropy are based on high order moments (>2). For example, Amari proposed a method based on the expansion of entropy with higher order moments for nearly gaussian distributions (hence, with minimal assumptions on the marginal distributions) [5].

Sparsity. The sparsity of coefficients required to describe the signal is a popular prior used to facilitate the estimation of *negentropy*. Popular measures for normalized distributions under the sparsity assumption are the fourth moment (kurtosis) [78] and the l_p norms:

$$||x||_p = \mathbb{E}(|x|^p)^{1/p}$$

with $1 \leq p < 2$ [164, 94]. The l_1 norm corresponds to the Laplace prior (see frame below). For speech data in particular, marginal distributions for various time-frequency decompositions have been shown to be well characterized by Laplace distributions [57]. This prior has been used multiple times for Independent Component Analysis applied to speech [91, 153].

Example: Generalized normal distributions

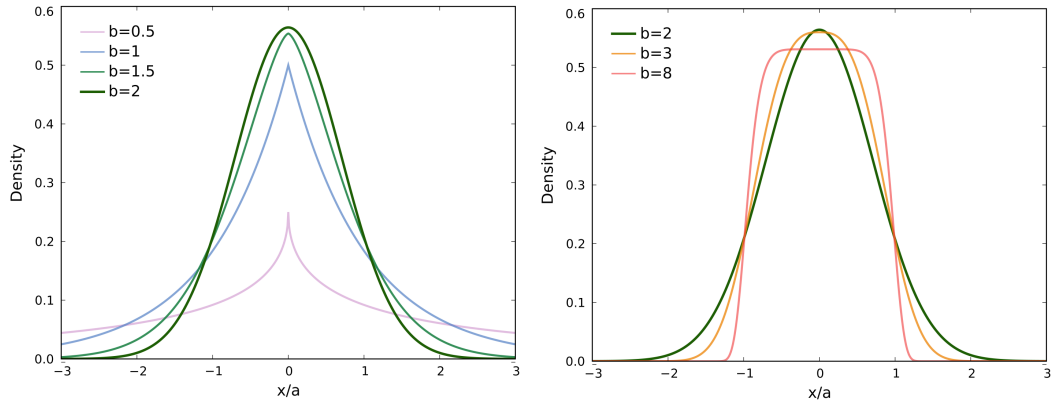


Figure 1.4 – The generalized normal distributions are a family of distributions parametrized with a (size) and b (shape), and defined by $p(x) \propto \exp(\frac{x}{a})^b$. **left:** distribution examples with $b \leq 2$ (*leptokurtic*), **right:** distribution examples with $b \geq 2$ (*platykurtic*)

The generalized normal distributions [116] are a family of probability distributions defined by:

$$p(x) = \frac{b}{2a\Gamma(1/b)} e^{-(|x-\mu|/a)^b}$$

where μ is the mean, $a > 0$ is a scaling factor and $b > 0$ is a parameter controlling the shape of the distribution. The normalization factor involves the gamma function Γ , defined by:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt \ ,$$

(special cases: where $x = n$, is an integer, $\Gamma(n) = (n-1)!$; $\Gamma(1/2) = \sqrt{\pi}$ is the Gaussian integral).
The variance of the extended normal distribution is:

$$\sigma^2 = a^2 \frac{\Gamma(3/b)}{\Gamma(1/b)} \ ,$$

hence the scaling parameter a can be replaced with $\sigma = \sqrt{\frac{\Gamma(3/b)}{\Gamma(1/b)}}a$ to obtain normalized distributions with respect to the second moment. The distributions are all symmetric, and therefore have zero odd-moments.

Notable cases: The generalized normal distributions includes both the *Gaussian distribution* ($b = 2$),

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(x-\mu)^2}{2\sigma^2} \ ,$$

and the *Laplace distribution* ($b = 1$) defined as:

$$p(x) = \frac{1}{2a} \exp^{\frac{|x-\mu|}{a}} = \frac{1}{\sqrt{2}\sigma} \exp^{\sqrt{2}\frac{|x-\mu|}{\sigma}} \ .$$

When b goes to $+\infty$, the density converges pointwise to a uniform density on $[\mu - a, \mu + a]$.

Entropy: The entropy associated with this family of distributions is:

$$H(X) = -\log\left(\frac{b}{2a\Gamma(1/b)}\right) + \mathbb{E}(\frac{|x-\mu|^b}{a}) = -\log\left(\frac{b}{2a\Gamma(1/b)}\right) + \frac{1}{b} \ ,$$

or, written with σ to obtain entropy terms related to fixed variance:

$$H(X) = \log 2\sigma - \log b - \frac{1}{2}\log\Gamma(3/b) + \frac{3}{2}\log\Gamma(1/b) + \frac{1}{b} \ .$$

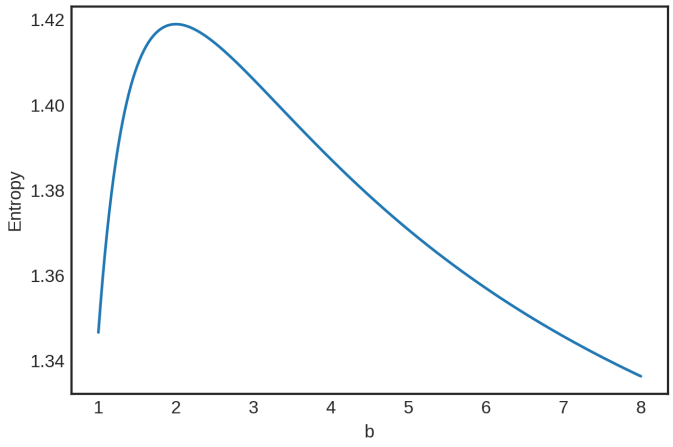


Figure 1.5 – Values of the entropy (in natural base) for fixed variance ($\sigma = 1$), as a function of the parameter b.

Naturally, the entropy has its maximum at $b = 2$, since the Gaussian normal distribution is of maximum entropy among the distribution of fixed variance, and is monotonic for $b < 2$ and $b > 2$.
Note that for any probability distribution q that we would want to approximate by an extended normal distribution, the cross-entropy is related to the l_p norm with $p = b$. In particular, in the case $b = 1$ (Laplace distribution), we have:

$$H(q,p) = \mathbb{E}_q(-\log p(x)) = \frac{1}{2}\log\left(2\sigma^2\right) + \sqrt{2}\ ||x||_1/\sigma\ . \tag{1.4}$$

Kurtosis: The kurtosis κ is the normalized fourth moment :

$$\kappa = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2}$$

The kurtosis quantifies the *tailedness* of a distribution, i.e. the ability for the associated process to produce large samples (relative to the standard deviation). For this specific family of distribution and other standard distributions, kurtosis also quantifies *peakedness*: distributions with high kurtosis have a narrow peak on their average value to compensate for the heavy tails.

Kurtosis separates the generalized normal distributions into two sub-families:

— *platykurtic* densities ($\beta > 2, \kappa < 3$): these densities have no tail and are concentrated on a interval.

— *mesokurtic* densities ($\beta = 2$, $\kappa = 3$): Gaussian normal distributions.

— *leptokurtic* densities ($\beta < 2, \kappa > 3$): these distributions have heavy tail and a peak on their mean value. When $\mu = 0$, most sampled values are close to zero with peaked distributions, hence these densities are used to model processes generating a *sparse* number of significant values.

The kurtosis has been used as a measure of non-gaussianity, sparsity [121], or *negentropy* [5].
Among the leptokurtic densities of the extended normal distributions, the log-concave distribution associated with minimum entropy and maximum kurtosis is the Laplace distribution $-\log p \propto |x| + cte$. This property makes the Laplace distribution a natural choice for modeling many stochastic sources satisfying the sparsity hypothesis.

Link between minimum entropy coding and *infomax*. Minimum entropy coding is related to the *infomax* criterion in the absence of noise [115, 21, 29]. Recall that, where there is no equivocation, $H(Z|X) = 0$, the *infomax* principle is simply to maximize the information of the output $H(Z)$, and that Z is related to Y by a non linear function defined component-wise: $Z_i = f_i(Y_i)$. This change of variable in the entropy yields:

$$H(Z) = -\int_z \ln p(z)\ p(z)\mathrm{d}z = -\int \ln\left(\frac{p(y)}{\prod f_i'(y_i)}\right)\ p(y)\mathrm{d}y\ .$$

We recognize here the KL divergence between two distributions:

$$H(Z) = -D_{KL}(p(y)||\prod f_i'(h_i)) \geq 0\ . \tag{1.5}$$

The case of equality (maximization of information) is obtained under two conditions, that correspond to two stages of processing:

— $p(y) = \prod_i p_i(y_i)$, the intermediary representation is factorial (independent features) – first stage of processing.

— $f'_i = p_i(y_i)$: the outputs are then normalized (*whitening*) – second stage of processing.

This analysis shows that the minimization of mutual information is largely redundant with the *infomax* principle, and corresponds to a first stage of processing (choice of the most appropriate representation). The infomax framework is richer, however, because it encompasses the normalization step, and noise can be included in the model.

1.2 – Algorithms and methods related

1.2.1 Independent Component Analysis

Overview of approaches for ICA. The search of independent features is translated in practice into a class of algorithms called Independent Component Analysis (ICA). ICA seeks a transformation – in most works linear – that makes the components of high dimensional data statistically independent. The term was introduced by C. Jutten and J. Herault in the end of the 1980s [85] in the context of *blind source separation* (BSS). Blind source separation is well known in speech & audio applications [167], as it is the mathematical translation of the "cocktail party problem": how to separate a mixture of speech signals from different speakers? ICA addresses this problem by leveraging the independence of the sources. For efficient coding, the view is reversed, with the neural coding units acting as if they were sources for the incoming signal. The initial algorithm of Jutten and Herault was motivated by the observation that decorrelation was a constraint too weak to recover independent sources. They derived a Hebbian-like rule in order to cancel nonlinear cross-correlation terms (using polynomial transfer functions). The other proposals that followed were also based on higher order cumulants (see Comon, 1994 [34] for a more detailed account of the early history of ICA).

There are many modeling approaches for ICA [78], all resulting in similar cost functions. The previous paragraphs have introduced the informatic-theoretic approaches of ICA: the minimization of mutual information [34] or minimization of entropy, and the *infomax* principle [99]. Other two common approaches of ICA are the maximum likelihood principle (closely related to information-theoretic approaches [29], see also next paragraphs), and the maximization of non-gaussianity [34]. The latter is a a kind of reverse version of the central limit theorem: gaussianity increases when random variables are mixed, hence gaussianity should decrease when these variables are separated.

Cost function. The goal of ICA is to find a matrix W such that the components $Y_1, \cdots, Y_m$ of $Y = W^T X$ are independent. We consider here that the matrix W is a square matrix ($m = n$). We have that $H(Y) = H(W^T X) = H(X) + \log|\det W|$. The first term is the quantity of information of the input and does not depend on W.

Therefore, we can derive a cost function for ICA from the constraint of minimization of mutual information (eq. 1.3) with :

$$h(W) = \sum_i H(Y_i) - \log|\det W| = \sum_{i=1}^{n} H(W_i^T X) - \log|\det W| \ .$$

ICA seeks to minimize the sum of entropy terms, with the constraint that W should not collapse ($\log\det$ term). However, as mentioned in previous paragraphs, the marginal entropy terms remain to be estimated, or a prior q on the marginal distributions must be introduced. In the latter case, we can replace the entropy terms with cross-entropy terms:

$$H(p_i) = H(p_i, q) + D_{KL}(p_i||q) \approx H(p_i, q) = -\mathbb{E}(\log q(y_i))$$

where p_i is the marginal distribution of the component Y_i (we ignore the term with the KL-divergence being the error of estimation). $h(W)$ becomes:

$$h(W) = -\sum_{i=1}^{n} \mathbb{E}\left(\log q(W_i^T X)\right) - \log|\det W| \ . \tag{1.6}$$

In particular, if the Laplace prior is used (eq. 1.4), the cost function is:

$$h(W) = \sum_{i=1}^{n} \left[\frac{1}{2}\log\left(2\sigma_i^2\right) + \sqrt{2}\ ||W_i^T x||_1/\sigma_i\right] - \log|\det W|$$

where σ_i is an estimated standard deviation of the i-th component.

Relation to a maximum likelihood formulation of blind source separation. The derivation above with cross-entropy terms seems very similar to a maximum likelihood formulation (for a parametric estimation of a single distribution, the log-likelihood is the opposite of cross-entropy). The "dual" view to the derivation above stems from the problem of blind source separation (BSS). We want to find $A = W^{-T}$, the *mixing* matrix, such that X is a *mixture* of *independent* sources $Y_1, \cdots, Y_n$ with $X = W^{-T}Y$. The posterior probability distribution of X is:

$$p(x) = \frac{\prod_{i=1}^{n} p_i(W_i^T x)}{|\det W^{-T}|} = |\det W| \prod_{i=1}^{n} p_i(W_i^T x) \ .$$

Hence, the log-likelihood of the model for a family of distributions q approaching p_i is [94]:

$$L(W, q) = \sum_{i=1}^{n} \mathbb{E}\left(\log q(W_i^T X)\right) + \log|\det W| \ ,$$

which is exactly the opposite of eq. 1.6.

Algorithms. Most implementations of ICA use a stochastic gradient descent to minimize the cost function (eq. 1.6). The gradient of h with respect to W is:

$$\nabla_W h : \mathbb{E}(X\psi(Y)^T) - W^{-T}$$

where $\psi = -q'/q$, called the score function, is applied component-wise on Y. The gradient descent can be accelerated by using second order schemes with an approximated Hessian matrix [4]. Many articles use the so-called *natural gradient method* [5], which is to multiply the gradient by WW^T. This method has also the advantage that the inverse matrix does not need to be computed. The iterations follow the scheme:

$$\Delta W = \mu W \left(I - \mathbb{E}(Y\psi(Y)^T) \right) \tag{1.7}$$

where $\mu > 0$ is the step size. The expectation is typically estimated from a batch of examples. If the Laplace prior is used, then we have the rule:

$$\Delta W = \mu W \left(I - \frac{\sqrt{2}}{\sigma} \mathbb{E}(Y \mathrm{sgn}(Y)^T)) \right)$$

where sgn is the sign function (-1 for negative numbers, +1 for positive numbers).

A popular algorithm, *FastICA*, follows a fixed-point iteration scheme instead of a gradient descent [77, 78]. It has been shown that FastICA is analogous to a second order gradient descent scheme [4].

1.2.2 Sparse coding methods

Many methods are related to the search of sparse representations. Independent Component Analysis itself, as seen in the previous paragraph, is often used with a sparsity prior on the source activations [91], and the term of *Sparse Component Analysis* has been coined [67].

Sparse dictionary methods. Sparse coding methods often comes with the notion of *dictionaries*. A dictionary $\mathbf{D}$ is a set of vectors $(d_1, ..., d_m) \in \mathbb{R}^{n \times m}$ whose elements are called the *atoms*. The goal of sparse coding is to represent the input signals in these dictionaries with a small number of atoms. Contrary to Independent Component Analysis, which works preferentially with complete dictionaries ($m = n$), sparse methods are most often used with overcomplete dictionaries ($m > n$). This is explained by the fact that the sparsity constraint enables undetermined linear systems to be well-posed – this fact is at the basis of the field of compressed sensing [42]. Atoms can be either fixed (e.g. Gabor dictionaries, wavelets [105], etc.) or learned from data. When dictionaries are fixed, atoms are chosen so that the dictionary has a 'structure' (e.g. the atoms of a Gabor dictionary are versions of the same basis function shifted in the time-frequency plane).

***Analysis* vs *reconstruction*.** In this thesis, I make a distinction between two paradigms (see also ref. 14 where a similar point of view is adopted):

1. The *analysis* paradigm – the decompositions are computed with the dot products between the signals and the atoms; meaning that the vector of interest is

$$Y = W^T X = \mathbf{D}^T X,$$

   as presented until now.

2. The *reconstruction* paradigm – only the most significant atoms are selected, so that the signal is the sum of the selected atoms. The vector of interest is a sparse vector $Y = g(X)$ such that the error of reconstruction is minimized:

$$\epsilon = ||X - DY||_2 \ .$$

The second paradigm is mathematically attractive, because it leads to well-posed problems, even in a overcomplete setting. For this reason, it is the prevailing paradigm in sparse coding methods. One limitation is there is an implicit hypothesis that every

input signal has to be a sparse sum of atomic components, but this condition is not always verified (for example, many speech sounds are noise sounds: fricatives, stops...). The *reconstruction* paradigm comes with techniques for reconstructing the signal from the atoms (see below). It exchanges an additional computational cost with a more sparse representation of input data. This thesis primarily focuses on the *analysis* paradigm: in other words, there is no attempt to reconstruct the signal from neural activations. To know which paradigm is the most relevant to sensory systems is still a mater of debate, but it can be argued that the *analysis* step is a necessary first step in any case, and that a complete reconstruction of the signal is generally a compute intensive task.

l_p-norm minimization. The most natural way to evaluate the sparsity of a vector is the l_0 "norm", defined by

$$||y||_0 = \lim_{p \to 0} ||y||_p^p = \lim_{p \to 0} \sum_{i=1}^{m} |y_i|^p$$

with the convention $0^0 = 0$ and $a^0 = 1$ for every other number. The l_0 "norm" is simply the number of non-zero coefficients of a vector. However, the l_0 "norm" is a non-convex, non-continuous function. The simplest sparse reconstruction problem (penalized form, $\lambda > 0$):

$$\min_y ||x - Dy||_2 + \lambda ||y||_0$$

is a NP hard problem [43]. One strategy is to *convexify* the l_0-"norm", that leads to replacing the l_0-"norm" by the l_1-norm [172]. The reconstruction problem

$$\min_y ||x - Dy||_2 + \lambda ||y||_1 \tag{1.8}$$

is called the *lasso* problem [158], which can be solved by several efficient algorithms [71].

Sparse dictionary learning. Sparse dictionary learning, in the *reconstruction* paradigm, can be formulated as an optimization problem [172]:

$$\arg \min_{\mathbf{D} \in \mathbf{S}, Y = g(X)} \mathbb{E}\left(||X - DY||_2^2 + \lambda P(Y) \right) , \tag{1.9}$$

where $\mathbf{S}$ is the set of admissible dictionaries and $P(Y)$ is a penalty term that ensures that the vector is sparse. P can be the l_p norm ($p = 0$ or 1). Algorithms of sparse dictionary learning typically update the dictionary D and the vectors y alternately. In the case $p = 0$, greedy algorithms are used to make an estimate of Y from X (matching pursuit [106]), then the dictionary D is updated following a given rule (e.g. gradient descent [149]). The same strategy can be used with $p = 1$, replacing the matching pursuit with a *lasso* algorithm [103].

Note that although I have presented ICA and sparse coding methods as the main methods for the investigation of statistical structure, they are not exclusive. For example, restricted Boltzmann machines (RBM) can be used as an alternative for ICA: when applied to speech, they yield the same kind of filters [82].

1.2.3 Dealing with overcompleteness

Independent Component Analysis works preferentially in a determined case (n the dimension is equal to m the number of sources). In this case, the entropy of the output vector $H(Y)$ in the definition of mutual information (eq. 1.3)

$$h = \sum_{i=1}^{m} H(Y_i) - H(Y) \qquad \text{(infomin, recall)}$$

can be replaced with $\log|\det W|$ to derive a cost function for ICA (eq. 1.6). However, complete bases have two disadvantages: a) they are not easy to manipulate. For instance, it is not straightforward to build orthogonal wavelet bases [105]. b) they do not correspond to actual coding. Models based on overcomplete bases are more realistic (this is discussed in next subsection, par. 'the neural code is redundant').

The $\log|\det W|$ penalty ensures that the directions defining the independent components are uncorrelated in addition to capture the statistical structure. The main difficulty in reproducing the derivation of the cost function when dealing with overcomplete bases is that there is no natural expression of "correlatedness" or redundancy for overcomplete bases that could substitute the $\log|\det|$ term [79]. For time-frequency frames, a common empirical measure of redundancy is related to the density of the frame (i.e. the number of filters for a finite representation), but even in this particular case, mathematical grounds for this quantity are scarce [72]. One solution, introduced in previous paragraphs, is to enforce the sparsity of the vectors y by trying to reconstruct the inputs x from y [95, 67]. The *reconstruction* paradigm alleviates the issue of correlations by selecting the atoms, leading to tractable optimization problems – it shows that the problem for overcomplete dictionaries is not so much how atoms are correlated but rather how they can be *selected* in further steps. The other solution, adopted in the rest of the thesis, is not to constrain the vector Y, but instead to constrain the representation, meaning that the matrix W (or $\mathbf{D}$) must be an element of a restricted set $\mathbf{S} = \{\mathbf{D}_\theta\}_\theta$, parameterized by θ. Structure (sparsity) and diversity (eq. 1.3) are then ensured differently:

— *Structure* (entropy minimization): minimization of the sum of marginal entropies

$$h(\theta) = \sum_i H([D_\theta]_i^T X) \qquad (1.10)$$

— *Diversity:* the dictionaries have similar 'structure' and atoms are distributed uniformly in space. For example, $\mathbf{S} = \{\mathbf{D}_\theta\}_\theta$ can be a set of Gabor dictionaries built from function bases $\mathbf{F} = \{f_\theta\}_\theta$ whose atoms are evenly distributed in time-frequency-phase.

1.3 – Evidence and limits

Evidence for the efficient coding hypothesis. Evidence for the efficient coding hypothesis is of two kind: a) empirical measures of information rate in the brain when naturalistic stimuli are presented, b) prediction of properties specific to a sensory system.
a) *Empirical measures*: One example of evidence is that coding efficiency is maximized at a neural level when natural sounds or sounds with naturalistic characteristics, are presented. In 1995, Rieke showed that information transfer is increased in the auditory nerve of

bullfrogs with sounds with the spectrum of animal calls compared to white noise [135]. In 1997, Attias and Schreiner demonstrated the same result with neurons in cats' inferior colliculus (midbrain) and for noises with naturalistic amplitude modulations [11]. It has been reproduced in the midbrain and in the auditory cortex of zebra finches with natural sounds and synthetic sounds with naturalistic time-frequency modulation statistics [75].
b) *Prediction power:* The efficient coding hypothesis can make predictions – or explain – specific characteristics of sensory systems by revealing the statistical structure of sensory data. This thesis shares this approach and some results have already been presented. Independent Component Analysis or methods of sparse coding on natural images result in Gabor-like oriented wavelets that resemble the receptive profiles of visual cells in V1 [120, 164]. Many other attributes of the visual system have been shown to be consistent with the efficient coding hypothesis: contrast sensitivity, colour coding, sensitivity to movement, etc. [145, 154]. For the auditory system, consistencies with speech statistics have been shown for the shape of the auditory filters (this is detailed in the present thesis), or for modulation filters in the midbrain [137, 30, 113].

Is the brain really analogous to a communication system? The efficient coding theory uses vocabulary elements from information theory (code, mutual information, channel capacity, redundancy...) and applies them to neural code that emerges from the activity of a large number of neurons. It is not really clear how these concepts are relevant to the description of neural activity [27]. Even considering a single or small number of neurons, how to measure the amount of information encoded by these neurons (see ref. 2 chap. 1 for a detailed analysis of this issue)? The simplest solution would be to count the number of action potentials (*spikes*) delivered (or equivalent, estimate the firing rate). This solution is similar to how the amount of information is computed in the case of sparse activations, at least for a simple model (see next chapter). However, it does not take into account temporal coding, that includes the time of firings, temporal (auto)correlations, etc. Considering neural coding from the sole angle of information theory, which quantifies information by probability distributions of a fixed quantity reflecting neural activity (e.g., firing rate), neglects the dynamic aspects and implications of neuronal connectivity. One effect on modeling is that the physical functioning of neurons necessarily introduces noise into the models of sensory processing. An example is given by inference of the contrast sensitivity of the visual system according to spatial frequency for low intensity stimuli [163, 10]. In this example, we know that the emission of action potentials, which are by nature discrete, introduces multiplicative noise, and this knowledge plays an important role in the analysis. To apply information theory results, it is necessary to have a stochastic model of the representation of the input on which the coding system is based. However, building a representation of the environment is not exactly the purpose of sensory processes which play a broader role in *perception-action* loops. All these reasons suggest that the vocabulary of communication systems should be applied to cognitive processes with sufficient caution, bearing in mind that it is above all a convenient way of abstracting the system.

Information bottleneck. The combination of information theoretic considerations of coding efficiency with learning models is necessary to model cognitive processes because the brain does not encode the total information that is perceived, but removes information during the sensory processing steps that is not relevant for higher level tasks (concerning sound and speech, these tasks can be speech recognition, speaker identification, acoustic scene recognition...). These high-level tasks require the transition from a continuous code to

discrete, categorical representations (e.g. phoneme recognition) [141]. These tasks use class invariants to separate categories from each other. Conversely, they get rid of information that corresponds to variations within the same class. This increasingly fine filtering of information is called *information bottleneck*. A formulation of the information bottleneck method in a similar fashion to the *infomax* criterion is to maximize the quantity

$$\max_{p(z|x)} I(H,Z) - \frac{1}{B} I(X,Z)$$

where H is the relevant information from the input (e.g. information of the category to which X belongs) [159], and B is a positive constant. Category coding has implications for low-level encoding. For example, if the categories are coded by a population of neurons, the receptive fields will be concentrated at the boundaries of these categories in the representation space [23]. However, if the signal is exploited by various high-level cognitive functions, each based on different parts of the signal, then the first step in sensory processing is to achieve a sufficiently broad representation of the input. Therefore, the hypothesis of efficient coding is especially relevant in the study of low-level peripheral sensory processing. For high-level processing, neural representations become more abstract and complex and the efficient coding paradigm is less appropriate when used on its own.

The neural code is redundant. A criticism of the efficient coding theory and the 'redundancy reduction' principle is that the neural code is highly redundant. An example often mentioned is the large number of neurons in the primary visual cortex V1 ($\sim 10^9$ neurons), far more than the number of retinal ganglion cells ($\sim 10^6$ cells) [18]. Redundancy is important for a couple of reasons: a) it makes the neural representation more robust to noise [99]; b) overlap between features is needed to capture details of the signal (e.g. edges of an image). In time-frequency analysis and the theory of Gabor frames, redundancy is known to be a unavoidable property for good time-frequency features [72, 134]. It has been argued that the high number of neurons in V1 would be the result of a compromise to obtain sparse representations of abstract features in higher level areas [121]. In 2001, H. Barlow wrote an article entitled *Redundancy reduction revisited* [18], in which he enumerated reasons why it is still relevant to seek minimally redundant representations despite the fact that the neural code is redundant. Overall, it can be argued that it is a general way to find good features of data in a non-supervised fashion.

Chapter 2

Sparse time-frequency representations

In the previous chapter, coding efficiency was related to criteria that apply to the statistics of decompositions (e.g. sparsity of response patterns). The nature of these decompositions, however, has not been specified. For sensory stimuli such as natural images or speech, the optimal decompositions perform a time-frequency analysis of the input signals. This chapter addresses the problem of sparse coding, from the perspective of time-frequency analysis, complementary to the information-theoretic approach. The input/output system is linked to quadratic time-frequency representations. The sparsity of response patterns is shown to be restricted by the uncertainty principle: time-frequency features cannot offer unlimited selectivity in time and frequency at the same time. The uncertainty principle tells that the key issue in finding an appropriate representation for a class of signals is to know the optimal time-frequency resolution trade-off, which can be constant or frequency-dependent. The theory of time-frequency analysis also underlines the specificity of Gabor filters, i.e. Gaussian-modulated sinusoids, that are associated with the best possible time-frequency resolution.

2.1 – Model

Definitions and notations.

— *Fourier transform:* In this chapter, the functions f or g are complex signals that belong to $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ (finite l_1 and l_2 norms). I will consider the Fourier transform defined with angular frequency ω:

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_t f(t) e^{-i\omega t} \mathrm{d}t \ .$$

The (extended) Fourier transform maps $L_2(\mathbb{R})$ onto itself and is a unitary operator (Plancherel theorem) .

— *Fundamental operations:* T_x and M_ω denotes the translation and modulation operators:

$$T_x f(t) = f(t-x) \ , M_\omega f(t) = f(t) e^{i\omega t} \ .$$

$\tilde{f}$ is the result of the reflection operator on f:
$\tilde{f}(t) = f(-t)$
The complex conjugate of v is denoted $\bar{v}$. The dot product between and f and g is $< f, g >= \int f \bar{g}$

— *Short-term Fourier transform (STFT):* The short-term Fourier transform or windowed Fourier transform is defined by:

$$\mathcal{V}_g(f)(x,\omega) = \frac{1}{\sqrt{2\pi}} < f, M_\omega T_x g >= \widehat{f\ T_x\bar{g}}\,(\omega) = \frac{1}{\sqrt{2\pi}} \int_t f(t)\overline{g(t-x)}e^{-i\omega t}dt\ .$$

$\mathcal{V}_g(f)$ maps $L_2(\mathbb{R})$ into $L_2(\mathbb{R}^2)$. It has a property of 'energy convervation' too. Let (f_1, g_1), (f_2, g_2) be two couples of functions of $L_2(\mathbb{R})$, then:

$$< \mathcal{V}_{g_1}(f_1), \mathcal{V}_{g_2}(f_2) >_{\mathbb{R}^2} = < f_1, f_2 > \overline{< g_1, g_2 >}\ .$$

g is called the *window function* or the *analysis function*. It is typically a log-concave function, which has most of its energy on a compact support centered around 0 (e.g. Gaussian function). If g is seen as an argument, the short-term Fourier transform is a quadratic function. Other formulations of quadratic time-frequency transformations are introduced further in text.

Model. In this chapter, I will only consider a simplified and abstract model of the input-output system (chap. 1, fig. 1.1) to provide explicit results. The stochastic source X emits signals which are shifted versions $M_\omega T_x f$ of a single basis function f in the time-frequency plane $(x, \omega) \in \mathbb{R}^2$. A consequence of this choice is that the input has complex values. I consider this case for simplicity, but it can be shown that considering instead real shifted signals $\sin(\omega \cdot + \varphi) T_x f$, $\varphi \in [0, 2\pi]$, which is more realistic, is equivalent under the assumption that the frequency spread of f is much lower than center frequencies ($\sigma_\omega \ll \omega$). The outputs $Y = W(X)$ are the dot-products between X and a infinite set of atoms, which are also shifted versions $M_\omega T_x f$ of an analysis function g. Another consequence of these (strong) assumptions is that the frequency selectivity of the input and analysis atoms remains the same along the frequency axis. This is unrealistic but makes it possible to derive a few mathematical properties and to gain intuition for the time-frequency coding problem. The last paragraph of this chapter introduces representations for which this constraint is released.
We will be interested in the cost function $h(X)$ defined by the mean of the activations $Z = |Y|$ (see also next paragraph). If we assume the time and frequency shifts of input and analysis atoms follow a centered Gaussian distribution of variance Σ_T and Σ_Ω (so that the total shift in the time-frequency plane follows a Gaussian distribution of variance $\mathrm{diag}(2\Sigma_T, 2\Sigma_\Omega)$), the expectation of $h(X)$ is:

$$\mathbb{E}\left(h(X)\right) = \frac{1}{4\pi\sqrt{\Sigma_T\Sigma_\Omega}} \iint |< f, M_\omega T_x g >| \exp\left(-\frac{x^2}{4\Sigma_T} - \frac{\omega^2}{4\Sigma_\Omega}\right) \mathrm{d}x\, \mathrm{d}\omega$$

and when the variances go to infinity, we have the equivalence:

$$\mathbb{E}(h(X)) \underset{\Sigma_T, \Sigma_\Omega \to \infty}{\sim} \frac{1}{2\sqrt{2\pi\Sigma_T\Sigma_\Omega}} ||\mathcal{V}_g(f)||_1\ . \tag{2.1}$$

Hence, the cost function is related to l_1-norm of the STFT.

Remarks on the activation function. In the preceding chapter, we have seen that the l_1 norm can be derived either from cross-entropy terms (or log-likelihood) for a Laplace distribution prior, either from the convexification of the l_0-"norm". Rather, the model that

has just been introduced stems from the concurrent view that the objective of an efficient representation is to reduce the activation of coding units. This appears as a natural way to take into account the need for biological systems to maintain low energy expenditures, since a lower activity of neurons translate into savings in metabolic resources [93]. A direct measure of the lack of sparsity is to *count* the number of neuron spikes. In the modeling of biological neurons, the activation function of a single neuron (or a population of neurons) is the expected firing rate as a function of the signal amplitude at its input. The choice of the absolute value (modulus) in eq. 2.1 corresponds to the symmetric version of a rectified linear unit (ReLU, fig. 2.1 a.), a common choice for the activation function of artificial neural networks [61]. ReLU approximates real activation functions with a zero value below a threshold, then a linear regime. However, the activation function of a single neuron is better described by a sigmoid that can also model the saturation regime. The ReLU, then, corresponds to the sum of activations of neurons with sigmoid input/output curves, with a uniform prior on the threshold value (above a fixed threshold) [117]. This view is better suited to auditory coding: auditory nerve fibers have sigmoid rate-vs-level curves at different thresholds. However, this is only an abstraction of the actual process, which differs from the ideal model (fig. 2.1 b) in several ways. For example, auditory nerve fibers are categorized according to their spontaneous rates (SR). Fibers with high SR have a low threshold, conversely low SR fibers have high thresholds (fig. 2.1 c.). These fibers do not have the same firing rates, moreover high SR fibers are more numerous. Another difference is that adaptation phenomena occur in the peripheral auditory system, resulting in changes in thresholds over time [175].
The choice of the absolute value as an activation function acts as if neurons coding for filters with opposite phase were counted at the same time with a ReLU activation. The modulus is the activation function of the scattering network [28], a multilayered neural network based on the wavelet transform that reproduces the first layers of convolutional neural networks.

2.2 – Quadratic time-frequency representations

The short-term Fourier transform (STFT) has already been defined. For $f, g \in L^2(\mathbb{R})$:

$$\mathcal{V}_g(f)(x,\omega) = \frac{1}{\sqrt{2\pi}} \int_t f(t)\overline{g(t-x)}e^{-i\omega t}\mathrm{d}t \ .$$

The STFT is also called the Gabor transform, especially when g is a Gaussian. If g is taken as argument, $\mathcal{V}_g(f)$ is a quadratic form. The input function and the analysis function play a symmetrical role in quadratic forms. It parallels the problem of speech coding insofar as speech is a stimulus produced by humans. Even if the context, here, is very simplified, it makes sense to consider the input as part of model parametrization, and not only as a given function, as evolution has brought some control over the speech signal – although it is subject to physical constraints.

The STFT has one potential drawback: the functions $t \mapsto f(t)\overline{g(t-x)}$ are typically off-center, and that adds a phase shift to the result. Two other versions of quadratic time-frequency distributions are also common:

— *(Cross-)ambiguity function:*

$$\mathcal{A}(f,g)(x,\omega) = \frac{1}{2\pi} \int_t f(t+\frac{x}{2})\overline{g(t-\frac{x}{2})}e^{-i\omega t}\mathrm{d}t = \frac{1}{\sqrt{2\pi}} e^{i\omega x/2}\, \mathcal{V}_g(f)(x,\omega)$$

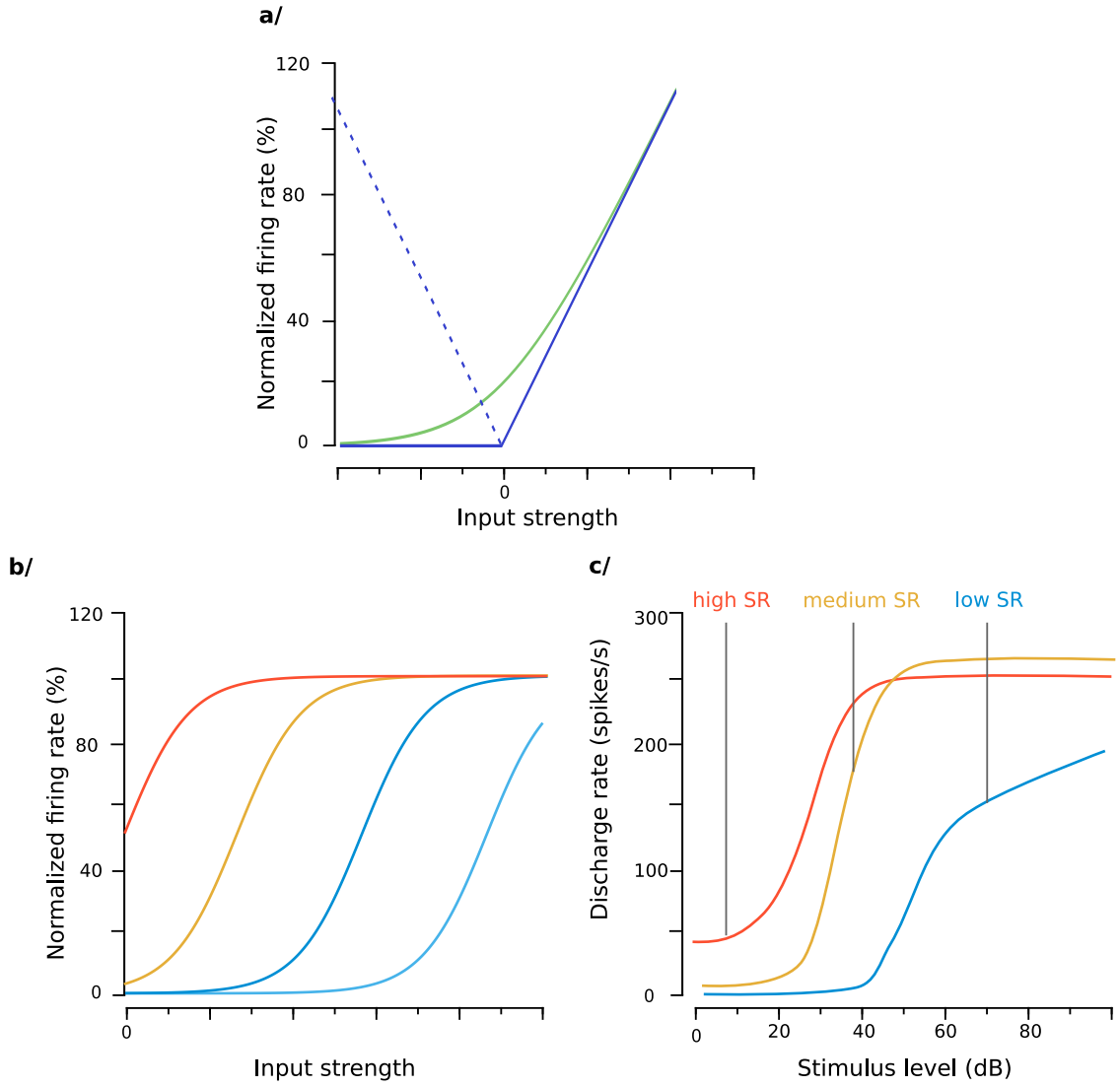


Figure 2.1 – **Activation functions:** **a/** *blue:* Rectified linear unit (ReLU), defined by $x \mapsto \max(x, 0)$ (*dash*: symmetric version, the absolute valute $x \mapsto |x|$). *pale green:* regularized version, the "softplus", defined by $x \mapsto \ln(1 + e^x)$. **b/** Underlying model for the softplus: it is the compound activity of sigmoids with thresholds uniformly distributed above 0. **c/** Schematic rate-vs-level curves of auditory nerve fibers: with high spontaneous rate (SR), medium SR, low SR *(adapted from Bharadwaj et al., 2014 [22])*

— *(Cross-)Wigner distribution:*

$$\mathcal{W}(f,g)(x,\omega) = \frac{1}{2\pi}\int_t f(x+\frac{t}{2})\overline{g(x-\frac{t}{2})}e^{-i\omega t}\mathrm{d}t = \sqrt{\frac{2}{\pi}}e^{2i\omega x}V_{\tilde{g}}(f)(2x,2\omega)$$

Cross-ambiguity functions and cross-Wigner distributions are related to each other with $\mathcal{W}(f,g)(x,\omega) = 2\mathcal{A}(f,\tilde{g})(2x,2\omega)$. The choice of the representation (Wigner-Ville distribution, ambiguity function, or STFT) does not have a big impact in our case since we consider the modulus of the output in a second step (eq. 2.1).

$\mathcal{W}(f) = \mathcal{W}(f,f)$ is the Wigner-Ville distribution of f. It is the most common choice for quadratric time-frequency representation since it has these additional properties:

— *Support property:* if f (resp. $\hat{f}$) is 0 outside a interval $[a,b]$, $\mathcal{W}(f)(\cdot,\omega)$ (resp. $\mathcal{W}(f)(x,\cdot)$) will be 0 outside this same interval.

— *Quasi-probability distribution:* $\mathcal{W}(f)$ is real and one has the following relations:

$$\text{a.} \int \mathcal{W}(f)(x,\cdot) = |f(x)|^2$$

$$\text{b.} \int \mathcal{W}(f)(\cdot,\omega) = |\hat{f}(\omega)|^2$$

$$\text{c.} \int_{\mathbb{R}^2} \mathcal{W}(f) = ||f||_2^2$$

In particular, if the l_2 norm of f is set to 1, these properties make the Wigner-Ville distribution similar to a posterior probability distribution.

Proofs (quasi-probability distribution). I denote $\varphi_x(f,g)$ the function:

$$\varphi_x(f,g) : t \mapsto f(x+\frac{t}{2})\overline{g(x-\frac{t}{2})} \tag{2.2}$$

We have $\mathcal{W}(f)(x,\omega) = \frac{1}{\sqrt{2\pi}}\widehat{\varphi_x}(\omega)$. $\varphi_x(f) = \varphi_x(f,f)$ is a Hermitian function, then its Fourier transform is real. In addition, $\int \mathcal{W}(f)(x,\cdot) = \frac{1}{\sqrt{2\pi}}\int \widehat{\varphi_x}(\omega)\mathrm{d}\omega = \varphi_x(0) = |f(x)|^2$, as a result of the inverse Fourier formula (equation a.). We get the symmetrical result on ω (equation b.) with the relation $\mathcal{V}_g(f)(x,\omega) = \mathcal{V}_{\hat{g}}(\hat{f})(\omega,-x)$ (Plancherel formula). Equation c. is straightforward from a. or b.

□

In reality, Wigner-Ville of normalized functions are not probability distributions for time-frequency concentration in general. An additional necessary and sufficient condition is that $\mathcal{W}(f)$ must be non-negative. The theorem below tells that it happens only in a very specific case.

Definition (Gabor Chirp)**.** *A Gabor 'chirp' is a function $f \in L^2(\mathbb{R})$ defined for $t \in \mathbb{R}$ by:*

$$f(t) = ||f||_2 \left(\frac{1}{2\pi\sigma_t^2}\right)^{1/4} \exp\left(-\frac{(t-t_0)^2}{4\sigma_t^2}\right) \exp\left(i(\omega_c t/T)t\right) e^{i\omega_0 t+\varphi} \tag{2.3}$$

σ_t is a positive number and $\omega_c/T, t_0, \omega_0, \phi$ are real parameters.

Here, ω_c/T controls the frequency chirp (if it is zero, the equation above defines a Gabor filter).

Theorem (Hudson's theorem [76])**.** *Assume $f \in L^2(\mathbb{R})$. $\mathcal{W}(f) > 0$ for all $(x,\omega) \in \mathbb{R}^2$ if and only if f is a Gabor 'chirp'.*

The demonstration of the Hudson's theorem is beyond the scope of this thesis and uses results from complex analysis (Bargmann transform). The property of positivity is closely related to the sparsity of Wigner-ville distributions, as shown in next section.

2.3 – Uncertainty principle

2.3.1 Heisenberg limit for time-frequency resolution

A natural question in the context of the efficient coding theory (chap. 1) is whether a particular input function and a window function can produce an efficient code, when the features coded are only time-frequency features. An ideal scheme would lead to a maximum sparse code (fig. 2.2 a.): for a signal localized at a time t_0 and a (angular) frequency ω_0, only one single unit would be activated in response to this input. This kind of unit that only activates when a very specific stimulus is presented has been called a 'grandmother' cell [121] (referring to the view that one individual would have a dedicated neuron for each people in his/her circle of acquaintances). This would also achieve a factorial code since each unit is activated for specific stimuli independently of its neighbors. However, the theory of time-frequency analysis tells that situation is highly unrealistic. If the dot products are computed at each point in the time-frequency plane (as for quadratic time-frequency distributions), the resulting representation necessarily includes an uncertainty about the localization of the response. More concretely, this means that a time-frequency distribution, even with localized input/window functions, typically presents a blur around the maximum response (fig. 2.2 b.), and that a single activation is not possible generally speaking.

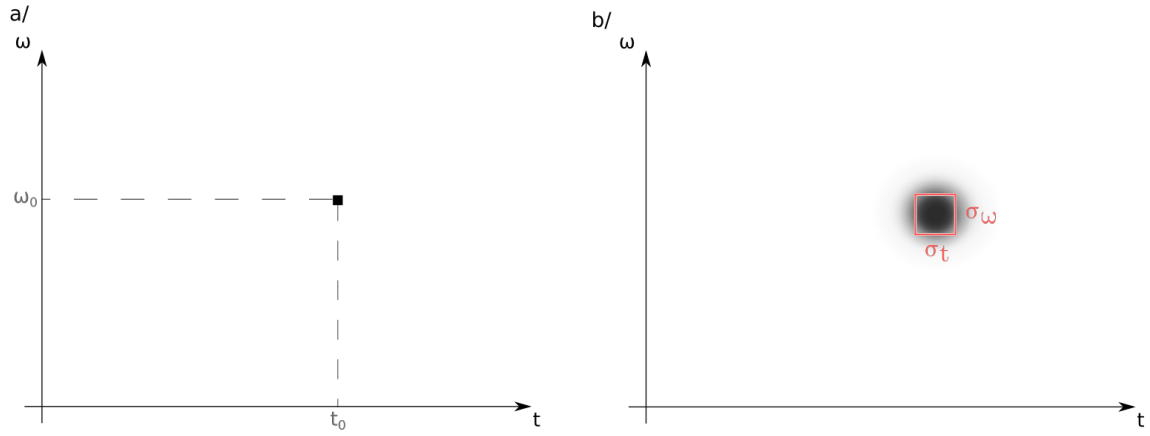


Figure 2.2 – **a/** Ideal time-frequency representation with a single unit activated. **b/** Actual time-frequency representation with a spread in the time-frequency plane (concentration of energy in a $\sigma_t \times \sigma_\omega$ box).

The spread in time-frequency plane can first be approached by the Heisenberg uncertainty principle. This principle is famous for its application in quantum mechanics: it states that the position and momentum (which is related to the Fourier transform of the wave function) cannot be known with a good accuracy with a single measurement.
Without loss of generality, I consider that f and $\hat{f}$ are centered around 0 and normalized such as $||f||^2 = ||\hat{f}||^2 = 1$. If f is not centered, f can be replaced with $T_{-\mu}f$ where μ is

defined by:

$$\mu = \int_t |f(t)|^2 \, t \, \mathrm{d}t \, ,$$

and the same can be done for its Fourier transform. The squared moduli define posterior probability distributions for time of frequency concentration. The standard deviations associated with t and ω are the time and (angular) frequency spreads σ_t and σ_ω. More explicitly, they are the positive values such as:

$$\sigma_t^2 = \int_t |f(t)|^2 t^2 \mathrm{d}t \, ,$$

$$\sigma_\omega^2 = \int_\omega |\hat{f}(\omega)|^2 \omega^2 \mathrm{d}\omega \, .$$

Heisenberg's uncertainty principle says that the time and frequency spreads are limited by the following inequality:

$$\sigma_t \sigma_\omega \geq \frac{1}{2}. \qquad \text{(Heisenberg-Gabor inequality)}$$

The equality is achieved if and only if f is a Gaussian function (possibly with a complex scaling parameter). Releasing the centering constraint, we have more generally that f is a 'Gabor chirp' (eq. 2.3).

Proof (in a simplified setting): Let's assume that σ_t and σ_ω are finite (this implies that f and its Fourier transform have derivative in $L^2(\mathbb{R})$. By Plancherel formula: $\sigma_t^2 = ||f \cdot t||^2 = ||\hat{f'}||^2$. By an integration by part, we have:

$$2Re\left(\int \hat{f}'(\omega)(\hat{f}(\omega)\ \omega)\mathrm{d}\omega\right) = \left[\frac{1}{2}\hat{f}^2\omega\right]_{-\infty}^{+\infty} - \int |\hat{f}|^2 \, .$$

Assume further that f' is in $L^1(\mathbb{R})$. This implies that $\left[\frac{1}{2}\hat{f}^2\omega\right]_{-\infty}^{+\infty} = 0$. The inequality of Cauchy-Schwarz gives the result:

$$\frac{1}{2} = Re\left(\int \hat{f}'(\omega)(\hat{f}(\omega)\ \omega)\mathrm{d}\omega\right) \leq \sigma_t \sigma_\omega \, .$$

The case of equality is an ordinary differential equation of first order, leading to a Gabor 'chirp' (eq. 2.3).

□

To obtain a version of the Heisenberg inequality that applies to the Wigner-Ville distribution [54], we note that:

$$x^2|f(x)|^2 = \int x^2 W(f,f)(x,\cdot),$$

$$\omega^2|f(\omega)|^2 = \int \omega^2 W(f,f)(\cdot,\omega) \, .$$

Therefore, we have (under the assumption that $||f||_2 = 1$):

$$\iint (x^2+\omega^2) W(f,f)(x,\omega) \, \mathrm{d}\omega \mathrm{d}x = \sigma_t^2 + \sigma_\omega^2 \geq 2\sigma_t\sigma_\omega \geq 1 \, .$$

The Heisenberg's uncertainty principles shows the importance of Gaussians and Gabor filters for time-frequency analysis, as they achieve the best time-frequency resolution among

all signals. The energy of Wigner-Ville distributions is concentrated in rectangles of size $\sigma_t \times \sigma_\omega$ for Gaussians [1]. These rectangles, called *Heisenberg boxes* [105], can favor accuracy in time or in frequency (fig. 2.3), but the box area remains constant and equal to one half.

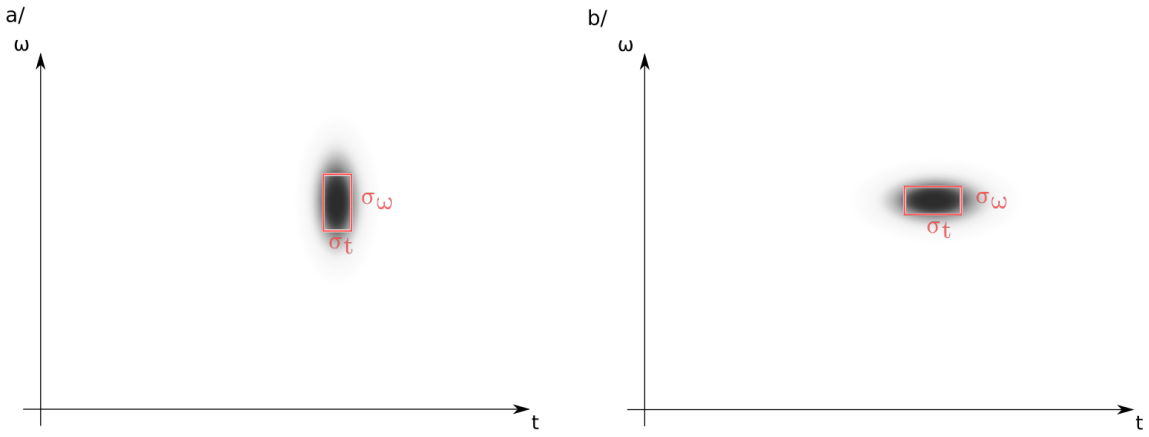


Figure 2.3 – *Heisenberg boxes* are an illustration of the time-frequency resolution trade-off: quadratic time-frequency representations can reduce uncertainty in time (**a/**) or in frequency (**b/**), but the concentration is limited by boxes of area 1/2.

2.3.2 Lieb's uncertainty principle

The Heisenberg's uncertainty principle gives a lower-bound of time and frequency spreads but does not have a direct interpretation in terms of coding efficiency or sparsity. Many other versions of the uncertainty principle exist [54, 134]. One formulation relevant to our context is the Lieb uncertainty principle that applies to quadratic time-frequency representations [98]. Let us go back on the cost function defined by eq. 2.1. The sparsity of the decomposition was related to the l_1-norm of a quadratic time-frequency representation. We can ask ourselves if a couple of input and window functions (f, g) can achieve the most sparse response, by minimizing the cost function:

$$\min_{f,g} ||\mathcal{W}(f,g)||_1 \,, \text{ subject to } ||f||_2 = ||g||_2 = 1 \,.$$

This problem for a fixed input f is called the problem of optimal window (which extends to l_p norms and other score functions) [90]. The solution to the quadratic version is given by the following result:

Theorem (Lieb's uncertainty principle [98])**.** *Assume $f, g \in L^2(\mathbb{R})$. The l_1-norm of the cross-Wigner distribution between f and g is bounded below by the product of the Euclidean norms:*

$$||\mathcal{W}(f,g)||_1 \geq ||f||_2 ||g||_2 \,. \tag{2.4}$$

The equality is achieved if and only if $g = \lambda f$ ($\lambda \in \mathbb{C}$) (up to modulation and translation) and f is a Gabor 'chirp' (eq. 2.3).

1. Note: Since Wigner-Ville distributions are quadratic, σ_t is in fact replaced with $\sqrt{2}\sigma_t$ (and σ_ω with $\sigma_\omega/\sqrt{2}$) compared to eq. 2.3

This result is closely related to the Prékopa–Leindler inequality. The latter is given below in a simplified context (the Prékopa–Leindler inequality stands with more general hypotheses).

Lemma (Prékopa inequality [128]). *Let f, g be two continuous functions of $L^2(\mathbb{R})$ and let h be defined for $x \in \mathbb{R}$ by:*

$$h(x) = \sup_t |f(x+t)\overline{g(x-t)}| \ .$$

Then

$$||h||_1 \geq ||f||_2 ||g||_2 \ .$$

Proof of Prékopa inequality. 1. Without loss of generality, I assume that f and g are real. For a measurable bounded set $A \subset \mathbb{R}$, I will denote $|A|$ the measure of A. The proof of the Prékopa inequality relies on the following inequality for two compact sets $A, B \subset \mathbb{R}$:

$$|A+B| \leq |A| + |B| \tag{2.5}$$

which is the Brunn-Minkoswki inequality in dimension 1.
Justification: $\forall (a,b) \in (A,B), \inf A + b \leq a + b \leq a + \sup B$ with equality i.i.f. $(a,b) = (\inf A, \sup B)$. Therefore, $A + \sup B$ and $B + \inf A$ define two subsets of $A+B$, of measure $|A|$ and $|B|$, and intersecting at a single point. The inequality follows directly from the inclusion.

2. We can write the l_1-norm of h as:

$$||h||_1 = \int_x \left(\int_0^{h(x)} 1 \right) \mathrm{d}x = \int_{\theta>0} |I_\theta(h)| \, \mathrm{d}\theta$$

where $I_\theta(h) = \{x | h(x) > \theta\}$. The following inclusion holds:

$$\frac{I_\theta(f^2) + I_\theta(g^2)}{2} \subset I_\theta(h) \tag{2.6}$$

Indeed, assume $(u,v) \in \big(I_\theta(f^2), I_\theta(g^2)\big)$, then

$$h(\frac{u+v}{2}) \geq |f(\frac{u+v}{2} + \frac{u-v}{2}) g(\frac{u+v}{2} - \frac{u-v}{2})| \ .$$

It follows from eq. 2.5 that:

$$||h||_1 \geq \int_{\theta>0} \left| \frac{I_\theta(f^2) + I_\theta(g^2)}{2} \right| \mathrm{d}\theta \geq \frac{1}{2} \int f^2 + \frac{1}{2} \int g^2 \ .$$

Using the inequality between the arithmetic and geometric means, we finally obtain the Prékopa inequality.

Note on the case of equality: If we assume that f and g are log-concave, then the case of equality is explicit. I assume further that f and g have their maximum at 0, are regular and normalized with $||f||_\infty = ||g||_\infty = 1$. The equality is achieved when the reverse inclusion for the level sets (eq. 2.6) also holds. Let (a,b) denote $(\sup I_\theta(f^2), \sup I_\theta(g^2))$ and (u,v) denote $(\frac{a+b}{2}, \frac{a-b}{2})$. The reverse inclusion implies that the function $t \mapsto f(u+t)g(u-v)$ has a maximum at $t = v$. It follows that $f'(a) = g'(b)$, and the reasoning extends to the lower bounds of the level sets. Because f and g are log-concave, the lower and upper bounds of the level sets describe $\mathbb{R}$, and we can verify that $f = g$. It can be shown that the case where f is log-concave and $g = \lambda T_x f$ for $\lambda, x \in \mathbb{R}$ is the only case of equality for the Prékopa inequality [44, 15].

□

Proof of Lieb's inequality.

$$||\mathcal{W}(f,g)(x,\cdot)||_1 = \frac{1}{\sqrt{2\pi}} \int |\hat{\varphi_x}(\omega)| \mathrm{d}\omega$$

with φ_x defined by eq. 2.2. We can also write:

$$||\mathcal{W}(f,g)(x,\cdot)||_1 \geq \sup_t \left| \frac{1}{\sqrt{2\pi}} \int \hat{\varphi_x}(\omega) e^{i\omega t} \mathrm{d}\omega \right| \ .$$

Using the inverse Fourier formula, we obtain:

$$||\mathcal{W}(f,g)||_1 \geq \int_x \sup_t |\varphi_x(t)| \, \mathrm{d}x \geq \int_x \sup_t |f(x+t)\overline{g(x-t)}| \mathrm{d}x \geq ||f||_2 ||g||_2 \ .$$

The last inequality is formally the Prékopa inequality.

I do not provide a formal proof for the case of equality. As stated above, the case of equality for the Prékopa inequality essentially gives $f = \lambda g$ up to modulation and translation. The case of equality in

$$||W(f)||_1 \geq ||f||_2^2 = \int_t \int_\omega W(f)(t,\omega) \, \mathrm{d}t \, \mathrm{d}\omega$$

is achieved if and only if $\mathcal{W}(f)$ is non-negative. This is only possible when f is a Gabor 'chirp' (Hudson's theorem).

□

Remarks:

1. The result by Lieb [98] is more general, as it includes all l_p norms for $1 \leq p < 2$:

$$||\mathcal{W}(f,g)||_p \geq \frac{1}{\pi} \left(\frac{\pi}{p} \right)^{1/p} ||f||_2 ||g||_2 \tag{2.7}$$

   The Gabor chirps are still the only functions that achieve the equality. The proof provided by Lieb follows a different strategy. First, the case $p > 1$ is proved, then the case $p = 1$ is inferred by taking the limit.
   To my knowledge, the demonstration presented in this chapter is original, although the link between the Prékopa inequality and Young's inverse inequality – which is a tool of the proof by Lieb – is well known [26].

2. Lieb's uncertainty principle stresses again the specificity of Gaussians and Gabor filters for time-frequency analysis. The interpretation in terms of sparsity is that Gabor filters are associated with the most sparse patterns for Wigner-Ville distributions. We also get the intuition that we would want g, the *analysis* function, resembling the input function f, at least near optimality [15], in order to achieve a sparse representation of the input.

2.4 – Gabor dictionaries

Standard Gabor frames. So far, time-frequency representations have been presented in the "continuous world", but that does not correspond to actual representations for practical applications, which are always computed on a discrete set of points of the time-frequency plane. The short-term Fourier transform can be approximated with a discrete sampling:

$$V_g(f)(m\Delta_t, n\Delta_\omega) =< f, M_{n\Delta_\omega} T_{m\Delta_t} g >, \quad m, n \in \mathbb{Z}$$

where Δ_t and Δ_ω are fixed positive parameters. These values are the dot products between f and the *atoms* of a Gabor system associated with a rectangular lattice:

$$\mathcal{G}(g, \Delta_t, \Delta_\omega) = \{M_{n\Delta_\omega} T_{m\Delta_t} g \,|\, m, n \in \mathbb{Z}\} \ .$$

The mapping

$$C : f \mapsto \{c_{mn}(f) =< f, M_{n\Delta_\omega} T_{m\Delta_t} g >\}_{m,n}$$

is called the coefficient map.

If the lattice on which the Gabor system is defined is too sparse, the input signals cannot be fully described by the coefficients. An important matter of discrete time-frequency analysis is to know when a Gabor system is complete and can provide further properties for a good description of the signal. Here is a glimpse of some results of the field of Gabor frames [68, 72]:

- If $\Delta_t \Delta_\omega > \frac{1}{2\pi}$, the the Gabor system cannot offer a complete representation.
- If g is a Gaussian and $\Delta_t \Delta_\omega \leq \frac{1}{2\pi}$, the Gabor system is complete.
- If g is Gaussian and $\Delta_t \Delta_\omega < \frac{1}{2\pi}$ (and only in this case), $\mathcal{G}(g, \Delta_t, \Delta_\omega)$ also satisfies the frame property, meaning that there exists $0 < A < B$ such that, for all $f \in L^2(\mathbb{R})$:

$$A||f||_2^2 \leq \sum |c_{m,n}(f)|^2 \leq B||f||_2^2 \ . \tag{2.8}$$

  The frame property provides many advantages to a Gabor representation (stability of the decompositions).
- At the critical sampling, $\Delta_t \Delta_\omega = \frac{1}{2\pi}$, the frame property is only possible at the extent of poor time and frequency localization for g (amalgam Balian-Low theorem). The interpretation of this property is that time-frequency representations with minimal properties of regularity and able to describe exhaustively any signal are necessarily overcomplete (redundant).

Flexible Gabor-wavelet atoms or α-*atoms* The decompositions that have been presented in this chapter have a unique resolution across frequencies: the window width is constant and the quality factor is linear with respect to frequency. But in some cases, we would want to have the resolution to be relative to the frequency, meaning that the window size is inversely proportional to the center frequency. Another family of decompositions allows this (multiresolution analysis): e.g. the constant-Q transform or the wavelet transform [105]. The Gabor transform and the wavelet transform have been studied in parallel, but mostly separately. They represent two distinct behaviors for the quality factor, and the intermediary cases have been less described. H. Feichtinger and M. Fornasier introduced the α-atoms to provide a consistent framework encompassing both Gabor and wavelet atoms [52]. Let the dilatation operator D_a be defined by

$$D_a(f)(t) = |a|^{-1/2} f(t/a),$$

then the $\alpha-$atoms are functions of the form

$$M_\omega T_x D_{\eta_{\alpha(\omega)}^{-1}} g \quad , (x,\omega) \in \mathbb{R}^2 \ .$$

$\eta_\alpha(\omega)$ is a function that fixes the resolution, defined for $\alpha \in [0,1]$ by:

$$\eta_\alpha(\omega) = (1+|\omega|)^\alpha \ .$$

Then the bandwidths are enlarged by the factor $\eta(\omega)$ and the quality factor satisfies:

$$Q(\omega) \propto \omega(1+\omega)^{-\alpha} \approx \omega^{1-\alpha} \ .$$

It corresponds to the parametrization used in the next chapters, with the relation $\beta = 1-\alpha$ (under the assumption $\omega \gg 1$). $\alpha = 0$ correspond to the standard Gabor transform and $\alpha \to 1$ tends to the wavelet transform. The flexible Gabor-wavelet transform is defined with the α-atoms by:

$$V_g^\alpha(f)(x,\omega) =< f, M_\omega T_x D_{\eta_{\alpha(\omega)}^{-1}} g > \ .$$

As in previous paragraph, V_g^α can be approximated with a discrete set of points (which is not anymore a rectangular lattice, since the discretization has to take into account the increasing bandwidths). Feichtinger and Fornasier showed that under some conditions (e.g. discrete set of points is sufficiently dense), the Gabor-wavelet coefficients satisfy the frame property (eq. 2.8).

Remark: Even if the transform is sampled on a discrete set of points, this set of points is infinite. What is more, the atoms are continuous functions. In numerical experiments, however, the atoms are also discrete by nature, and the integrals are approximated by a finite sum. Another point of difference with the following chapters is that atoms will be selected (quasi)randomly in the time-frequency space, and will not form a regular lattice.

Chapter 3
Statistical structure of synthetic signals

The purpose of the next two chapters is to describe the statistical structure of speech in relation to the acoustic properties of the speech signal. In this chapter, I study more specifically artificial signals that are expected to share the same structure than speech signals. The aim is to provide a first intuition of the most relevant acoustic features for the statistical structure of speech. Next chapter will focus on real speech data.

This chapter introduces the sparse dictionary method which is also applied in the following chapters. The analysis is based on the β parameter that describes the general behavior of frequency selectivity in the high frequency range.

Two kinds of artificial signals are analyzed. The first kind is time-modulated or frequency-modulated noise. This first kind of synthetic sounds relate to consonants. The second kind of artificial signals is sounds emitted by a uniform cylindrical waveguide with various radii that are similar to vowels.

3.1 – Methods

3.1.1 Overview

The general scheme of the method, common to the next chapters, is as follows:.

1. Create a set of overcomplete dictionaries W_β of Gabor wavelets from $\beta = 0.3$ to $\beta = 1.2$, uniformly distributed in time-frequency-phase.
2. Generate 16ms-slices X from speech (or speech-like) waveforms.
3. Pre-process the slices with filtering (high-pass filter at $f_c = 1.5kHz$) and normalization (root mean square value set to a constant).
4. Compute an average score reflecting (the lack of) sparsity of decompositions based on the l_1-norm (as motivated in previous chapters):

$$h(\beta) = \mathrm{E}\left(||W(\beta)^T X||_1\right)$$

5. Select $\beta^\star = \arg\min_\beta h(\beta)$.

Control parameter u: Specific to this chapter, I consider a control parameter u that tunes a characteristic of the signal. With this parameter, we can track the behavior of $\beta = \beta(u)$ with respect to this signal characteristic.

Analyses for synthetic data. Statistical analyses are made easy, or even superfluous, for artificial signals since the properties of structure are made evident by the generation. The

properties of structure for real data are generally less clear and are more of a trend that a general rule. The scores need to be averaged over a bunch of samples to make these properties explicit: this procedure will be detailed in next chapter.
There is still some averaging for the synthetic data. The two simulations that will be considered (first, modulated noises, then synthetics vowels) generate 200 samples associated with the control parameter going from $u = 0$ to $u = 1$. The first step is to decompose the samples in the Gabor dictionaries and to compute the scores. At the end of this step, we obtain a 200×30 score matrix $h(u, \beta)$. The matrix is then smoothed with a Gaussian filter ($\sigma = 1$) on the u-axis, and the values of $\beta^\star(u)$ are computed as the arg min of each row.

3.1.2 Gabor dictionaries

The candidates for the most sparse representations are a set of 30 overcomplete dictionaries whose atoms are Gabor filters (fig 3.1). Each dictionary is composed of 600 filters uniformly distributed in time, frequency and phase (low-discrepancy random sequence). The choice of Gabor filters is motivated mathematically (chapter 2) and is also consistent with the filter shapes found empirically with ICA [84, 94, 91]. When ICA is applied to sufficiently broad subclasses of speech sounds, the Q_{10} factor of the learned filters plotted against center frequency is well fitted by a line on a *log-log* scale. The intercept was found redundant with the slope of the regression in previous studies, since most of the lines cross around the point ($f_0 = 1kHz, Q_0 = 2$) for various speech data at the input of ICA [153, 47] [1]. The regression slope of Q_{10} on f_c is the β parameter that was introduced earlier in this thesis. It is a synthetic parameter of the representations obtained with ICA.

Each dictionary corresponds to a value of β. The range of β values is [0.3, 1.2], chosen so as to encompass all the values taken by β in the previous work by Stilp and Lewicki [153]. To ensure more diversity of filters, some randomness is added to the Q-factor with multiplicative noise. It follows that the quality factor of the Gabor filters is set by:

$$Q_{10}(f) = \log Q_0 + \beta(\log f - \log f_0) + 0.04\eta \tag{3.1}$$

where the log is taken in base 10 and η is i.i.d. noise drawn from the normal distribution. As for the other parameters which are uniformly distributed, the ranges were respectively [1–6.5 kHz] ,[2–14 ms] and [0, π] for center frequency, time shift and phase. The time shift did not cover the full range of the time window ($T = 16ms$) in order to avoid potential boundary effects.

1. Three values for Q_0 were tested : 1.5, 1.75, and 2. $Q_0 = 2$ was chosen as it provided the most sparse decomposition of actual speech data taken as a whole.

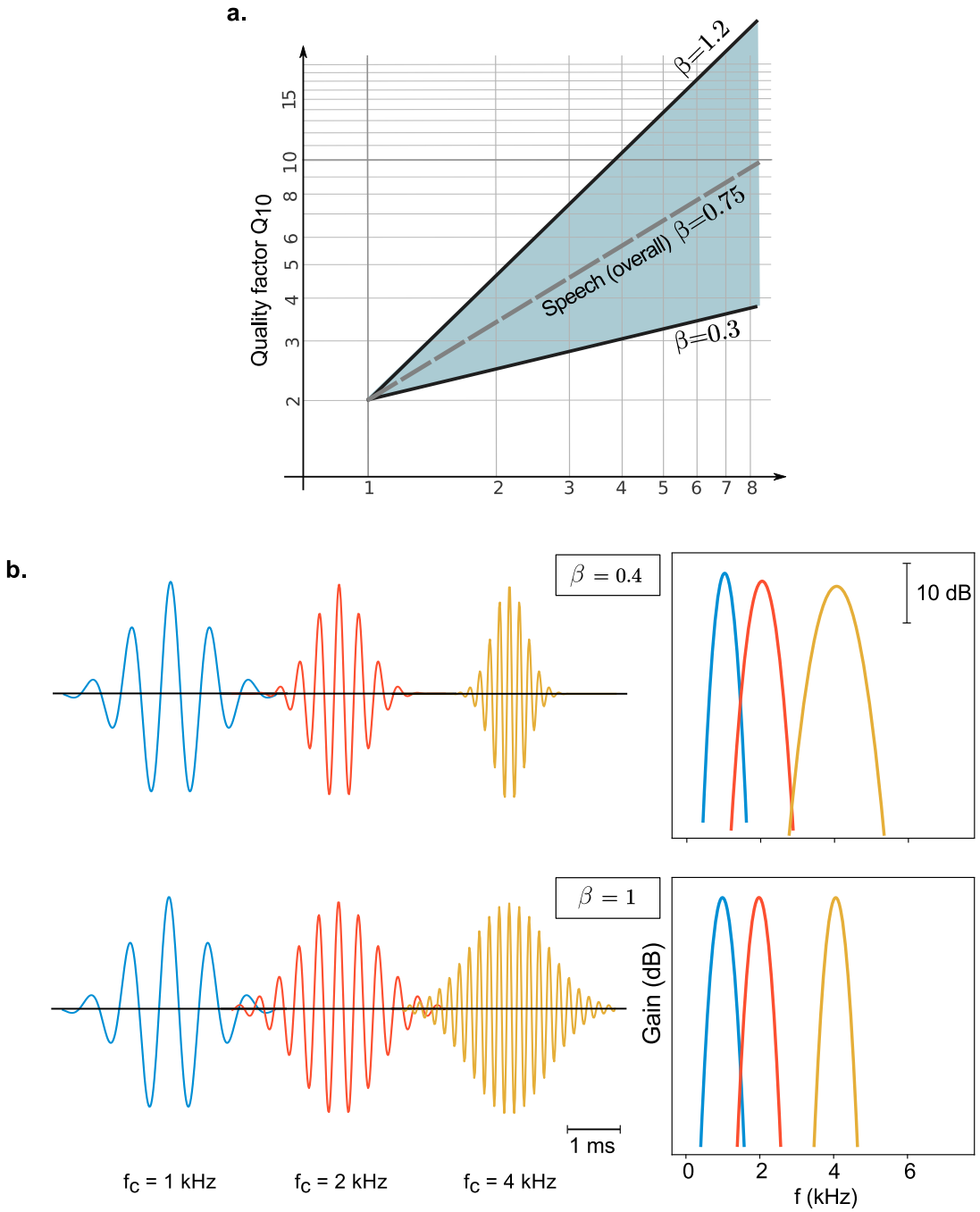


Figure 3.1 – **a.** The dictionaries follow different power laws for Q_{10}, with β ranging from 0.3 to 1.2.
b. Examples of Gabor filters used in the dictionaries.
left: time waveforms for two values of β at three values of f_c .
right: Corresponding frequency responses.

3.1.3 Cost function

In this chapter, I consider $n-$dimensional vectors X generated from a synthetic process. In next chapter, the input vectors will generated from the slicing of real speech data. The vectors X represent the time waveforms of preprocessed data on 16 ms. Preprocessing includes a filtering step (high-pass $f_c = 1.5kHz$) and a normalization step, so that all waveforms have the same root mean square value (the preprocessing step is described more in details for real data in next chapter). The goal is to select the best set of filters $W_\beta = (W_1, ..., W_m)$ among the Gabor dictionnaries indexed by β. I denote $Y_\beta = W_\beta^T X$ the output vectors that are produced by decomposing the input vectors in the Gabor dictionnaries. A raw measure of sparseness is expressed by the sum of response activities:

$$h_\beta(X) = ||Y_\beta||_1 = \sum_i |Y_{\beta,i}| \ . \tag{3.2}$$

This measure is the l_1-norm of the output vector. This cost function has been motivated in previous chapters. The actual cost function also includes weights as normalization factors (+2.5dB/octave): I explain how these weights were chosen in next chapter when I deal with real speech data. h_β is then averaged over a bunch of samples, and the cost function are normalized with h_β set to 1 for the less sparse signals.

The best β value minimizes the averaged cost function:

$$\beta^\star = \arg\min_\beta h_\beta \ .$$

In the rest of the thesis, β refers to $\beta^\star$, the optimal choice of the parameter, when there is no possible confusion with the other values.

Interpretation of the cost function. The cost defined in eq. 3.2 is a measure of the lack of structure (chap. 1 showed that it can be derived from a measure of entropy). Average values of h over the set of Gabor dictionaries were considered simultaneously with $\beta = \beta^\star$ in the analyses. β describes the general behavior of the optimal representation in relation to frequency selectivity, while h quantifies the computational cost to decompose the signal. Low values of h characterizes sounds that present structure, typically vowels. On the contrary, maximum values of β characterize sounds that are related to noise (obstruents: fricatives, stops...). Another interpretation of h that applies to many speech sounds is that it is a measure of localization. A signal with a single peak is associated with a minimum cost h: if this peak is on the time axis, β will also be minimal, if localization is in frequency, then β will be maximum. This interpretation is illustrated by the first simulation of artificial signals, on windowed noises.

A measure of the significance of $\beta^\star$ is **contrast**, defined by the relative difference between h_{max} and h_{min} over all values β:

$$h_{min} = \min_\beta h_\beta, \ h_{max} = \max_\beta h_\beta,$$

$$c = \frac{h_{max} - h_{min}}{h_{max}} \ . \tag{3.3}$$

Contrast c is small for a flat score function and big when the score function has a clear minimum.

3.1.4 Relation to other methods

This paragraph explains how the sparse dictionary method compares to other methods, and present some practical advantages. The limitations will be further discussed in next chapters.

The cost function is of the form of the general equation 1.10, with a Laplace prior on the marginal distributions. It was showed in chap. 1 that this cost function is similar to the entropy minimization formulation of Independent Component Analysis (eq. 1.3), without the $-\log|\det|$ penalty that only makes sense in the determined case. ICA has been the preferred method to investigate the statistical structure of speech in previous work. ICA is a non-parametric method that offers a rich analysis of specific data, generating a decomposition with filters of arbitrary shapes that are fully adapted. The results, however, are often difficult to interpret as such, and this difficulty is even greater when comparing representations learned on different parts of data. The parametric approach takes advantage of the synthetic parameter β that has proven to summarize many speech representations learned with ICA. It consists in using a constrained representation model based on the β parameter, instead of the entire ICA procedure. The interest of the parametric approach is to provide a unified and convenient framework for comparisons between representations of speech or speech-like data, given that the power law model fits very well the overall statistical structure of speech, or sufficiently broad classes of speech sounds.

An alternative method to compute β is to use ICA as a first step, then to estimate β directly on the learned representation, as was done in Stilp and Lewicki, 2013 [153]. However, the procedure is not easily automated since ICA, a gradient descent algorithm, is run each time we consider a new subset of data. The learned filters can also deviate from the parametric model in several ways, making it more difficult to provide a consistent interpretation of the estimated values. Instead, the approach of the dictionary method is to compare directly the data with the underlying model. An intermediate choice between the dictionary method and ICA would be to use a gradient descent to minimize the cost function with respect to β [122]. The interest of this method would be to obtain a more precise value of β, but it still has the disadvantage of relying on a gradient descent algorithm. The exact value of β depends on several experimental settings: the precise value of Q_0, defined in next subsection, the preprocessing of data, and the weighting strategy. I considered that what is important is not to obtain the most accurate values of β but to describe how values are distributed in the (β, h) plane when considering a variety of speech sounds.

The dictionary approach has several practical advantages. Unlike ICA, a non-parametric method that requires to estimate a large matrix, the dictionary method needs little data to obtain an estimate of β. While the results presented for real speech data will be averaged over hundreds of samples for statistical significance, a rough estimate of β can be computed on single examples, and we can see the evolution of β over time for a sentence. Some phenomena that will be discussed are already visible on single examples, such as the lowering of β at the onset of stops. Another practical interest is that the computations of the scores and the statistical analyses on β can be separated (see next chapter). Another advantage of the dictionary method is to have access to the global cost function. The measure of contrast, defined in previous subsection based on the global cost function, gives an idea whether the value of $\beta^\star$ is significant or if the best representations are poorly discriminated.

3.2 – Windowed noises

The first kind of generated signals is noise windowed in time or frequency. The purpose of this first simulation is to see the behavior of β on noise sounds localized in time or in frequency.

Motivation. Many speech sounds are noises that are the results from turbulent airflow occurring in the vocal tract. These are the sounds that are produced by obstructing the airflow, therefore called the obstruents: they are the fricatives, the stops, and the affricates.

Turbulent sources. In speech production, the two main sources of sound are:

1. vocal-fold vibration (e.g. for vowels) characterized by pressure drops at regular intervals, but the resulting airflow is laminar (vowels are the object of the second simulation).
2. turbulent noise

These two sources can be present at the same time (e.g. voiced fricatives [z], [ʒ]). Turbulence noise is the result of chaotic changes in air velocity. The two factors that determine whether the airflow is turbulent or not are the size of the channel and the volume velocity of the airflow. Turbulence noise is produced when the air has to flow through a narrow constriction. For example, the tongue almost blocks the passage of the air against the alveolar ridge (fig. 3.2) in the production of the fricative [s]. Turbulence noise can also be generated when constriction is wider but airflow has a higher rate: this is the case in particular of the aspirate [h] which is produced by the passage of the air through the glottis. Turbulence without any additional filtering typically generates a pink noise with a spectral tilt of -6dB/octave [83].

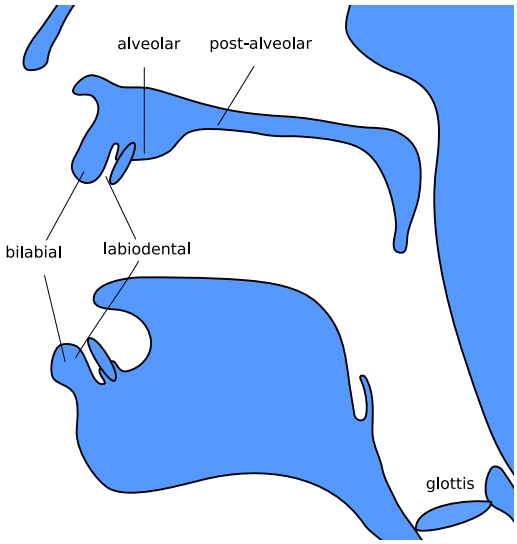


Figure 3.2 – Main places of articulation mentioned in the analyses. *Adapted from Wikimedia Commons.*

Fricatives. Fricatives are static turbulent noises made by narrowing the vocal tract at a given location. Depending on the place of constriction (fig. 3.2), fricatives have different signal properties. The main reason of these differences is that most often, the noise that is

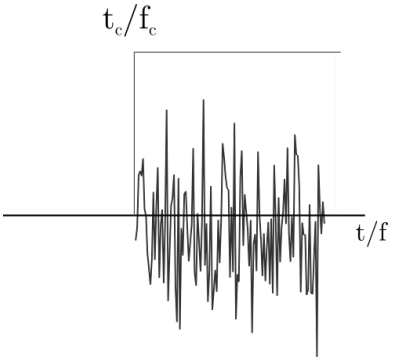
produced by the turbulence is then filtered by the vocal tract. The aspirate [h] produced at the glottis undergoes the same filtering as a vowel, therefore has a similar formant structure. For other places of articulation, fricatives are most affected by the front cavity, extending from the narrowed channel to the lips [151, 83]. The resonances can be computed as for a cylindrical waveguide (eq. 3.12, derived further for the second similutation on vowels): $f_j = \frac{c}{4L}(2j+1)\,,\quad j = 0, 1, 2\cdots$, but as the length of the cavity is short, the first resonance is already in the high frequency range (e.g. 6kHz for L=1.5cm). As a consequence, a simple model for sibilant fricatives (place of articulation behind the teeth) is high-pass filtered noise.

Stops and affricates. Stops and affricates have a more complex structure than fricatives because their characteristics change during production. They have the following temporal pattern: closure - release burst - release transition. During the closure, no sound is emitted (or a low frequency sound only as in [b]). We are not interested in low frequency information and we can always ignore this part. After the closure, the release can be divided into two phases. The first phase is the burst following the instant of the occlusion release. During this phase of small duration (few milliseconds), the pressure impulse creates one or several significant jumps in intensity, and the power spectrum is typically flat. In the second phase – the opening or aspiration phase – there is still some obstruction at the place of articulation and/or aspiration at the glottis, resulting in sounds similar to fricatives [83]. Voicing can also start to occur (transition to vowel). Affricates have their opening phase that is more easily associated with a fricative, and typically longer than stops (in the exception of aspirated stops, who have a similar time pattern, but are considered apart). Therefore, affricates can be seen as the concatenation of a stop and a fricative, but whose transition is not enough clear to perceive two distinct phonetic units. The opening phase of stops and affricates can be modeled as for fricatives. On the other hand, because of the sudden increase in intensity at stop bursts, a model for the onset of stops is noise windowed in time.

3.2.1 Mathematical modeling

I provide a simple mathematical heuristics for the coding of windowed noises. The purpose is not to explain thoroughly the results of the numerical simulations, but to provide an insight into the expected behavior.

Model. The input is Gaussian white noise multiplied by a semi-infinite rectangular window in time or in frequency (see image on the right). I limit the analysis to finite intervals in time ($[-\Delta t/2, \Delta t/2]$) and in frequency ($[0, \Delta f/2]$). The signals are real in time, conjugate symmetric in frequency. I ignore the difficulty of complex and symmetric signals in frequency: hence, the analysis is made with the assumption that signals are windowed in time but the reasoning could be translated in the frequency domain as well. The *cut-off* frequency f_c (or time t_c) is considered to be localized around the mid-point of the interval. Noise can occur after or before with the same probability.



Modeling with Gaussian filters of varying selectivity. We consider Gaussian filters (without any oscillation) arranged on the time (or frequency axis) as in fig. 3.3, normalized with respect to the l_2-norm, and of constant width $\sigma = \sigma_t$ (or σ_f). Suppose that the cut-off is at $t = 0$, the response of a Gaussian filter at $t = \mu$ is a Gaussian process of variance

$$\chi_{\sigma,2}(\mu) = \int_{-\mu}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp -\frac{t^2}{2\sigma^2} \mathrm{d}t = \frac{1}{2} + \frac{1}{2}\mathrm{erf}(\frac{\mu}{\sqrt{2}\sigma}) \leq 1$$

with the standard deviation of the noise set to 1. The error function erf is antisymmetric, therefore the mean variance does not depend on σ_t:

$$\frac{1}{\Delta T} \int_{-\Delta T/2}^{\Delta T/2} \chi_{\sigma,2}(\mu) \mathrm{d}\mu = 1/2 \ .$$

On the other hand, the absolute moments p of the Gaussian processes are related to the second moments by:

$$\chi_{\sigma,p}(\mu) = \sqrt{\frac{2^p}{\pi}} \Gamma\left(\frac{p+1}{2}\right) \chi_{\sigma,2}(\mu)^{\frac{p}{2}} \ .$$

The root breaks the symmetry of the variance function. For $1 \leq p < 2$, we use that and if $x \in [0,1]$, $x^{p/2} \geq x$ with equality i.i.f $x = 0$ or $x = 1$. Therefore:

$$\frac{1}{\Delta T} \int_{-\Delta T/2}^{\Delta T/2} \chi_{\sigma,p}(\mu) \mathrm{d}\mu \geq \sqrt{\frac{2^p}{\pi}} \frac{1}{2} \Gamma\left(\frac{p+1}{2}\right) \ ,$$

in particular, in the case $p = 1$, corresponding to our setting,

$$\frac{1}{\Delta T} \int_{-\Delta T/2}^{\Delta T/2} \chi_{\sigma,1}(\mu) \mathrm{d}\mu \geq \frac{1}{\sqrt{2\pi}}$$

with equality in the limit $\sigma \to 0$ (we can show in addition that the left sides of the inequalities above are increasing functions of σ).

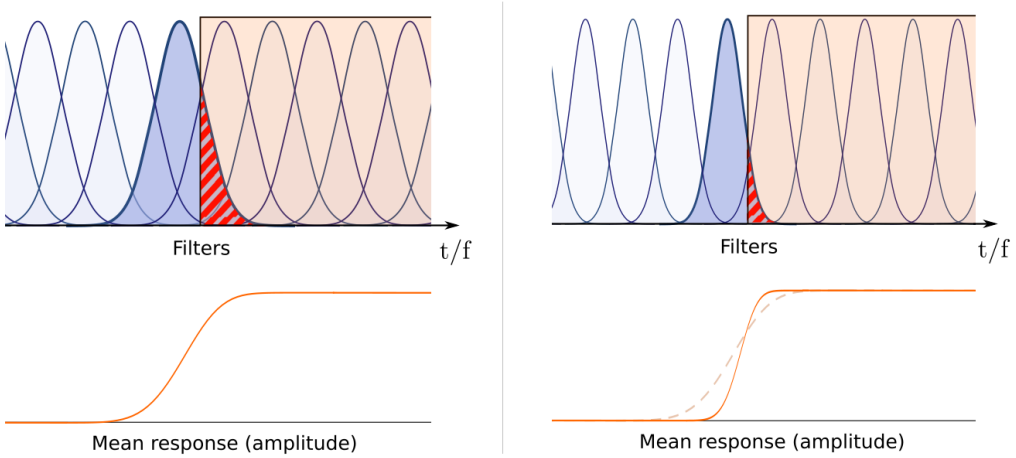


Figure 3.3 – Examples of two situations with Gaussian filters arranged on the time (or frequency) axis. The Gaussian filters differ by their selectivity (left: wide filters, right: narrow filters). The red areas represent the windowed noise. Bottom: Mean response (amplitude) as a function of the position of the filter. Selective filters improve the sparsity of the response.

As shown in chap. 1, the minimization of the l_p-norm corresponds to the minimization of cross-entropy terms with a generalized Gaussian model, in particular a Laplace distribution in the case $p = 1$. In the case of windowed noises, the coding strategy for reducing entropy in neural channels amounts to reduce the mean amplitude value at the output of each component. This simple model shows that the most efficient strategy for a transformation that maintains the same power as the input, is to have filters with no response along with filters at full output. This strategy increases the sparsity of the signal decompositions as it puts as many outputs as possible with zero response.

3.2.2 Generation

Overview. 200 samples of 16 ms were generated with the following procedure. Initially, the noise sounds are samples of Gaussian noise filtered by a low pass filter of order 1 (cut-off at 2 kHz). The parameter u controls localization in time or in frequency ($u = 0$: samples localized in time, $u = 1$: samples localized in frequency).
At $u = 0$, the noise is windowed by a Gaussian of time deviation $\sigma_t = 0.01 \times T = 0.16ms$. At $u = 1$, the sample of noise is convolved by a Gaussian filter of frequency deviation $\sigma_f = 0.01 \times f_s/2 = 80Hz$. The sounds of intermediate values are shifts of these two configurations. From 0 to 0.5, they go through time modulation essentially with σ_t increasing. From 0.5 to 1, they go through frequency modulation with σ_f decreasing. $u = 0.5$ is not very different from non-modulated white noise. Some values for the modulation widths are given in Table 3.1. Next paragraph provides further details on the generation, and figure 3.5 shows some examples of generated samples.

Table 3.1 – **Correspondence between the control parameter u and modulation widths (windowed noises).**

u	σ_t (ms)	σ_f (Hz)
0	0.2	$\star$
0.2	1.0	$\star$
0.4	4.5	$\star$
0.6	$\star$	2000
0.8	$\star$	470
1	$\star$	80

Details on the generation. The two types of modulations are:

— *time modulation:* multiplication by a Gaussian in time. The time width of the Gaussian, σ_t, is linked to u by the equation:
$d_t(u) = \frac{\sigma_t}{T} = A + \frac{B}{1-u/u_{max}}$
where $T = 16ms$ (time of the window).

— *frequency modulation:* multiplication by a Gaussian in the frequency domain, equivalent to a convolution in time by a Gaussian of time width $\sigma_t = 1/(2\sigma_\omega)$. σ_ω is linked to u by:
$d_\omega(u) = \frac{\sigma_\omega}{\Omega} = A + \frac{B}{u/u_{max}}$
where $\Omega = 2\pi f_{NY} = \pi f_s$ ($f_s = 16kHz$).

The values of the constants were set to $A = -0.070, B = 0.080, u_{max} = 0.52$. This corresponds to $d_t(0) = d_\omega(1) = 0.01$ and $d_t(0.5) = d_\omega(0.5) = 2$. The modulation widths are plotted in fig 3.4. The modulations were applied on each sample in a random order. The central times and center frequencies were chosen randomly in the ranges of 4 to 12 ms and 2 to 6 kHz.

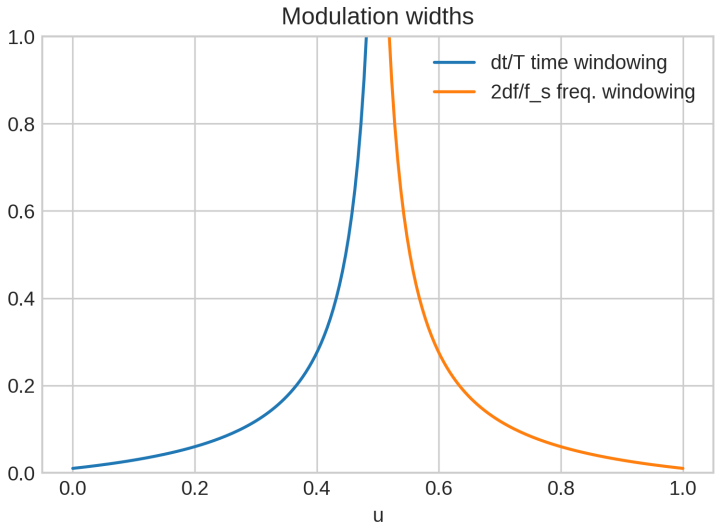


Figure 3.4 – Modulation widths as a function of the control parameter.

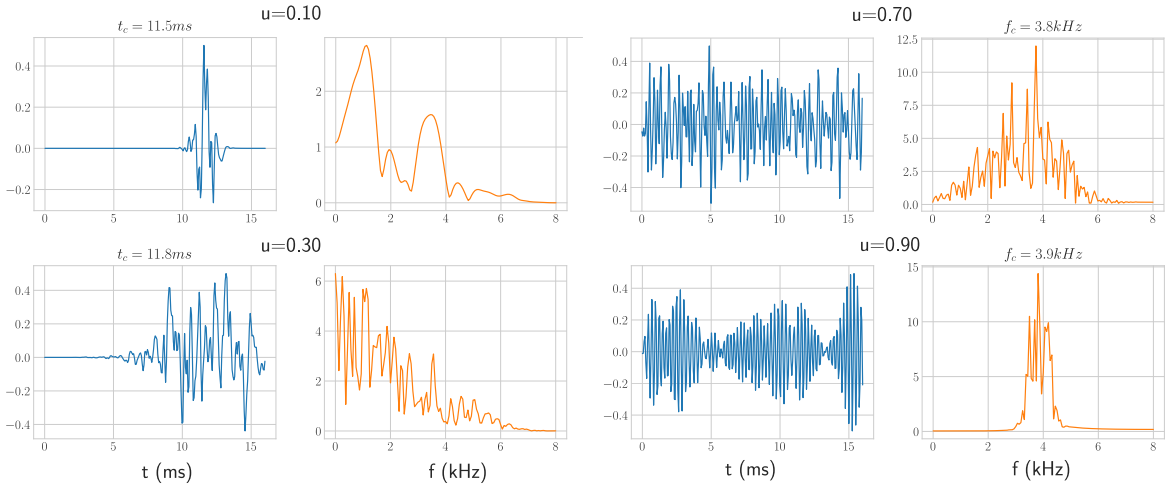


Figure 3.5 – Examples of generated samples for four values of u.
Blue: time waveforms.
Orange: frequency profiles.

3.2.3 Results

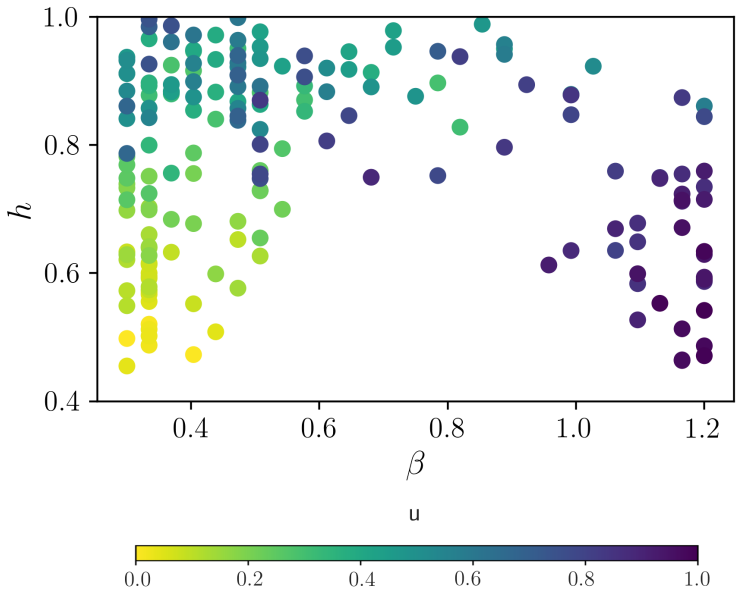


Figure 3.6 – β (offering the best decomposition) and h as a function of the control parameter u for windowed noises ($u = 0$, time windowing, $u = 1$, frequency windowing).

The mathematical modeling showed that we would want to have selective filters on the axis (time/frequency) of the modulation. More generally, the optimal decomposition of noise sounds shifts toward a time (or frequency) decomposition if it has a sharp power increase/decrease in the time (or frequency) domain. The simulation on modulated noises (fig. 3.6) illustrates this fact. β takes the lowest value (poor frequency selectivity, time representation) when the noise is multiplied by a Gaussian function localized in time. Then, β increases up to a median value as the Gaussian expands. At the same time, h increases because any structure is lost. Halfway through the simulation, at $u = 0.5$, the generated samples lose most structure, and h is maximum. β has a rather erratic behavior and the score function becomes flatter as indicated by the low contrast value (fig. 3.7). The symmetrical pattern occurs when the simulation goes on frequency modulations. At $u = 1$, β is maximum.

3.3 – Synthetic vowels

The second kind of generated signals is sounds emitted by a uniform cylindrical waveguide with different values for the cross-sectional area. The goal is to analyze vowel-like sounds with various bandwidths.

Motivation. The reflections of the acoustic waves in the vocal tract result in resonant frequencies that are evenly arranged on the frequency axis (modes). These peaks on the spectrum, visible for any vowel, are called the formants. The formants, especially the first formants F1–F3, vary as a function of the vocal tract configuration, and characterize the categories of vowels. The formant values are not the first variables of interest. What is

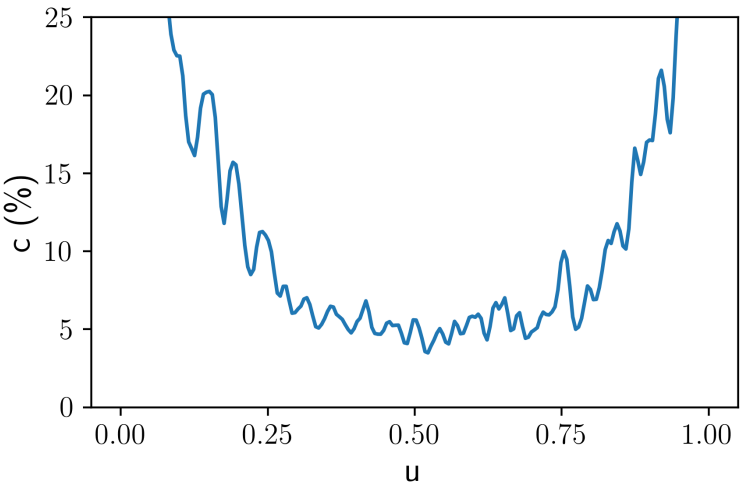


Figure 3.7 – Contrast as a function of u for windowed noises

more significant for the statistical structure of the signal is formant bandwidths, as larger bandwidths will correspond to a lower quality factor for the best decomposition.

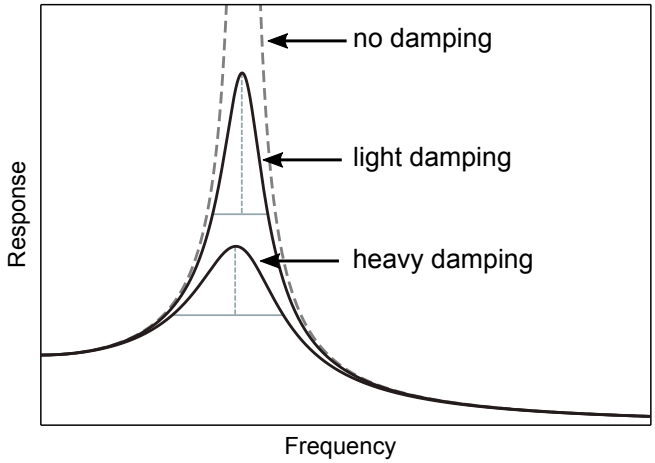


Figure 3.8 – The quality factor (or bandwidth) of a filter is related to the degree of damping for a physical system. Damped systems have less sharp resonances. Examples for a 2nd order low-pass filter.

Formant bandwidths are determined by the level of damping (fig. 3.8), caused by acoustic losses. To make this explicit, let's consider a second order band-pass filter, defined by its (canonical) transfer function:

$$H(jx) = \frac{jH_0/Qx}{1 - x^2 + jx/Q} = \frac{1}{1 + jQ(x - 1/x)}$$

where $x = \omega/\omega_0$, $j^2 = -1$. The other variables are parameters. For such a system, the quality factor at X dB is related to the parameter 'Q' with:

$$Q_{XdB} = \frac{Q}{\left[10^{X/10} - 1\right]^{1/2}} \ .$$

In general the Q_{3dB} factor is chosen ($10^{X/10} \approx 2$) and identified as the Q factor. The solutions of the homogeneous equation corresponding to the transfer function are damped oscillations. The envelope has an exponential decay $A\exp^{-\omega_0\zeta t}$. The damping ratio ζ is related to the Q factor by

$$\zeta = \frac{1}{2Q} \ .$$

A generalization of this relation is the rule of thumb [152] :

$$Q = \omega \frac{\text{total stored energy at frequency f}}{\text{power dissipated at frequency f}} \ .$$

The contribution of each type of acoustic losses can be evaluated following this rule. The important factors are wall losses at low frequencies and radiation losses at high frequencies [48, 152]. Losses at the glottal opening can also be significant but are limited in high frequencies >2kHz. The other losses (heat conduction, viscosity...) have a smaller impact. Since higher formants play a greater role in the determination of β, a key factor for the statistical structure of vowels is likely the degree of acoustic radiation at the lips. The purpose of the synthetic examples is to simulate this effect, in the simplified setting of a uniform cylindrical waveguide. The level of acoustic radiation depends on the termination impedance, which increases with frequency and lip opening [53, 7]. The estimation of acoustic radiation with aperture radius, however, is only an approximation, in particular it does not take into account inner reflexions [48].

3.3.1 Generation

Basics of acoustics. Before describing the simulation of sounds emitted by a uniform cylindrical waveguide, some elementary notions and definitions of acoustics are briefly introduced.

Wave equation.
In linear acoustics, we are interested in the evolution of two variables: the local deviations of the pressure p (different from the static pressure p_0 which remains overall constant) and particle velocity $\mathbf{u}$. These variables obey to the following equations [24]:

— *From the equation of state:* a relation between p and the local deviations of the density ρ:

$$\frac{\rho}{\rho_0} = -\frac{\delta^2 V}{\delta V} = \kappa p$$

where κ is the compressibility of the medium.

— *Conservation of mass:*

$$\frac{\partial \rho}{\partial t} + \rho_0 \operatorname{div} \mathbf{u} = 0 \ .$$

— *Newton's second law:* applied locally

$$\rho_0 \frac{\partial \mathbf{u}}{\partial t} = -\nabla p \ , \tag{3.4}$$

which is in fact obtained from the linearization of the Euler's equation for the quantity of movement.

The combination of the three equations above leads to the wave equation:

$$\Delta p - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = 0 \tag{3.5}$$

where $c = \frac{1}{\sqrt{\rho_0 \kappa}}$ is the phase velocity of the waves (in air, speed of sound is 343 metres per second). Its version in the frequency space is the Helmholtz equation:

$$(\Delta + k^2)\hat{p} = 0 \tag{3.6}$$

with $k = \omega/c$.

Acoustic impedance.
A solution of eq. 3.5 (in one dimension) is the plane, harmonic, progressive wave:

$$\underline{p}(x,t) = A\exp\left(i(\omega t - kx)\right) \ . \tag{3.7}$$

Here, I used the complex representation of the scalar wave, with $p = Re(\underline{p})$ (p is identified with its complex representation in the following). Substituting in eq. 3.4, we obtain:

$$\rho_0 i\omega \underline{u} = ik\underline{p} \Rightarrow \underline{p} = Z_0 \underline{u}$$

with $Z_0 = \rho_0 c$, the characteristic acoustic impedance of the medium. The notion of acoustic impedance (denoted Z) can be extended to more complex configurations. Its value (in Pa.s/m) is defined by the ratio between pressure and velocity. Z plays the role of a transfer function in many situations. The calculations in acoustics have many similarities to electrical problems, using Z as an analogy to electrical impedance and the relation between p and u as an analogy to Ohm's law. The real part of Z is called the *resistance*, the imaginary part is called the *reactance*.

Cylindrical waveguide.
Let's consider a cylindrical waveguide of length L and of cross-sectional area $S = \pi r^2$ (fig. 3.9). The left end is a rigid boundary with imposed vibrations $u = u_g$ (glottis) and the right end is open, imposing $p(L) = 0$. We study longitudinal plane waves in the cylinder.

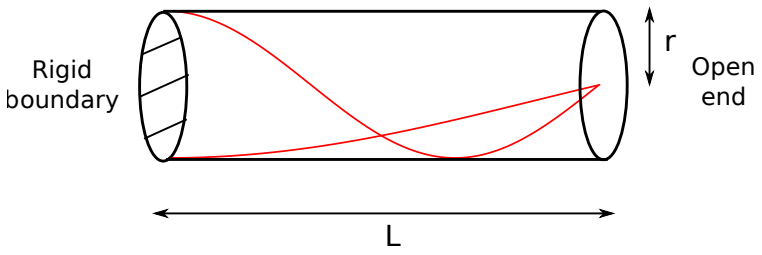


Figure 3.9 – Cylindrical waveguide with one open and one closed end. *In red:* the two first modes associated with the acoustic pressure.

We look for solutions of the form $p(\mathbf{x}) = f(x)g(r,\theta)$.
We can find the solutions with the method of separation of variables. Equation 3.5 gives:

$$\frac{1}{f}\left(\Delta_x f - \frac{1}{c^2}\frac{\partial^2 f}{\partial t^2}\right) = \frac{\Delta_r g}{g} \tag{3.8}$$

where Δ_x is the second partial derivative with respect to x and Δ_r the Laplacian in polar coordinates. Each side depends on different variables, therefore constant. If the dependence of r and θ can be separated, the solutions to the eigenvalue problem

$$\Delta_r g = k_r^2 g$$

are solutions of the Bessel's equations [24, 114]. The detailed computation is beyond the scope of this thesis. The rigid boundary imposes $\mathbf{u} = 0$ on the edges of the cylinder, limiting the set of solutions to a discrete set. The modes can be denoted by two integers (m, n), where m is the number of nodal planes and n is the number of nodal cylinders parallel to the axis (fig. 3.10).

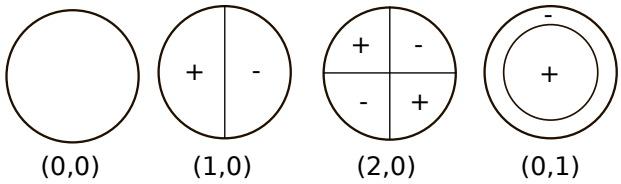


Figure 3.10 – First transverse modes for a cylindrical waveguide. The boundaries between positive and negative sides are the nodal planes and cylinders.

The second equation (left-hand side in eq. 3.8) depends on the mode with:

$$\frac{\partial^2 f}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 f}{\partial t^2} - k_r^2 f = 0 \ .$$

Moving to the frequency domain, we get the modified Helmholtz equation:

$$(-k_x^2 + w^2/c^2 - k_r^2)\hat{f} = 0$$

and the dispersion relation:

$$\omega^2/c^2 = k_x^2 + k_r^2 \ .$$

Only the (0,0) mode ($k_r = 0, g = \text{cte}$) can propagate at all frequencies and is associated with a linear dispersion relation. The other modes only propagate above cut-off frequencies. I will now consider only the first mode. For realistic simulations of the vocal tract, however, higher modes can have an impact on the computation of the radiated energy for frequencies above 5kHz for the largest apertures at the lips [8].

Considering only the first mode, the wave equation is equivalent to the 1D equation:

$$\frac{\partial^2 p}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = 0 \ .$$

The solution with the boundary conditions on the ends of the cylinders is a standing wave. Each mode (fig. 3.9) is the sum of two harmonic plane waves (eq. 3.7) : u_+ moving forward and a reflected wave u_- moving in the opposing direction. We have:

$$u = u_+ + u_-, \quad p = Z_0(u_+ - u_-) \ .$$

For harmonic waves, the amplitudes and phases of the two waves can be recovered from a system of 2 equations given the pressure and velocity at one point or the pressure at any couple of points x and $x + dx$ on the cylinder axis [62]. We also have the relation:

$$\begin{bmatrix} Zu(x+dx) \\ ip(x+dx) \end{bmatrix} = R\left(\frac{\omega dx}{c}\right) \begin{bmatrix} Zu(x) \\ ip(x) \end{bmatrix} \tag{3.9}$$

where $R(\theta)$ denotes the matrix of rotation of angle θ. Extending to the total length of the waveguide, we get:

$$\begin{bmatrix} Z_0 u(L) \\ 0 \end{bmatrix} = R\left(\frac{\omega L}{c}\right) \begin{bmatrix} Z_0 u_g \\ ip(0) \end{bmatrix} \ . \tag{3.10}$$

From the second equation, we get a relation between $p_g = p(0)$ and u_g, that allows us to define the equivalent impedance of the waveguide as seen from the closed end (glottis):

$$p_g = Z_{eq} u_g$$

$$Z_{eq} = iZ_0 \tan\left(\frac{\omega L}{c}\right) \ . \tag{3.11}$$

We can find from this equation the resonances (formant frequencies):

$$f_j = \frac{c}{4L}(2j+1), \quad j = 0, 1, 2 \cdots \tag{3.12}$$

The vocal tract does not have a uniform cross-sectional area. If the cross-sectional area $S(x)$ is allowed to change, the acoustic impedance must be replaced with $Z = \frac{\rho_l}{S(x)}$ in eq. 3.9 where ρ_l is the linear density [41, 62]. Changes in the configuration of the vocal tract affect the values of the formants (in particular the first formants), producing the vowels, associated with different perceptual cues. Because the purpose of the simulation is to recover the overall statistical structure of speech for vowels and not the formant structure specific to some vowels, I stick to the simpler uniform cylindrical waveguide in the simulations.

Acoustic radiation.
The mechanical vibrations of a localized object produce sounds that propagate in the air forming in general spherical waves. This process is called radiation. Insight into radiation problems can be gained from examining spherical wave solutions to the Helmholtz equation (eq. 3.6). A solution is:

$$p(r) = \frac{A}{r} \exp\left(i(\omega t - kr)\right) \ . \tag{3.13}$$

From eq. 3.4, we found the corresponding velocity field:

$$\mathbf{u}(r) = \frac{i}{\omega\rho}\nabla p = \frac{1}{\rho c}(-\frac{i}{rk} + 1)p \ \mathbf{n}$$

where $\mathbf{n}$ is a radial unitary vector. The impedance associated with this setting is:

$$Z = Z_0 \left[\frac{(kr)^2}{1+(kr)^2} + i\frac{kr}{1+(kr)^2}\right] \ . \tag{3.14}$$

We can note that $Z \to 0$ when $k \to 0$. The radiation process is similar to passing a high-pass filter on the vibrations (to obtain the pressure field). Small objects have difficulty radiating energy in low frequencies. We are used to experiencing this phenomena: for example, small loudspeakers do not reproduce bass sounds well. Radiation occurs at the

open end of the acoustic waveguide (at the lips) and results in acoustic losses which are greater in high frequencies.

Generation model. For the simulation of vowel-like sounds, I generated 200 samples sounds emitted by a cylindrical waveguide. A simple model for acoustic losses due to radiation is to place a radiating half-sphere with a cross sectional area equal to the lip opening at the end of the cylinder [150, 53]. We can use the impedance of a radiating spherical object (eq. 3.14) as a termination impedance for the acoustic waveguide. Equation eq. 3.10 becomes:

$$\begin{bmatrix} Z_0 u(L) \\ iZu(L) \end{bmatrix} = R\left(\frac{\omega L}{c}\right) \begin{bmatrix} Z_0 u_g \\ ip(0) \end{bmatrix} \ . \tag{3.15}$$

and the equivalent impedance seen from the glottis is:

$$Z_{eq} = Z_0 \frac{Z\cos(kL) + iZ_0 \sin(kL)}{Z_0 \cos(kL) + iZ\sin(kL)} \ . \tag{3.16}$$

Multiplying the first row by $\cos(kL)$ and the second raw by $\sin(kL)$ in eq. 3.15, and by summing, we get the pressure at the lips:

$$p(L) = Z \frac{Z_0}{Z_0 \cos(kL) + iZ\sin(kL)} u_g \ . \tag{3.17}$$

I used this last relation and a model of the glottal flow u_g (see next paragraph) to generate the samples. To account for other surface losses in the waveguide, k was substituted with $k = \omega/c - j\alpha$, $\alpha = 1.2e-5\sqrt{\omega}/0.01$ as in Hanna et al. 2016 [69]. The waveguide length was similar to the length of the vocal tract (in average 16.5 cm). The control parameter u controls linearly the radius r of the cylinder ranging from $r = 0.2cm$ ($u = 0$) to $r = 1.3cm$ ($u = 1$). Some values of u and r are given in Table 3.2. Figure 3.11 shows the power spectrum of two sounds generated with different radii.

Table 3.2 – **Correspondence between the control parameter u and aperture radius (synthetic vowels).**

u	r (cm)
0	0.20 ([ʊ])
0.2	0.42 ([u], [ɔ])
0.4	0.64 ([ʌ], [ɪ])
0.6	0.86 ([i], [ɛ])
0.8	1.08 ([æ])
1	1.30 ([ɑ])

For reference, I give examples of vowels that have comparable sectional area in the hypothesis of a circular aperture (see Story, 1996 [155] for more accurate data).

Glottal flow. The last paragraph described the generation of the samples but the signal u_g was not specified. I used the *Lijencrants-Fant* model (*LF model*) [51, 49] to simulate the glottal flow. This model provides two equations for the flow derivative:

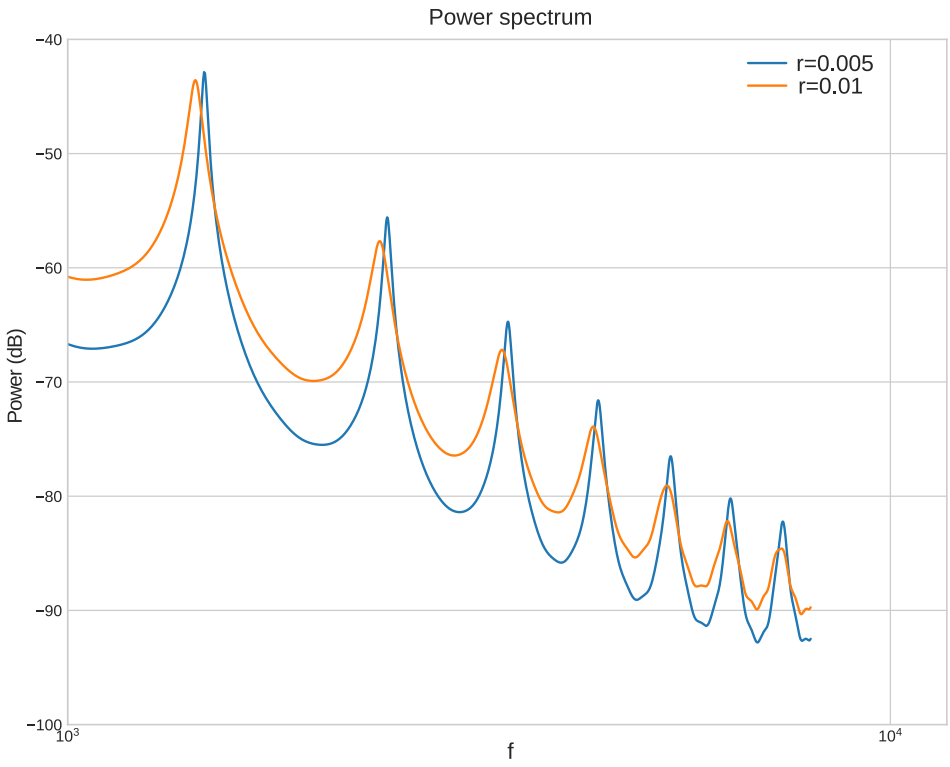


Figure 3.11 – Power spectrum of sounds emitted at the output of a cylindrical waveguide with two different radii. *Blue:* r=0.5cm. *Orange:* r=1cm.

$t < t_e$:

$$\frac{dU(t)}{dt} = E(t) = E_0 e^{\alpha t} \sin(\omega_g t)$$

$t > t_e$:

$$\frac{dU(t)}{dt} = E(t) = -\frac{E_e}{\epsilon t_a}\left[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}\right]$$

ϵ satisfies $\epsilon t_a = 1 - e^{-\epsilon(t_c-t_e)}$. For t_a small, $\epsilon = 1/t_a$.
The model is fully characterized by four parameters (fig. 3.12): the times t_p, t_e, t_a and E_e. It does not include t_c which is the glottal period, or considered infinite if t_a is small. t_p sets ω_g with $\omega_g = \pi/t_p$. α is found from the other parameters using the constraint that the flow is zero at the end of the cycle.

The glottal flow has three characteristic phases:

— $t < t_p$: glottal opening. The flow increases slowly.

— $t_p < t < t_e$: glottal closure. The flow decreases abruptly for a short amount time. This is the most important part as the glottal flow presents its steepest slope at this instant: it is the main excitation for high frequencies, and the steepness of this phase controls the spectral tilt (~-12dB/octave).

— $t > t_e$: return phase. It is the end of the glottal closure but with a slighter decrease.

Parameters for the simulation: t_a = 0.10 ms, t_p= 3.5 ms, t_e= 4.5 ms.

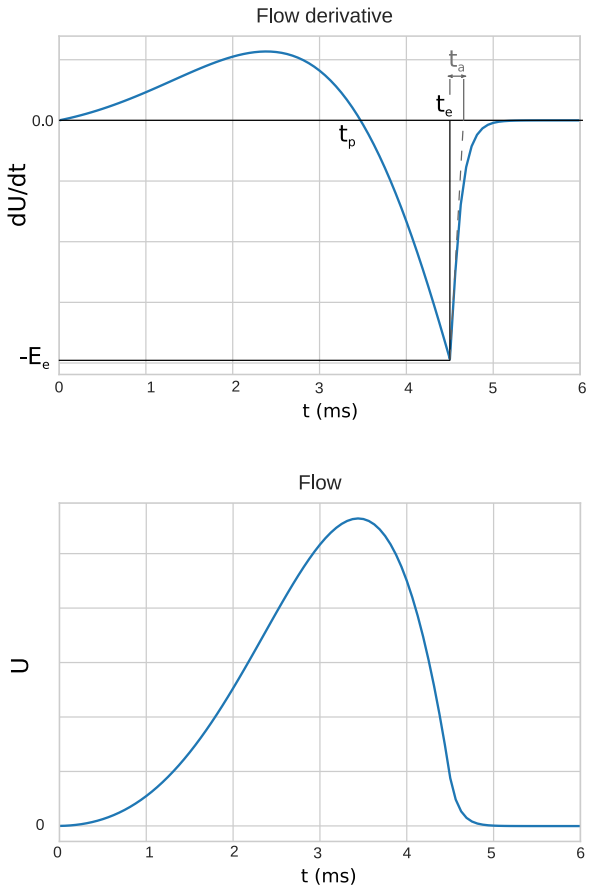


Figure 3.12 – The LF-model: the glottal flow is characterized by four parameters: t_p, t_e, t_a and E_e.

3.3.2 Results

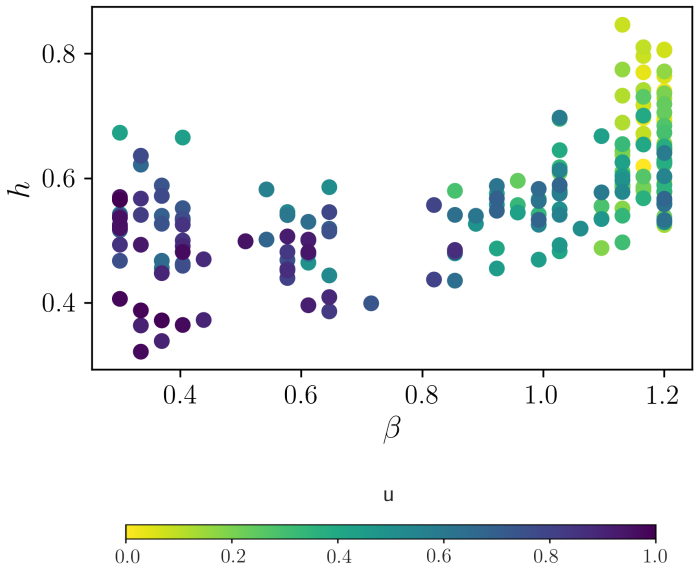


Figure 3.13 – β and h as a function of the control parameter u for sounds emitted by a cylindrical waveguide with different radii ($u = 0 : r = 0.2cm$, $u = 1 : r = 1.3cm$).

The simulation shows the impact of aperture radius on β with synthetic vowels generated by a uniform cylindrical waveguide (fig. 3.13). We obtain a maximal value of β with a small aperture ($u = 0$) associated with low acoustic losses and narrow bandwidths. We get the opposite for a large aperture ($u = 1$). The two end points are separated by a steep phase transition occurring between $u = 0.5$ and $u = 0.7$ (or $r = 0.7\,cm$ and $r = 0.9$ cm), where contrast is minimum (fig. 3.14). Note that the correction of the wave number that is used in the simulation corresponds to a low estimate of surface losses. Hanna et al. proposed instead to increase this correction by a factor 5 [69] : I found that this change makes the transition to be very close to $u = 0$ (data not shown).
A concurrent trend revealed by the simulation is that h decreases at the same time as β. The most likely reason for this phenomenon is that narrow bandwidths (high β values) fill the time-frequency domain with longer tails while damped signals (low β values) are localized in time, therefore sparser.

3.4 – Summary

The analysis carried out on synthetic signals gives several clues about the most important acoustic factors for the statistical structure of speech:

— For obstruents (fricatives, affricates, stops), which are similar to noise sounds: the behavior of β is expected to be related to the localization of the noise on the time or frequency axis. Stops exhibit a sharp increase in intensity at onsets (bursts), and are expected to be associated with low β values. On the contrary, fricatives present a sharp increase in power spectrum in high frequencies, and we expect high β values. Affricates borrow from both behaviors.

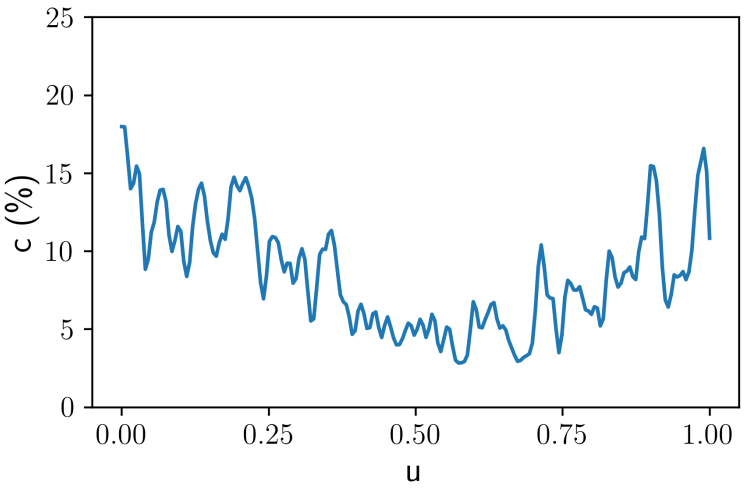


Figure 3.14 – Contrast as a function of u for synthetic vowels

— For vowels, one important acoustic factor is certainly the opening at the lips, which controls the degree of acoustic radiation. A larger opening means greater acoustic losses and wider bandwidths. As a result, we expect lower values of β for vowels with the largest aperture at the lips.
Lip opening is similar to the notion of openness (vowel height) but not equivalent. The vowel height refers to the position of the tongue and the jaw, but some open(-mid) vowels, such as [O], are pronounced with rounded lips, then with a small aperture at the lips.

The analysis on synthetic signals showed the importance of the distinction between structured and non-structured sounds. Structured sounds can be well approximated by a sparse number of atomic components (vowels, nasals, some semivowels), whereas non-structured sounds relate to noise (obstruents: most consonants). The latter are characterized by poor time and/or frequency structure. It does not mean that consonants have no structure at all on a larger time scale (e.g. stops have a clear time pattern closure - burst - opening phase). It means, however, that non-structured sounds are more easily related to noise on small times scale of about 10 ms. The distinction is relevant to our analysis because the factors determining β are different for each type. A characterization of statistical structure of speech based on features found in all samples is difficult because of this dichotomy. Structured and non-structured sounds will be described separately in next chapter for real speech data.

Chapter 4

Fine-grained statistical structure of speech

This chapter describes the statistical structure of speech based on real data. The behavior of β, which controls the frequency selectivity in the high frequency range, is analyzed for subclasses of speech sounds, each time at a finer level (broad phonetic categories, then phonemes, then phoneme parts). All classes are based on the TIMIT database [56] which provides sentences of American English along with phonetic information segment by segment.
The distribution of beta values are interpreted by highlighting the most important acoustic factors. We will be helped in this task by the analyses of the previous chapter on synthetic signals.

4.1 – Methods

The essential part of the method has been described in previous chapter. I use the same set of Gabor dictionaries to estimate the optimal value of β for many subsets of speech data. The data samples (speech slices) are decomposed in the dictionaries, associated with different values of β ranging from 0.3 to 1.1, and a score h reflecting the lack of sparsity is computed for the resulting decompositions. The dictionary that minimizes the score averaged over selected samples offers the most sparse representation and is associated with the best value $\beta^\star$ for the corresponding data. The results are then presented by mapping the values of $\beta^\star$, along with the h scores, on the (β, h) plane. The process is repeated each time at a finer level of speech (broad categories of phonemes, phonemes, then parts of phoneme). Recall that β refers to $\beta^\star$, the optimal choice of the parameter, when there is no possible confusion with the other values.

The following paragraphs present the specifics of the methods for the analysis of real data.

4.1.1 Data

The speech data was retrieved from the TIMIT database [56]. It provides audio examples of sentences in American English as well as information on their phonetic content by segment. Slices of 16 ms of speech were considered, representing 256 samples at $f_s = 16kHz$. The examples were preprocessed with filtering and then normalization. The filtering was done with a high pass Butterworth filter of order 8 and a cut-off frequency at $1.5kHz$. The use of a high cut-off frequency is not common for speech analysis as much of the phonetic information is in the low-frequency part, but the focus of this study is only on the high frequency region above the intersection point of the power laws at $f_0 = 1kHz$. In addition, the dictionaries are all the same at 1kHz, meaning that the region of discrimination is in even higher frequencies. The normalization was done by dividing each slice by its root mean square (RMS value). The TIMIT database indicates the time of releases for stops and affricates. This information was used several times in the analyses, in particular the closure part, which contains no high frequency information, was always ignored.

4.1.2 Weighting strategy

The actual cost function includes weights:

$$h_\beta(X) = \sum_i \gamma(f_i)|Y_{\beta,i}| \ . \tag{4.1}$$

where f_i is the center frequency of the i-th filter and $\gamma(f)$ typically is an increasing function of the frequency. I define three weighting strategies:

— *Strategy A* (raw scores): I make no difference between the components, setting $\gamma(f) = \gamma_0$ to a constant.

— *Strategy B* (spectral whitening): I set $\gamma(f)$ to be inversely proportional to the amplitude spectral density (for speech: +5dB/octave).

— *Strategy C*: This is a balance between the two strategies defined above. The spectral whitening is performed with a slighter gain of +2.5dB/octave.

Strategy B is mathematically justified: it corresponds to the normalization term that appears in cross-entropy terms with a Laplace prior (chapter 1). It ensures that medium frequencies do not override high frequencies due to the natural decrease in energy along the frequency axis. However, it has the opposite effect on high frequency sounds (e.g. sibilant fricatives), leading to erratic behaviors for these classes of sounds. The more naive Strategy A is also interesting in this respect because it considers the response patterns without any assumption about the global power spectrum. The results presented in the following were obtained with the intermediary strategy (*Strategy C*). The central figure on phonemes is also presented, obtained with Strategies A&B (fig. 4.4 and 4.5). The choice of the weighting strategy does not have a big impact on the main results and their interpretations, however an issue of *Strategy B* is that fricatives are associated with very large confidence intervals.

4.1.3 Bootstrapping

To provide a measure of the statistical variability inside the subclasses, bootstrap confidence intervals were computed along with each value of β.

The analyses were carried out in two steps. The first step was to compute the decompositions and the scores for each sample. At the end of this step, all the scores are saved in a table. The second step was to estimate β by minimizing the score averaged over the selected samples. Given the summary scores, the second step could be done with any recombination of the data without the need to recompute the decompositions. In particular, the computation of the bootstrap confidence intervals were simplified by this means. The use of relatively small datasets (800 samples for broad categories of phonemes and 400 samples for single phonemes) provides meaningful confidence intervals: I show later on that large confidence intervals correspond to high acoustic variability within the class.

The bootstrap confidence intervals were obtained by repeating the estimation of $\beta^\star$ 3 000 times with re-sampled versions of the (400 or 800) slices with repetitions. The bootstrapping procedure was the "smoothed percentile bootstrap". After resampling, we get a first histogram (fig. 4.1, left) that is smoothed with a Gaussian filter ($\sigma = 0.015$) to obtain the final bootstrap distribution (fig. 4.1, right). The mean value and the bootstrap confidence intervals are then extracted from the distribution (e.g: to get the 70% CI, 15% of the samples on the left and right are excluded).
The values plotted in this chapter are the means and the 70% confidence intervals of the bootstrap distributions. Alternatively, bootstrap distributions can be represented by box plots as done in fig. 4.2.

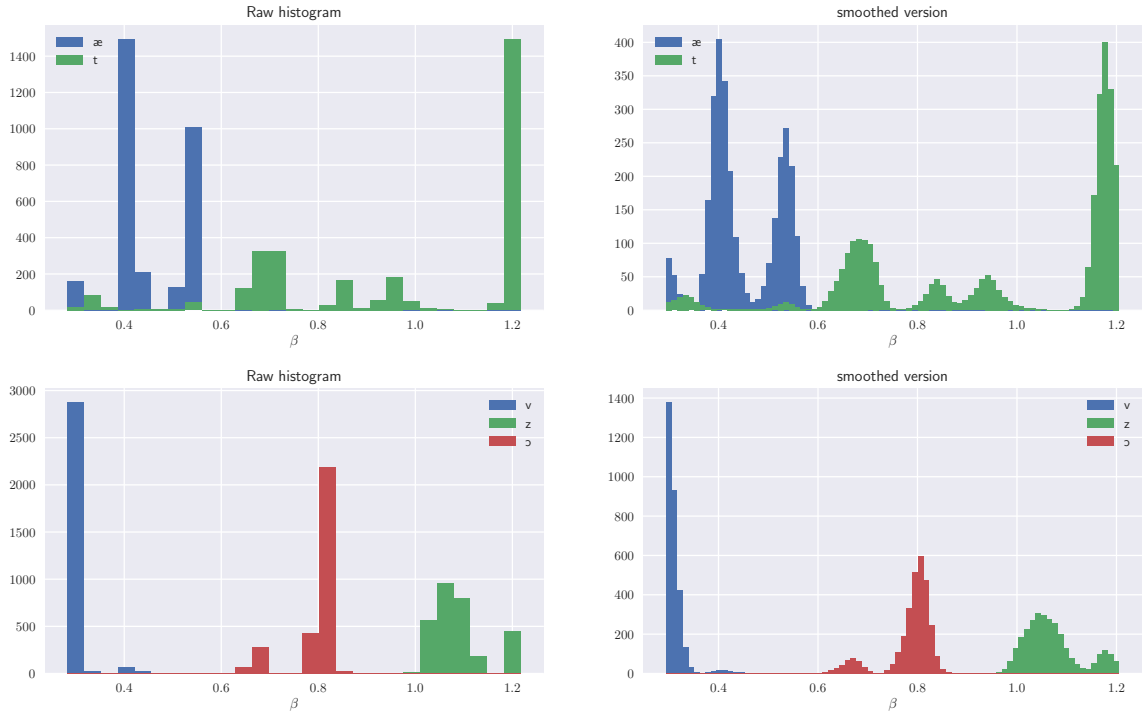


Figure 4.1 – Bootstrap distributions for a few phonemes based on 400 samples. On the right is the smoothed version of the bootstrap histogram. Phonemes: æ,t (up), v, z and ɔ (bottom). *Color online.*

Table 4.1 – **The division of phonemes used for the broad phonetic categories, similar to the division in Stilp and Lewicki, 2013.**

Category	**Phonemes**
Fricatives	s,ʃ,z,ʒ,f,θ,v,ð
Affricates	dʒ,tʃ
Stops	b,d,g,p,t,k,ɾ
Semivowels	l,r,w,j,h,ɦ
Nasals	n,m,ŋ
Vowels	i,ɪ,ɛ,eɪ,æ,ɑ,aʊ,aɪ,ʌ,ɔ,ɔɪ,oʊ,ʊ,u,ɝ,ə,ɚ

4.1.4 Analyses

Analysis 1: The purpose of the first analysis is to estimate the best value of β for different classes of speech sounds. Occurrences were retrieved from the TIMIT database for each class of speech sounds, randomly sampled from throughout the database: 400 occurrences (for single phonemes) or 800 occurrences (for broad phonetic categories: fricatives, stops, vowels...). Since the broad phonetic categories contain a greater diversity of sounds, I used 800 samples to get more robust estimations of $\beta^\star$. A 16 ms slice was selected at random for each occurrence. The scores h_β were computed and averaged over the slices, as described previously. The scores were then smoothed with a Gaussian filter ($\sigma = 0.03$) along the β axis, and the minimum score was obtained at $\beta = \beta^\star$. For the broad phonetic categories, the phonemes were divided into vowels, stops, fricatives, affricates, laterals/glides (semivowels) and nasals, as in Stilp and Lewicki 2013 (table 4.1). I do not represent the confidence intervals on h because the variations were not significant (standard deviation of order 0.01).

Analysis 2: The purpose of the second analysis is to describe the variations of $\beta^\star$ on a finer time scale. The motivation behind Analysis 2 is that some phonemes like affricates or stops are subject to acoustic changes even within an occurrence. We expect to find time patterns on $\beta^\star$ inside some phonetic units. We retrieved 400 occurrences from the TIMIT database for each phoneme, as in Analysis 1. This time, we considered eight 16ms slices at regular intervals for each occurrence, possibly with some overlap, instead of a single slice by occurrence. As the occurrences do not have the same duration, the eight steps represent relative time rather than absolute time (1 is the start of occurrence, 8 is the end). The procedure for the estimation of $\beta^\star$ was the same as described for Analysis 1. The $\beta^\star$ values computed for each step (1-8) describe the temporal evolution of $\beta^\star$.

4.2 – Results

It is recalled that in the following, β refers to $\beta^\star$, i.e. the power law satisfied by Q_{10} offering the best decomposition of the data.

Structured and non-structured sounds. The distribution of values β obtained for the broad phonetic categories (fig. 4.2) is globally consistent with the distribution of exponents found by Stilp and Lewicki (discussed further in next section). The addition of the average value of the cost function h, reflecting the lack of structure, is consistent with the separation of speech slices between structured sounds ($h < 0.7$: semivowels, vowels, nasals) and non-structured sounds ($h > 0.7$: stops, affricates, fricatives). Most sparse signals are the approximant [ɹ] ($\beta = 0.82$) and the related vowels [ɝ] ($\beta = 0.85$) and [ɜ] ($\beta = 0.88$) with $h = 0.47$ for the three phonemes, not represented in fig. 4.3. h is more than twice larger for the least sparse sounds being the fricatives [f] ($h = 1.00$) and [ʃ] ($h = 0.95$).

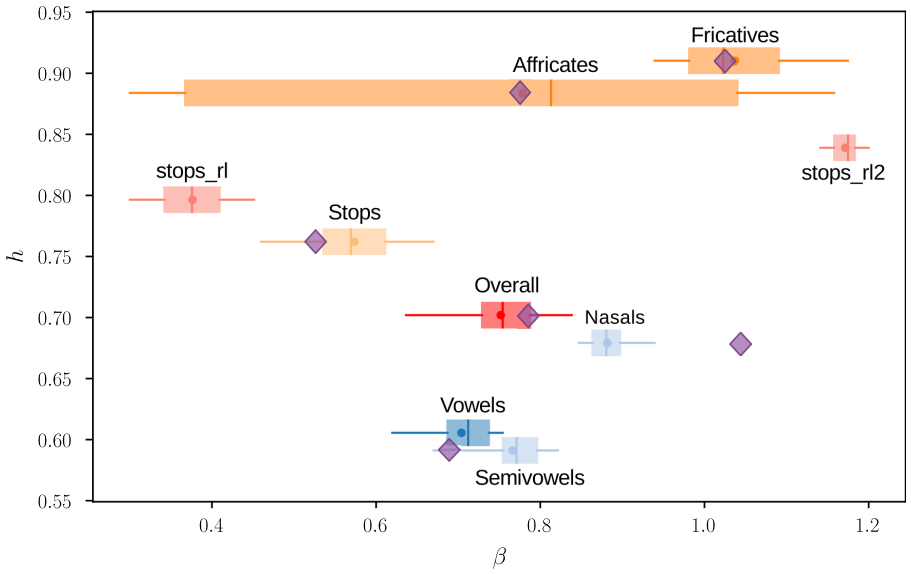


Figure 4.2 – Distribution of American English broad phonetic categories in the (β, h) plane. Box plots show quartiles Q1, Q2, Q3, [5%, 95%] percentiles (whiskers) and mean (dot) of bootstrap distributions based on 800 occurrences for each category. *stops_rl* and *stops_rl2* are for first parts and second parts of stop releases (see text). The diamonds show the values found by Stilp and Lewicki for the same categories with Independent Component Analysis [153].

Consistency of phonetic categories. The phonetic categories are unequal in terms of variability. The bootstrap confidence intervals on β reflect the diversity of acoustic features within a category. For example, affricates, which borrow acoustic features from both stops and fricatives, are characterized by large confidence intervals covering almost the entire range of values. The consistency of bootstrap confidence intervals is further confirmed when I describe more in details the statistical structure of the phonetic categories. The more detailed figure on phonemes (fig. 4.3) shows that variability can sometimes be explained by opposite values of β within a class. Most of the fricatives are in the region $\beta > 1$, but we find other fricatives ([v], [ð], [f]) in the opposite region $\beta < 0.5$. Other times, the same

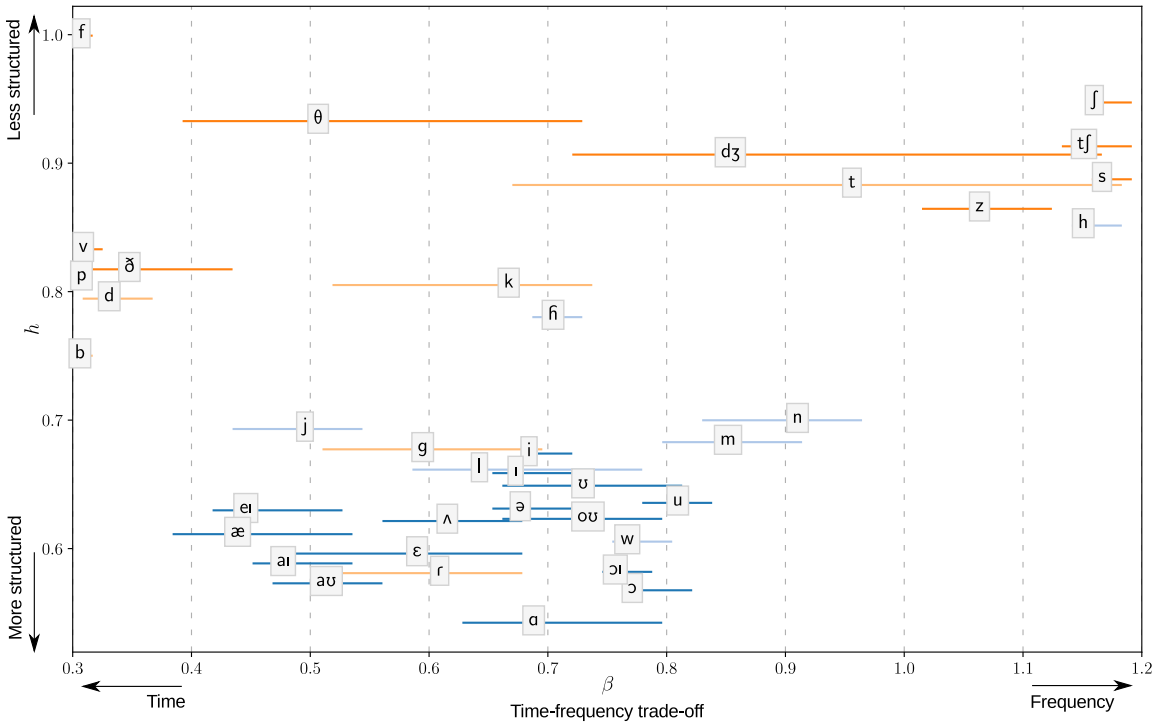


Figure 4.3 – Distribution of American English phonemes in the (β, h) plane, obtained with *Strategy C* chosen as the main strategy. Labels are positioned on bootstrap distribution means, lines represent 70% bootstrap confidence intervals. Bootstrap distributions are based on 400 occurrences for each phoneme and 3000 repetitions.

variability is again observed at the level of phonemes. For example, the affricate [dʒ], the fricative [θ] or the stops [t] and [k] still have large confidence intervals. Although there is some scattering among the fricatives and stops, phonetic categories form consistent groups. Most of the time, phonemes that are close in the acoustic space are also close in the (β, h) space. However, the phonetic categories that are used in fig. 4.2 do not always offer the best clustering of the data for statistical structure. Some phonemes appear to belong to a cluster different from their attributed category. Some examples are the aspirant [h] with the cluster of fricatives, the fricatives [v] and [ð] with stops, the stop [g] with the laterals [j] and [l], the *flap* [ɾ] with approximants.
Another measure of the significance of β is contrast, which is the relative difference between the minimum and maximum of the cost function over the values of β (eq. 3.3). $c = 1\%$ for speech as a whole when the scores are averaged over all the samples. The phonetic categories are in increasing order of c: affricates (0.4%), stops (0.7%), fricatives (1.7%), vowels and approximants (1.8%), and nasals (2.1%). Contrast again indicates a strong variability for stops and affricates which requires to be examined at a finer level. This is done in the next paragraphs.

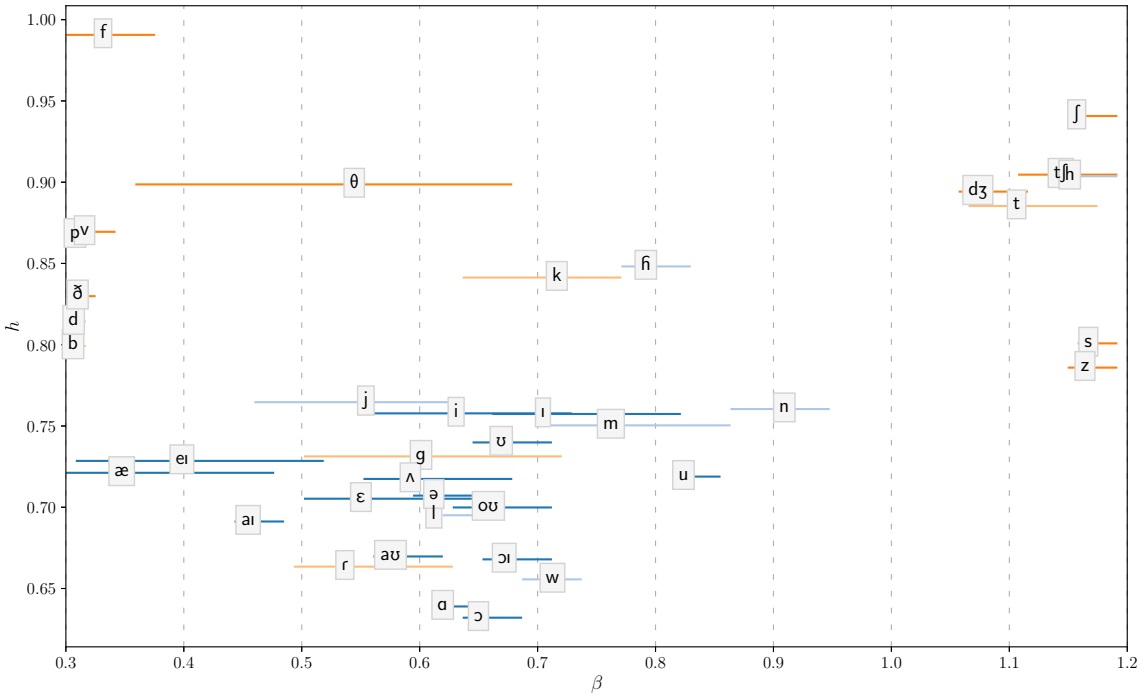


Figure 4.4 – Distribution of American English phonemes in the (β, h) plane with *Strategy A*: raw scores (no spectral whitening). Not represented: r, ɝ, ɚ ($\beta = 0.93, h = 0.58$).

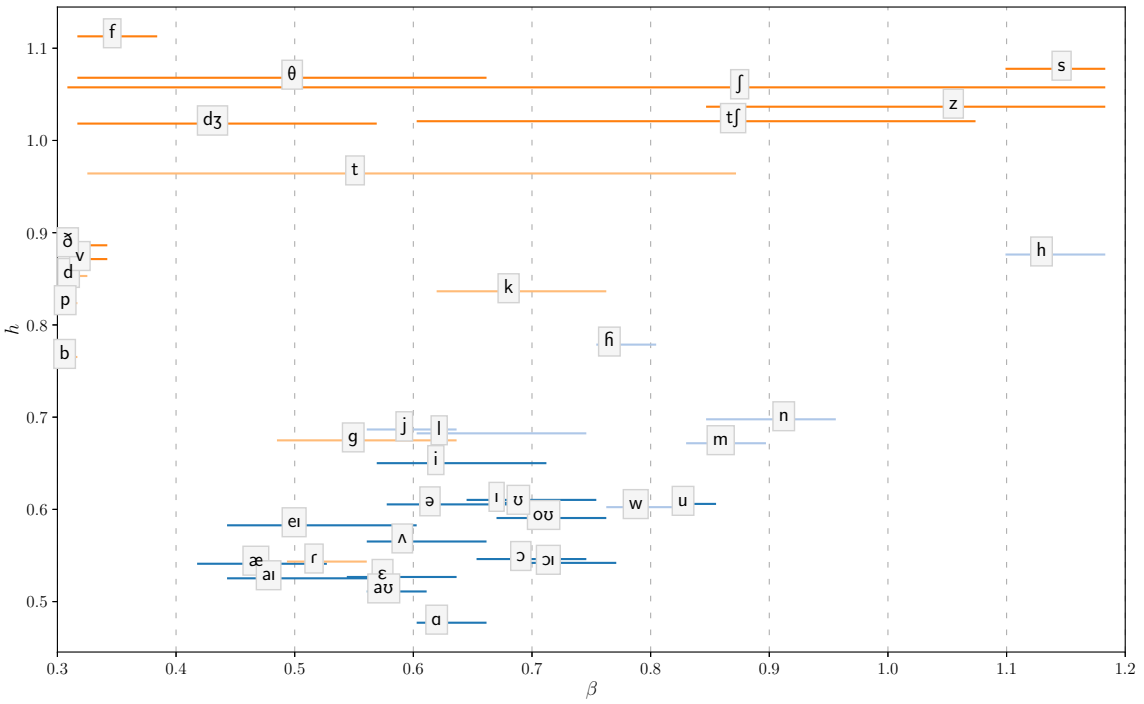


Figure 4.5 – Distribution of American English phonemes in the (β, h) plane with *Strategy B*: spectral whitening +5dB/octave on weights. This setting is unreliable for high frequency sounds (e.g. fricatives). Not represented: r ($\beta = 0.86, h = 0.37$), ɝ($\beta = 0.91, h = 0.36$), ɚ ($\beta = 0.84, h = 0.37$).

4.2.1 Stops, fricatives, and affricates

Stops, fricatives, and affricates are non-structured sounds (associated with the highest values of the cost function h). It means they are easily related to noise, for which some behaviors have been explained in previous chapter (simulation of noises windowed in time or in frequency). Their statistical structure is described first with the distribution of β values at the level of phonemes (fig. 4.3), then by characterizing the behavior of β over time within single phonemes.

Static description. For a stop, the modulation function can be thought as a gate function in the time domain with random time for the burst. This is close to simulation and time windowing when the modulated noises show a rapid increase in intensity for a short amount of time. Based on this very simplified model, we expect the time representation to be optimal. Stops are indeed associated with low β values (fig. 4.3 from Analysis 1), but the large bootstrap confidence intervals indicate a more complex behavior described further on. Fricatives are more explicit for now with Analysis 1 because they can be well approximated by stationary processes. Fricatives are the result of turbulent airflow occurring at a constriction in the vocal tract. Some noise is produced and then filtered by the vocal tract, similar to the generated sounds passing through frequency modulations. Most fricatives are characterized by values of β close to 1 consistent with this frequency description (fig. 4.3). It is at least true for the sibilant fricatives. Sibilant fricatives are filtered with a short cavity after the alveolar ridge and therefore present sharp increase/decrease of power in the high frequency range. It is a clear trend for the hissing alveolar fricatives [s] (c=4%) and [z] (c=3%) and it remains valid for the hushing post-alveolar fricative [ʃ] (c=1%). However, labial or dental fricatives [f], [θ], which are less affected by vocal filtering resulting in wide-band noise, are associated with lower β values (c=1%, c=0.6%, resp.) and maximal h values. Voiced fricatives are an intermediary case for what has been seen as they are affected by both time and frequency modulations. In addition to vocal filtering, the sound intensity follows the repeated openings and closures of the glottis. The coincidence of time and frequency modulations is likely to explain why the points move to bottom-left on the (β, h) plane when replacing the unvoiced versions of the fricatives by the voiced ones (compare [s], [h], [θ], [f] with [z], [ɦ], [ð], [v], resp.). The shift is also visible on stops to some extent (compare [t], [k], [p] with [d], [g], [b], resp.).

Dynamic description. The large confidence intervals on stops and affricates indicate an inner variability that requires to be handled by a finer description. The dynamic aspects of stops and affricates have to be taken into account. The closure parts were ignored in the analyses, but the release parts can be separated into two phases (chap. 3): the burst following the closure, then the opening phase, which is similar to a fricative sound. I conducted a specific analysis to determine if the biphasic nature of stops and affricates has an impact on the parameter β. I performed the procedure described Analysis 1, but this time I separated the releases into two parts of equal duration. The results, reported in Figure 4.6, show that the dual nature of stops and affricates is also revealed by β. The first and second parts of the releases are indexed by the suffixes $\_rl$ and $\_rl2$, respectively. While the $\_rl$ parts containing the bursts are characterized by minimal values of β, the $\_rl2$ parts are characterized by higher values close to 1. Values for stops as a category are also reported in fig. 4.2 with the same suffixes. This analysis shows that the opening phase of stops is in reality similar to fricatives with regard to statistical structure.

We can describe further the temporal behavior of β for stops and affricates with Analysis 2. Figure 4.7 provides the time evolution of β within some phonemes. The parameter β is stable for vowels or nasals – apart from diphthongs. But it increases during the occurrences of stops and affricates, joining the extreme values. This behavior is consistent with the description in the previous paragraph. However, this transition is more or less abrupt depending on the nature of the opening phase. The stop [t], whose opening phase is similar to the sibilant fricative [s], has a fast transition after the burst. In contrast, the stop [p] has a more gradual transition. We explain this gradual transition by the fact that the opening phase of the stop [p] is similar to the low β fricative [f] on which some formant structure appears gradually when the back cavity plays a role again (as for the high β fricative [h]). Note that although the change is less pronounced, fricatives have also an upward shift of β at their onset.

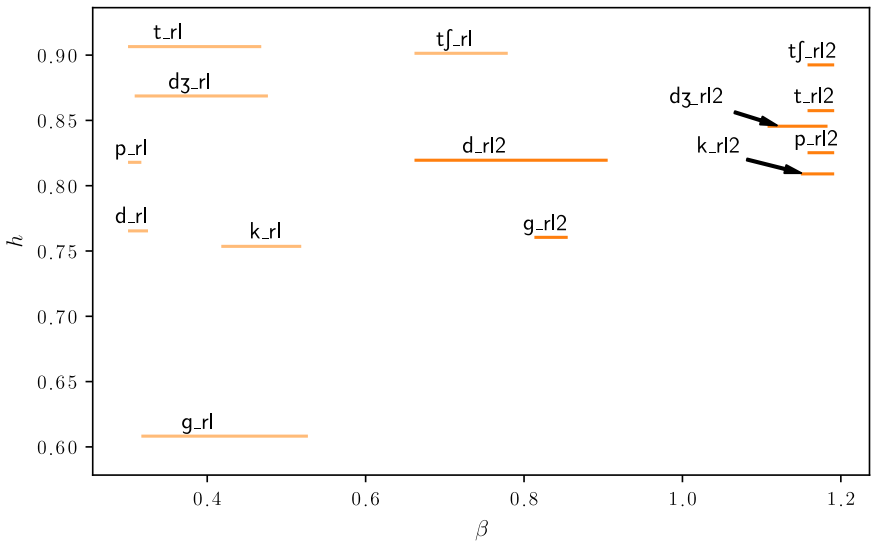


Figure 4.6 – Detailed distribution of stops and affricates in the (β, h) plane. When stop or affricate releases are separated into two parts of same duration, first parts *_rl* (including the bursts) are best represented in a dictionary with a low β value (time representation) but second parts *_rl2* are best represented in a dictionary with a high β value (frequency representation). Plot shows distribution means and 70% bootstrap confidence intervals.

4.2.2 Vowels, semivowels, and nasals

Vowels. The above reasoning for non-structured sounds does not apply to structured sounds, in particular vowels. We have to look for the acoustic properties that determine the structure of the signal for these phonemes. The structure of vowels can be seen both in time and in frequency. Along the frequency axis, vowels are characterized by a few spectral peaks arranged at almost regular intervals (~1 kHz): these are the formants and correspond to the resonances of the vocal tract (chap. 3). Note that the harmonics of *F0* are not resolved in the high frequencies, on the time scale we consider, and therefore have no effect on frequency structure. On the time axis, the signal presents peaks of intensity at the instant of glottal closure remaining true if the signal is band-passed around

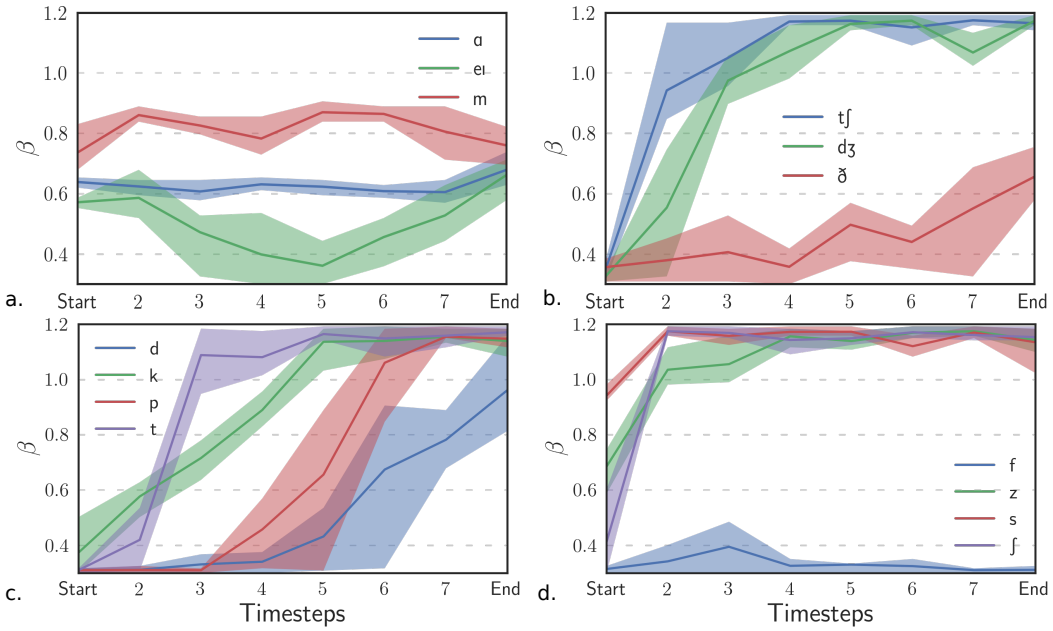


Figure 4.7 – Temporal evolution of β for some phonemes. The timesteps represent the relative times between the beginning (*Start*) and end (*End*) of the occurrences at regular intervals. The region filled represents 70% confidence intervals. Stop onsets are associated with minimal values of β (time representation), but the opening phases are associated with higher values. **a.** Vowel: ɑ, diphtong: eɪ, nasal: m **b.** Affricates: tʃ, dʒ, fricative: ð **c.** Stops: d, k, p, t **d.** Fricatives: f, z, s, ʃ. *Color online.*

formants. The latter statement stands at least for the first formants as higher formants can be excited at other instants, especially at the glottal opening [70]. On one glottal cycle, a naive image of the underlying structure in the time-frequency plane can be a comb shape whose *teeth* represent the formants. The complete structure cannot be perfectly covered by a single Gabor filter bank: a representation is always a compromise between time and frequency favoring either the glottal pulse or the tails of the formant oscillations. From a frequency point of view, it means that either the wideband parts of the signal associated with low response and low group delay or the narrow bands associated with the formants, are somewhat neglected. This competition between the pulse and the formant oscillations has a visible effect on the quality factor of the efficient coding filters at medium frequencies, at 1 kHz and up to 5 kHz. When ICA is performed on the front vowels which have high second formants, the quality factor goes from 1.3 at 1 kHz to 3 at 1.7 kHz. On back vowels, the quality factor increases at 1 kHz more steeply (see fig. 1 of ref. 153).
The optimal representation, in any case, is expected to be related to formant bandwidths. The simulation on synthetic sounds at the output of a uniform cylindrical waveguide with different radii (chap. 3) suggested that a key factor was the degree of acoustic radiation at the lips and lip opening. It showed that the sounds with the largest aperture present the largest bandwidths are associated with the lowest values of β. The same trend can be observed for real data although the transition is more gentle and less apparent. The vowels form a tight cluster around $\beta = 0.6 \pm 0.2$, but the unrounded vowels and diphthongs [æ], [ɛ], [eɪ], [aʊ], and [aɪ] give lower values than the rounded vowels [u], [ʊ], and [ɔ] (fig. 4.3). The intermediate sounds [ə], [ʌ], [i], and [ɪ] are in the middle. The vowel [ɑ], however, does not match the rest of the distribution (this may be the consequence of the constriction

at the back of the vocal tract weakening the effect of acoustic radiation at the lips). The simulation revealed the concurrent trend that h decreases at the same time as β. However, the trend is not sufficiently clear on phonemes to conclude that this rule applies to actual data.

Nasals and semivowels. The nasals are found in the continuity of the vowels, with higher values of β and h, meaning that nasals are better described in frequency. I explain this fact by the presence of antiresonances surrounding the formants which have the effect of cutting the bandwidths of the nasals. This is rather contrary to the known fact that nasals have wider bandwidths due to greater surfaces losses. This is not contradictory since the region of interest is in the high frequency range and the values of wide bandwidths (e.g. 10dB bandwidths) are more significant here than the usual narrower 3dB bandwidths.
Semivowels are within the same range of values of β and h as vowels. The rhotic approximant and r-colored vowels occupy the lower right part of the cluster ($\beta = 0.8, h = 0.47$) in the (β, h) plane. A likely explanation for the low score h for r-sounds is that they present a strong frequency decrease in high frequencies, hence the underlying structure for the high-passed filtered signal is essentially a prominent peak in frequency close to 1 kHz. I found that all the vowels that are related to the [ɹ] sound tend to be close to the location of the [ɹ] sound on the (β, h) plane. This effect could also apply to the back vowels [ɑ] and [ɔ] and offset the effects demonstrated by the simulation for low back vowels.

4.3 – Interpretation

4.3.1 Consistency with previous work

The parameter β corresponds to the slope of the quality factor Q_{10} against center frequency f_c, on a logarithmic scale, for the best decomposition of the data. The overall distribution of β values for the broad phonetic categories (fig. 4.2) is in agreement with the regression slopes found by Stilp and Lewicki using Independent Component Analysis. In particular, the slope β is found between 0.7 and 0.8 for speech data as a whole, with both ICA and our method. Stilp and Lewicki used several subcategories for vowels but always found a β value in the range 0.7-0.85 (in current analysis: 0.7). The most noticeable gap is for nasals (0.9 compared to 1.05 in Stilp and Lewicki [153]). Phonemes that are close acoustically are found together in the (β, h) plane, showing the robustness of β in relation to acoustic features. The detailed distribution at the level of phonemes (fig. 4.3) shows the consistency of the phonetic categories as defined in table 4.1. We have seen that these categories can still be adjusted, for example the aspirant [h] belongs to the cluster of fricatives rather to the cluster of semivowels.

4.3.2 Relationships between the parameter β and acoustic features

In previous chapter and in previous paragraphs, I inferred the main acoustic factors that affect β based on the distribution of β values at or below the level of phonemes and based on the simulations. Some of these properties coincide with previous proposals, but others are new or clarify some previous ideas.

In 2002, Lewicki examined whether the spectral tilt – the natural decrease of power spectrum density – could explain the power law satisfied by the quality factor [94]. His conclusion was that there is no connection between the two. The average power spectrum density has a low impact on signal structure, because the efficient coding filters are localized in frequency – it has an effect on the weighting between midrange and high frequencies but not on the atomic components. An exception is that the addition of a decrease or increase in the frequency power spectrum leads to the emergence of frequency structure in the case of non-structured sounds. We have seen this phenomenon on fricatives: high-pass filtered hissing sounds [s] and [z] are associated with a higher value of β and a lower value of h compared to broadband noise or the [f] sound. The symmetric case for time structure is that stop onsets are associated with a low value of β due to the sudden increase in intensity.

In 2013, Stilp and Lewicki listed three others acoustic factors affecting the value of β: *harmonicity*, *acoustic transience* and *bandwidths* [153].

I argue that the *F0* periodicity plays little if no role because the efficient coding filters are shorter than the period length. *Harmonicity* in the usual sense (*F0* harmonics) does not add any frequency structure on this time scale (unlike formant structure). More generally, I argue that acoustic changes of characteristic time greater than the duration of a glottal cycle (e.g. coarticulation, formant transitions) do not have a significant impact on the efficient coding filters as such. However, I have found that an acoustic factor close to harmonicity and significant for the statistical structure of speech is *voicing*. The fact that voiced sounds are characterized by scarce time-localized excitations has the effect of enhancing time localization, and decreasing both β and h. Consequently, vowels have been shown to be associated with relatively low values of β, a result that could appear counterintuitive. Vowels are sustained sounds that are often believed to be better captured by a frequency representation. This view might be biased by the source-filter model that focuses on the resonances in the frequency space and makes extensive use of Fourier analysis. The statistical structure of speech supports the opposite view that a time decomposition, i.e. characterized by a low quality factor, would be more appropriate for the efficient coding of vowels.

I agree with Stilp and Lewicki that *transiency* is a key acoustic factor for the statistical structure of speech, since the lowest values of β are reached during stop bursts because of the sudden increase in intensity. The current description of the statistical structure of speech also supports the hypothesis that β is related to *formant bandwidths* for vowels and nasals. It suggests that two key acoustic factors are vowel openess (but more specifically lip opening) and the existence of antiresonances. The value of β is increased by greater opening but increased by antiresonances, however these two parameters alone do not explain the entire distribution of values for vowels and nasals.

The fact that β is bound to a few acoustic properties means that the analysis could be replicated on natural sounds other than speech. The reasoning on non-structured sounds would probably apply to many environmental sounds, and the reasoning on vowels would

apply to other animal vocalizations as well.

Chapter 5
Nonlinear sparse representations of speech

After having described the fine-grained statistical structure of speech, we can ask ourselves if it can motivate nonlinear representations of speech sounds as a basis for efficient coding schemes. This question is also motivated by the fact that the cochlear filtering has a nonlinear behavior, since cochlear frequency selectivity decreases with sound intensity at medium and high intensity levels. In this chapter, I show that the statistical structure of speech can be matched by a level-dependent filtering scheme. A decrease in frequency selectivity is shown to be consistent with the description provided in the previous chapters, supporting the hypothesis that nonlinear cochlear filtering is adapted to the statistics of speech. I discuss the limitations of the current method that impede a thorough verification of the hypothesis. These preliminary results call for further research on the relationship between the statistical structure of speech and cochlear tuning, both at a theoretical and experimental level. I also present more general limitations of the parametric approach and the use of Gabor dictionaries, for a comparison with peripheral auditory coding.

5.1 – Overview

Motivation. The description of the fine-grained statistical structure of speech showed that the best representation has to be adjusted to match subclasses of speech, with different time-frequency resolution trade-offs. For instance, sibilant fricatives (e.g. [s]) are best captured by a frequency decomposition (quality factor linear with respect to frequency). On the other hand, transients parts, especially stop bursts, are best captured by a decomposition with a good time resolution (quasi-constant quality factor). A natural idea to better adapt the representation to the statistics of speech is then to adjust the representation and the frequency selectivity to the input signal.
So far, I have only discussed the case where the representation is obtained by a linear transformation of the input. The output vector Y was the product of a constant matrix W and the input vector X (fig. 1.1, chap. 1):

$$Y = W^T X \ .$$

A representation that changes with the input means that the linear transformation is replaced by a more general transformation g:

$$Y = g(X) \ . \tag{5.1}$$

The linear case is included in this setting, but it offers much more possibilities with $g(X)$ being able to change nonlinearly with some characteristics of the input.

Achieving adaptive time-frequency representations is a natural motivation for many classes of signals whose characteristics of generation change over time. This type of nonlinear decompositions has been called *nonstationary decompositions* to emphasize the dependence of the variables on time:

$$Y(t) = g(t)(X(t)) \ .$$

One example is the use of nonstationary Gabor frames for the decomposition of sounds generated by isolated music instruments (e.g. glockenspiel) for which the percussive part requires better time resolution that the sustained tails [13, 14]. In this example, time adaptivity (changes of time resolution along time) and frequency adaptivity (changes of frequency resolution along frequency) were considered separately as symmetric cases. Both were considered as a way to provide *flexible* time-frequency resolutions adapted to specific signals.

Although the term of *adaptive representation* is convenient to characterize this nonlinear framework, I try to use it sparingly in the following because it could be misleading. In information theory, *adaptive coding* means that the coding scheme can change over time as the systems gets more information about the input, and the term is particularly suitable when dealing with streaming data. While this problem is not totally disconnected with our subject, a nonlinear representation that is adapted to several types of inputs does not necessarily mean that the coding scheme has to evolve over time. The distinction may seem thin, but makes sense when thinking about the auditory system. The auditory system exhibit temporal changes in behavior, on several time scales, that are akin to adaptation to stimuli [125]. This applies in particular to the early stages of sensory processing. It does not mean, however, that if the type of acoustic input is changed and exposure to speech ceases, all the characteristics of the peripheral auditory system will change to adapt to the new type of input. An *adapted* system is not the same as an *adaptive* system. The question I address in this chapter is whether the peripheral auditory system, whose coding properties are considered fixed, is adapted to the statistical structure of speech. How the system has adapted to its environment through evolution is also a worthy question but out of the scope of this thesis.

Requirements of a nonlinear representation. The nonlinear setting allows a high degree of freedom and flexibility to adapt the representation to speech sounds at a fine level. However, this additional adaptability can in turn add complexity to the coding scheme. There is here an exchange between the efficiency of the representation and computational efficiency. One naive solution based on the study of statistical structure of speech at the phonetic level would be to define the representation as a function of the phonetic category to which a sound belongs. The function g would be linear when restricted to a phonetic category, and we can return to the linear case:

$$Y = W(P(X))^T X$$

where the matrix W depends however on the phonetic category $P(X)$. This scheme would provide an appropriate representation for every speech sound, but now a prior analysis of the signal is required to know what is the phonetic category. It seems to be a vicious circle: the gain in representation is obtained in exchange for an additional cost to analyze the signal, and the time the phoneme has been recognized, it can already be too late to adjust the signal decomposition. There are other difficulties for the nonlinear setting: for example, a representation that is defined at a too fine level would not be robust to small changes in the structure of the input. We want the nonlinear scheme to meet a few requirements:

— *Representational efficiency:* the nonlinear representation has to capture the overall statistical structure described in the previous chapters. The finer the level of adaptation, the better.
— *Simplicity*: we want the representation to depend on simple characteristics of the signal:

$$Y = W(Q(X))^T X \tag{5.2}$$

where $Q(X)$ is obtained from X with a minimal computational cost.
— *Power of generalization:* since the acoustic stimuli can vary in many ways, the representation has to be broad enough. From a 'learning' point of view, this means that the representation must be able to generalize to other sounds than the ones the system has be trained on (it should avoid *overfitting*).
— *Regularity:* We want the functions $W(q)$ and $Q(X)$ to be regular enough. The constraint of regularity is motivated by these two requirements:
  — *Consistency:* the representation has to be suitable for higher level processing tasks, meaning that the time-frequency features must not change radically from one time to another.
  — *'Mechanical' cost:* thinking about the cochlea, a change in the representation means that the mechanical properties of the basilar membrane must change (see next section). A too irregular behavior would not be suitable for the mechanical system. The mechanical constraints also impose regularity on the atomic components of W (the filter shapes must be regular as a function of center frequency).

Following these requirements, it appears that the simplified representation model used in previous chapters provides an appropriate framework to investigate possible strategies, or at least to understand the overall behavior of plausible efficient coding schemes. The power-law model for the Q-factor, summarized by the β parameter, is a good compromise between the ability to adjust to speech on very short time scales and the quality of being general enough, since it fits well both the overall statistical structure of speech and data on cochlear tuning. The questions remains what must the control parameter(s) $q = Q(X)$ should be. An adaptive code based on phonemes would not be feasible in practice because it would take too long to recognize a phoneme before the representation could be adjusted. In this chapter, I will focus on one control parameter in particular being sound intensity level. Including the intensity level as a control parameter provides a simple strategy to adapt dynamically the representation to the input, in addition reflecting the nonlinear behavior of the cochlea, as described in next section. Sound intensity level could serve as a basis for efficient coding schemes as it is an indicator that can be captured almost effortlessly by the auditory system, or other speech processing systems. The choice of this framework is summed up by the equation:

$$Y = W^T_{\beta(I(X))} X \tag{5.3}$$

where $\{W_\beta\}$ is the family of Gabor dictionaries indexed by β, and β is a function of the intensity level $I(X)$.

Link with auditory coding. Most works on the efficient coding of speech – and other natural sounds – and how it relates to auditory coding, has only focused on the linear setting. In 2002, Lewicki showed that Independent Component Analysis applied to speech yields a bank of filters characterized by the same power law for the quality factor as

the cochlea [94]. He proposed the hypothesis that speech evolved to be optimally coded by the mammalian auditory system. He also suggested that the median β value was explained by the diversity of acoustic factors in speech production providing the right balance between transient and sustained sounds. As a piece of evidence for this explanation, he showed that the same agreement with physiological data was obtained with a mixture of environmental sounds and animal vocalizations [94]. However, the scattering of β when ICA is performed on subclasses of speech could imply a more efficient coding scheme based on a nonlinear representation. Stilp and Lewicki suggested that the distribution of the β values is congruent with the diversity of time-frequency trade-offs found in the characteristic responses of the neurons of the cochlear nucleus [153]. However, they admitted that this property of the cochlear nucleus stands for single tones but maybe not for complex stimuli. It can be added that recombining filters is a computationally intensive task that can not be integrated in an efficient coding scheme easily. Instead, I argue that if an efficient coding strategy is implemented physiologically to adjust the neural representation to the input, it should be at the level of the peripheral auditory system. The assumption that the auditory filters are fixed and independent of the input is only an approximation of cochlear tuning. The active mechanism of the cochlea makes auditory filtering highly nonlinear. This behavior is complicated to describe exhaustively as it implies to have a good model of the traveling wave along the basilar membrane of the cochlea [100, 101]. However, the main behavior of this nonlinearity can be characterized as a level-dependence of the cochlear decomposition. This is the focus of the next section.

5.2 – Level-dependent filters

5.2.1 Level-dependent auditory filters

Physiological mechanisms at the origin of compressive nonlinearities. Nonlinear cochlear tuning is a product of the active mechanism of the cochlea, also known as the cochlear amplifier. The cochlear amplifier was predicted by T. Gold in 1948, who noted that the degree of damping of a passive model of cochlear filtering was not compatible with the frequency selectivity of the inner ear [63]. His argument was that acute frequency selectivity was only possible on the condition that additional mechanical energy was supplied to the cochlea. Although the high frequency tuning of auditory nerve fibers has long been known, it was not until the late 1970s that the idea that it was also a mechanical feature of the cochlea became widely accepted. In 1971, W. Rhode used an optical method, the Mössbauer technique, to show that the frequency selectivity of the basilar membrane was much higher than previously thought (von Bekezy's model of the passive cochlea) [132]. A few years later, T.D. Kemp discovered that the ear spontaneously emits sounds [86]. These emissions are called otoacoustic emissions and are direct evidence that the cochlea possesses an active mechanism.

The main components of the cochlear amplifier are the outer hair cells (OHC). They provide mechanical energy to the membranes of the organ of Corti (fig. 5.1). This action has two concurrent effects: a) amplify the vibrations of sounds along the cochlea, b) enhance the frequency selectivity of cochlear filters. However, the OHC are prone to a nonlinearity: when the displacements of the basilar membrane exceed a threshold, their activity saturates and the cochlear amplifier becomes weaker. The effects of this behavior are compressive

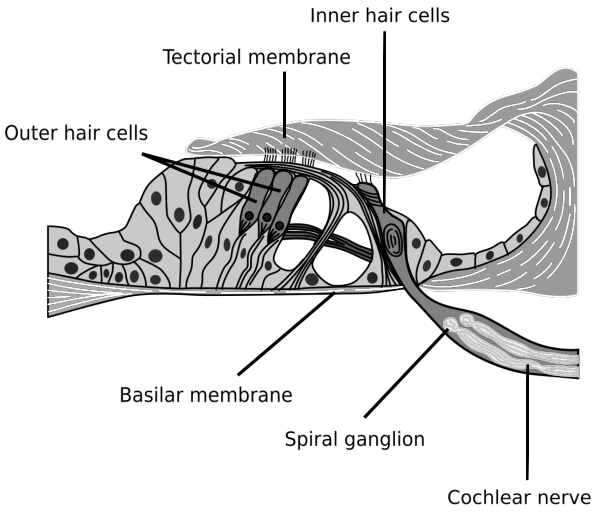


Figure 5.1 – Diagram of the organ of Corti (cross section of the cochlea, part between the tympanic and vestibular ducts). The organ of Corti is the receptor organ for hearing, and the place where the mechanical energy is transduced with the emission of action potentials caused by the vibrations of the hair cells. The hair cells are of two kinds: inner hair cells (IHC), which constitute most of the afferent connexions, and outer hair cells (OHC), responsible for the active mechanism. *Adapted from Wikimedia Commons.*

nonlinearities on the amplitude and the frequency selectivity of the response: a) the main consequence is that the active mechanism amplifies low-level sounds but compresses higher level sounds. As a result, the dynamic range of the mechanical vibrations of the basilar membrane is reduced compared to the dynamic range of acoustic waves. b) frequency selectivity is enhanced for low-level sounds, but decreases with the amplitude of the mechanicals vibrations above a threshold.

The compressive nonlinearities on the frequency selectivity are what we are interested in this chapter. They are mainly characterized by a decrease in the quality factor with sound intensity level. Interestingly, the strength of this nonlinearity increases with frequency [171, 124, 166] consistent with the initial assumption that the variations of the quality factor are small at 1 kHz but large at 8 kHz.

Models of nonlinear cochlear filtering. Several models for mammalian or human auditory filtering include the effects of compressive nonlinearities [141]. One example is the dynamic compressive gammachirp [80, 81], an extension of the linear model of the gammatone. As mentioned before, level-dependence is only an approximation of a complicated behavior resulting from the interaction between the traveling wave and the nonlinear behavior of the outer cells positioned along the cochlea. Exhaustive models of auditory filters take into account this coupling, either with the use of (at least) two parallel filterbanks (e.g. the DRNL model [100], or the phenomenological model of Carney et al. [171], recently augmented and refined [174, 173]), either with a model of cascade filters/transmission line [101]. These models are fitted on psychophysical or neural data, or even directly on responses of the basilar membrane, but the dependence on level is often based on scarce data at single or a few frequencies since this dependence is difficult to assess (see next paragraph). For the most exhaustive models, the compressive nonlinearities are more detailed than a simple dependence on intensity level, and therefore can account for other

psychophysical phenomena. These phenomena include in particular two-tone suppression (the broadening of filters is more prominent when the input is made up of two close frequency components) and the fact that compressive effects occur at a lower intensity threshold when the input has significant low frequency energy.

The experimental difficulty of assessing level-dependent cochlear frequency selectivity. Cochlear frequency selectivity can be assessed by different means. It can be estimated directly from the tuning curves of auditory nerve fibers [133, 111], or from the shape of filters obtained with reverse-correlation techniques [31]. Another way is to use responses of the basilar membrane to tones [139], but this method is limited for low and medium frequencies (<8 kHz). All the techniques mentioned above have been used in mammals (e.g. cats), however they are not applicable to humans as they are invasive. In humans, estimates of cochlear tuning are primarily based on psychophysical experiments that have been ingenuously set up for this purpose (e.g. the notched-noise masking method [123]). Some researchers have tried to get indirect but more objective estimates of human cochlear tuning based on non-invasive measurements, including otoacoustic emissions [144], and electrocochleography (ECochG) [165].

In addition to the difficulty of finding experimental procedures suitable for humans, there are other complications for the assessment of level-dependent frequency selectivity. The first obstacle is that these effects are relatively small, especially at mid-frequencies. The second obstacle is that most methods are confounded by the nonlinearities of the cochlea, as they require to test different levels for the stimuli. Because the input must be of sufficient intensity, frequency selectivity in response to low-level sounds is the most difficult to estimate, in particular in humans. A workaround is to exploit a well-known temporal phenomenon of the auditory system: forward masking (the fact that exposure to a pre-stimulus reduces the response of the auditory system) [123, 166]. Another confusing element is that the frequency selectivity is evaluated with the quality factor Q_{10}, but more often with another quality factor called the Q_{ERB} (ERB = equivalent rectangular bandwidth). The latter requires to fit an auditory model (e.g. roex filter) on psychophysical or physiological data. The behaviors of Q_{ERB} and Q_{10} with frequency are found different, with Q_{ERB} typically being almost constant [166]. If we consider Q_{ERB} instead of Q_{10}, β would take the value of 0.3 [123]. The illustration of the relative confusion surrounding cochlear tuning is that the debate on the specificity of human auditory tuning, compared to other mammals, is still not settled [107]. The prevailing view is that humans are not very different from unspecialized mammals regarding auditory tuning, but would benefit from an enhanced frequency selectivity, particularly in response to low level sounds.

5.2.2 Agreement with the statistical structure of speech

Following the behavior described in the previous paragraph, the proposition that nonlinear peripheral auditory processing matches the fine-grained structure of speech would make sense if β was negatively correlated to sound intensity. Using the same method as previous chapters, I found that this is indeed the case. This trend is reinforced when the first parts of stops and affricates containing the bursts are ignored (fig. 5.2). The thorough testing of this proposition is not the purpose of this thesis, but the fine-grained statistical structure of speech as described in the last chapters explains why the agreement with cochlear nonlinearities is plausible. The left part of fig. 5.2 contains the low-level sounds of speech, which are mainly non-structured sounds (stops, affricates, and fricatives). We

have seen that, at least if we ignore the onset parts, non-structured sounds are better decomposed with a high value of β close to 1. This explains the higher value of β observed in the left part of fig. 5.2. The right part of fig. 5.2 contains the high-level sounds of speech, which are mainly vowels. For vowels, formant bandwidths and sound intensity level increase at the same time with lip opening, explaining the decrease in β observed for the highest intensity levels. This is congruent with cochlear compression, which expands the bandwidths of auditory filters as sound intensity increases. Figure 5.2 was reproduced with the synthetic vowels generated by a cylindrical waveguide (second simulation in chap. 3). We find the same decrease in β with sound intensity level, but the transition is more abrupt (fig. 5.3). The general trend described above is also visible at the level of phonemes (fig. 5.4). Contrast for the results in fig. 5.2 is under 1% for the left part and above for the right part (fig. 5.5).

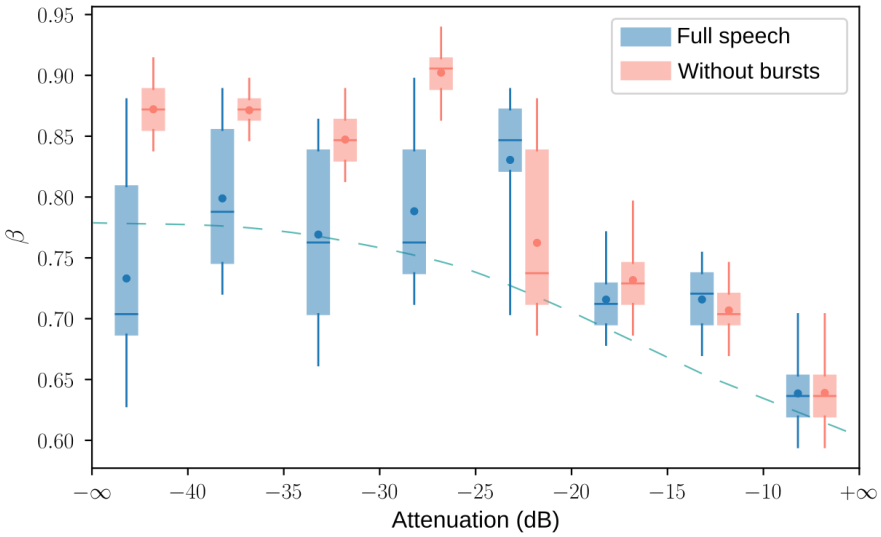


Figure 5.2 – Exponent β associated with the most sparse decomposition of speech sounds of same intensity in 5dB intervals (ref:max). **In blue:** Full speech, **in red:** same but with the first parts of stop and affricate releases removed. Box plots show quartiles, [5%, 95%] percentiles (whiskers) and mean (dot) of bootstrap distributions obtained from 2 500 16ms-slices of speech. Dashed line: indicative value of β estimated from electrophysiological measurements in cats (based on Verschooten et al., 2012 [166], range 25-65dB).

Figure 5.2 on the theoretical behavior of β predicted by speech statistics also shows the trend for physiological measures of cochlear tuning (dashed line) for reference. However, the reader should be aware that these are only indicative values. The dash line joins the regression slopes obtained from the figure of Q_{10} as a function of center frequency and intensity (in the range 25-65dB) in Verschooten et al., 2012 [166]. However, this data is too scarce to have good confidence on the regression slopes. These values are based on electrophysiological measurements in cats (compound action potentials, CAP) in a forward-masking setting.

Although the best match with actual data is the blue box plots (analysis based on the whole speech data), the pattern obtained with the red box plots (analysis based on first parts of stop releases removed) is also interesting. Because the onsets of stops and affricates are more efficiently decomposed with a low value of β, the pattern before the knee is a plateau with a high value of β close to 1 (good frequency selectivity). The

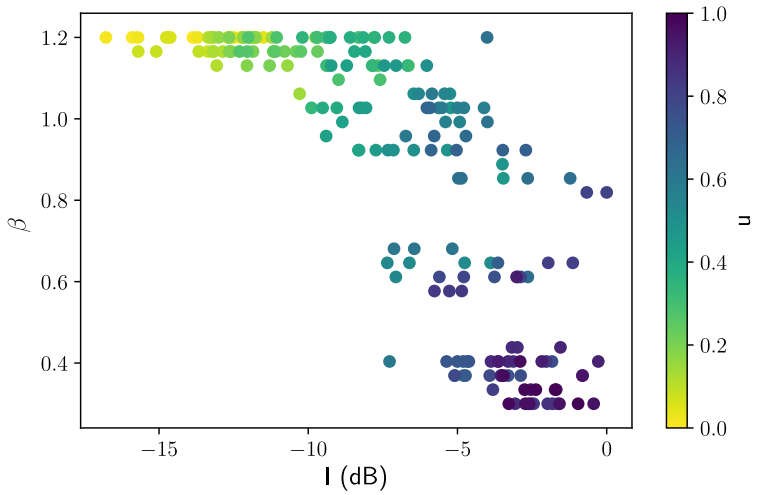


Figure 5.3 – Scatter plot of simulated samples on the (I, β) plane: exponent β against intensity I in dB (ref:max). Each point is a sample of the second simulation (chap. 3) on synthesized vowels. The control parameter u determines the aperture of the cylindrical waveguide from $r = 0.2cm$ $(u = 0)$ to $r = 1.3cm$ $(u = 1)$.

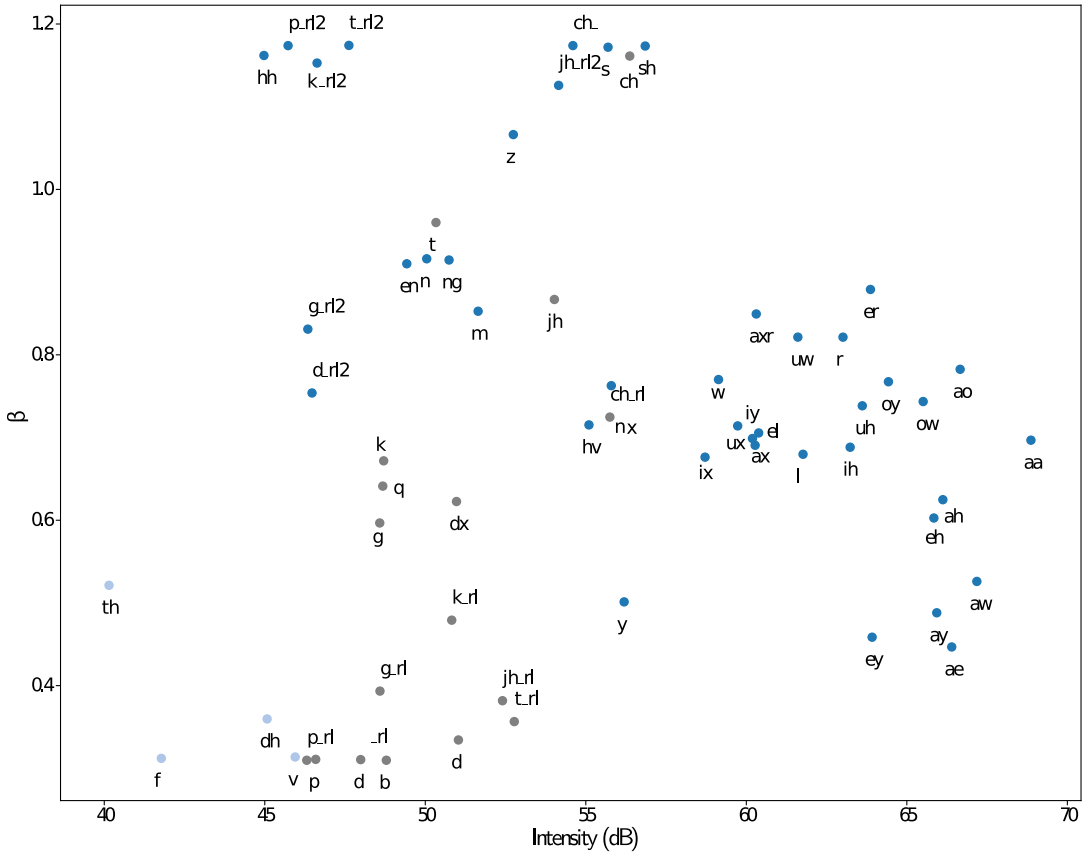


Figure 5.4 – Distribution of phonemes (American English) in the (I, β) plane. The negative correlation between β and intensity (in dB) is a general trend among the phonemes (ARPABET notation). The second parts of the releases of stops and affricates have been considered instead of the entire occurrences (gray dots). Outliers: broadband noise fricatives (in light blue). Color online.

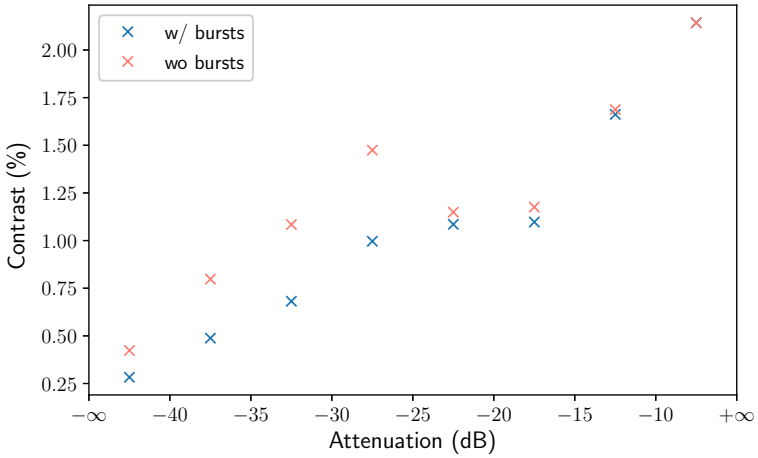


Figure 5.5 – Contrast as a function of intensity (corresponding to fig. 5.2), with (blue) and without (red) first parts of releases for stops and affricates.

frequency selectivity declines after a threshold. The same kind of pattern is seen with cochlear compression, but it is unclear whether separating the onsets of sounds would make sense in comparison with auditory coding. It would require further investigation to know if the specificity of onsets reflects the temporal processing of the inner ear, but it has already been reported that sound onsets are analyzed with poorer frequency selectivity [140].

5.3 – Limitations of the model

Limitations of the parametric approach. One limitation of the parametric method is that the optimal filtering can depart from the power law model when considering specific classes of speech sounds. The model will have to be loosened in future work if one wants to investigate the statistical structure of speech in even greater detail. The parametric model is still convenient because speech sounds can be compared at a fine level with a single parameter. The parametric approach is a tool that comes not in replacement of Independent Component Analysis but as a simple method to investigate the variations of β on short time scales. The method can suffer from several biases since it depends on a few experimental settings (weighting strategy, choice of Q_0, normalization and filtering), but the overall description that has been presented is robust to changes in these parameters and consistent with previous work. The benefit of this method has been to draw up the "big picture" of the fine-grained statistical structure of speech, in relation to key acoustic factors, and to provide an insight into plausible efficient strategies for short-time scale speech coding. However, this method covers only one aspect of speech coding, and other machine learning techniques should be involved in the study of the statistical regularities of speech as well. In particular, the correlations that intervene in the determination of β are under 10 ms, regularities at higher time scales have also to be exploited by efficient speech coding systems to be exhaustive.

Asymmetry of auditory filters. One limitation of Gabor filters for the comparison with the auditory system is that cochlear filters are not symmetric [31], in particular in response

to high intensity sounds (fig. 5.6). The filters at the output of ICA do not present a strong asymmetry neither, but this can be obtained if sparse response patterns are reinforced by a matching pursuit (MP) algorithm [149] (although it has been reported that gammatone filters learned with MP have longer time characteristics than auditory filters [156]). This remark recalls the discussion on the opposition of paradigms between *analysis* and *reconstruction* (chap. 1, sec. 1.2.2). The non-gaussianity of filters may be the mark of a departure from the *analysis* paradigm. If the atoms are selected, as in MP, they are less constrained by the spread in the time-frequency plane, and have more freedom to fit the signal closely. *Reconstruction* is feasible at a reasonable cost if the inputs can be reconstructed with only a few atoms (sparsity assumption). The asymmetry of auditory filters is consistent with the overall statistical structure of speech:

— Low-level sounds are mainly non-structured sounds (obstruents). For these sounds, the *analysis* paradigm may be more relevant, and we expect impulse responses similar to Gabor-like filters.

— High-level sounds are mainly vowels, which are structured sounds. For these sounds, the *reconstruction* paradigm may be more relevant. The asymmetric impulse responses are consistent with the glottal excitations followed by damped oscillations with an exponential decay. Gammatone filters model this asymmetry with the equation:

$$g(t) = At^{n-1}e^{-2\pi bt}\cos(2\pi ft + \phi)$$

where n is the order of the filter, and b is a measure of 'damping' [80]. If the auditory systems attempts to 'reconstruct' the input, or at least 'select' time-frequency atoms, the question remains, what are the mechanisms (implemented at a central or peripheral level) behind this ability?

Another feature of auditory filters which has not been discussed is that they present a frequency chirp [80].

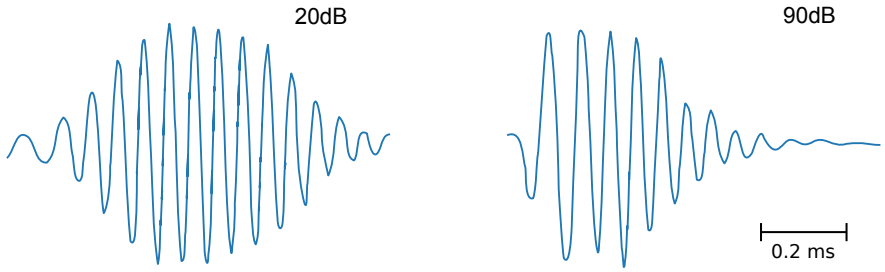


Figure 5.6 – Impulse responses of the basilar membrane (center frequency at approx. 20 kHz) at two input levels: 20dB (left) and 90dB (right). Impulse responses present a strong asymmetry for high intensity levels. *Adapted from De Boer and Nutall, 2002 [39].*

5.4 – Future research

The perspectives of future research are twofold:

1. Confirm and precise the description of the fine-grained statistical structure of speech. In particular, the connexion between statistical structure and acoustic features will have to be examined further.
2. Explore further the relationship between the statistical structure of speech and nonlinear cochlear signal processing (in the context of the efficient coding theory)

The parametric method presented in this thesis is well suited to get the overall behavior of the quality factor for a good decomposition of the data, but these two aspects will require a less constrained model. In addition to computational developments, the investigation of the link between speech statistics and nonlinear cochlear filtering could require new experimental data.

In the following, I present possible avenues of research.

— *Development of a non-parametric model for Q_{10} as a function of frequency and intensity.* The power-law model has proven useful in approaching the fine-grained statistical structure of speech and in formulating new hypotheses. However, this model is too restrictive in order to explore the statistical structure of speech in greater detail. Deviations from the power law model were observed for few phonetic categories in the work of Stilp and Lewicki based on ICA (e.g. vowels) [153]. It was found several times that the quality factor of cochlear filters in humans does not follow exactly a power law in the region 1-8kHz but rather a 'U' shape [124, 165]. These observations raise the question of the appropriate pattern for the quality factor as a function of center frequency when the statistical structure of speech is examined at a sufficient fine level. For example, a non-parametric estimation of Q_{10} as a function of frequency and intensity could be carried out . The same type of model could be used to extend the description of the statistical structure at the phonetic level.

— *Converging efficient coding models and computational models of cochlear signal processing.* The level of cochlear compression being dependent of sound intensity level is only an approximation. The real behavior is actually more complicated because the level of compression depends on the amplitude of the cochlear traveling wave, then on a bandpass filtered version of the signal at each place (this accounts for suppression effects). This property of the cochlear system should be included in the model, at least with a simplified account of this phenomenon. A second possible issue is the question of sound onsets. In figure 5.2, I showed that the onsets of stops and affricates lower significantly the β parameter. Is the special treatment of onsets a feature of the auditory system? The development of the model could make explicit efficient strategies consistent with advanced descriptions of the statistical structure of speech. Ideally, these two aspects (auditory coding models and speech statistics) will benefit each other. All these analyses can be done with Gabor filters as they facilitate the analysis, but, in a final step, this constraint should also be released. As discussed in the previous section, the asymmetry of the auditory filters could be the mark of the enforcement of sparsity in the response patterns (selection of time-frequency atoms). This hypothesis could be tested with algorithms of sparse dictionary learning.

— *Supplement and clarify experimental data on cochlear frequency selectivity.* This last

possibility of future research concerns rather the experimental aspects, although the theoretical developments could also help in clarifying existing data if they are found consistent with the predictions. Frequency selectivity data in the high frequency range 1-8Khz and for different sound intensity levels is still scarce – especially for low-level sounds that are problematic to many experimental settings. This apply in particular to humans, as most measurement techniques are invasive and not suitable for individuals. A new promising (not too invasive) method of measurement proposed recently uses electrocochleography together with a forward-masking procedure [166, 165]. This method could be optimized in certain aspects. A prior clarification would be to distinguish between the Q_{10} and Q_{ERB} factors, that are defined by different calculation procedures and have different behaviors: what accounts for these differences? Eventually, both experimental and theoretical developments will bring new elements to the debate on the specificity of human cochlear tuning. If robust variations are observed, one can hope to answer the critical question: has the human ear evolved to better adapt to speech?

Chapter 6

Characterization of speech rhythm with summary statistics

The last chapter of this thesis is about characterizing speech rhythm based on summary statistics. Although the motivation of this work was initially related to the topic of the efficient coding of speech, which is the central focus of this thesis, it has evolved into a largely independent work. Therefore, the motivation and methods of this work are presented separately in this chapter.

The aim of this work is to provide a quantitative basis to the traditional categorization of languages according to their rhythmic properties. Languages are classically divided into two (or three) rhythmic families: syllable-timed languages and stress-timed languages (the third category being mora-timed languages). However, there is little justification for this division based on explicit measures, and its validity has been challenged several times. Here, I propose to bring additional evidence for the notion of "speech rhythm" through a task of language classification. The classifier is a recurrent neural network trained on a large number of sentences in 15 different languages. The inputs are the time courses of several summary statistics including intensity. The output is a probability vector whose maximal value is the predicted language. The abstract knowledge learned by the neural network can be visualized by producing proximity maps from output vectors.
The first results are promising, and show that characteristics of speech rhythm, or phonological properties related to the perception of speech rhythm, can be captured by an artificial neural network. At the end of the chapter, I discuss how these preliminary results could be improved by future research work.

6.1 – Motivation

The existence of language categories according to speech rhythm is a classic issue of linguistics. Research in this field is also motivated by the fact that rhythm seems to be a prominent characteristic noticed by young infants, therefore possibly playing a role in language acquisition.

The notion of speech rhythm. The classical division of languages between syllable-timed languages and stress-timed languages is due to K. Pike (1945). It was inspired by an earlier account of speech rhythm by L. James who characterized some languages as having a *machine gun* rhythm (syllable timing), and others a *Morse code* rhythm (stress timing). An explanation for this dichotomy was proposed later by D. Abercombrie in the late 1960s, who introduced the concept of *isochrony*. According to him, the duration of every syllable would be almost equal for syllable-timed languages (Italian, Spanish, French, Finnish...) whereas the interval between two stressed syllables would be equal for stress-timed languages

(English, Dutch, German, Swedish...). However, the theory of isochrony was confuted by experiments in the early 1980s, in which statistical measures of interval durations were derived from recordings in several languages (for example, the experiments by P. Roach [136]). R. Dauer proposed instead that it is a group of several phonological properties that explains the difference of perception between the two families [38]. The most prominent phonological properties attributed to stress-timed languages are vowel reduction/reduction of unstressed syllables and the complexity of syllabic structure (i.e. stress-timed languages have both small and large consonant clusters, as in the word '(str)o(k)e'). She also disputed the fact that there were two distinct families, claiming that there exists rather a *continuum* from syllabic-timed to stress-timed languages. A third class of speech rhythm often mentioned is the mora-timed languages, with Japanese being its the most noticeable member.
Despite the relative success of rhythmic metrics based on Dauer's intuition (see next section), the weak empirical evidence of the existence of different speech rhythm classes led researchers in the field to call for a 'new paradigm' that would include a more rigorous and more flexible definition of speech rhythm [87, 9, 160, 119]. An idea common to these articles is that new research on speech rhythm should focus less on explicit timing, but more on the role of perception. We tend to perceive periodic patterns even though the exact timing does not present a regular pattern. It has been proposed that the neural processes act as a 'regularizer' of speech timing, both at the level of perception and motor control. This idea is materialized by the coupled-oscillator model [60]. New descriptions should also encompass a wider range of factors assumed to influence the sense of speech rhythm (e.g. intensity contrasts, intonation), rather than only duration features, and to evaluate the contribution of each factor. Research should also clarify whether speech rhythm is only an epiphenomenon of phonological differences between languages, or whether it plays a role in communication performance [87] or speech acquisition (as suggested by the behavioral data presented in the following paragraph).

Infant capacities for language discrimination. High-amplitude sucking experiments carried out on infants (see frame below) showed that individuals begin to learn statistics of their mother tongue prenatally, and shortly after birth. At their birth, infants prefer their mother tongue on other languages [110], they can also discriminate some foreign languages without previous exposition [118, 109, 129]. Prosodic cues more than phonetic cues are believed to explain these discrimination abilities. One argument is that high frequencies above 800 Hz are filtered when the infants are exposed to sounds prenatally, making the phonemes less distinguishable [110]. Further experiments suggest that infants abilities for languages discrimination could be based on speech rhythm: newborns can distinguish English from Japanese or English from Italian but not English from Dutch [109]. They can also make the difference between English + Italian against Dutch + English but not English + Italian against Dutch + Spanish [118]. Speech rhythm does not describe the full phonological differences that could be figured out by infants, nevertheless it seems that newborns are particularly sensible to phonological properties – like syllable complexity – that contribute to the rhythmic characteristics of a language. Attention to speech rhythm could help very young children to learn some phonological and syntactic properties of their own language (phonological/prosodic 'bootstrapping'). Two or three months after their birth, babies less notice rhythmic differences between foreign languages while being better at recognizing their mother tongue [109]. As children progress in the acquisition of speech, their abilities become more and more specific to their own language. From six to twelve

months, they lose their ability to discriminate non-native phonemic contrasts while having a robust categorical perception of the phonemes used in their language [89, 88, 126].

High-amplitude sucking experiments

Many experiments for the demonstration of infants speech abilities rely on the technique of High Amplitude Sucking (HAS), also called non-nutritive sucking [25]. The protocol was proposed by Siqueland and Delucia in 1969 for visual stimulations [146]. Its first use in the context of speech dates back to 1971 with an experiment by PD Eimas et al. [46]. The experiment consists in delivering sounds to Infants each time they produce a strong or high-amplitude suck. The rate of high-amplitude sucks is recorded during the time of experiment. Newborns rapidly realize the connection between the production of a sound stimulus and their sucking, and they have more frequent high-amplitude sucks as their interest in the stimulus increases. The assumption behind HAS experiments is that infants show more interest when a new type of stimuli is detected, hence the protocol can help to evaluate speech-related discrimination abilities in infants. In 1986, DeCasper and Spence adapted the method of high-amplitude sucking so that newborns could choose their preferred stimulus. With this technique they showed that infants have a strong preference for a text that was recited before birth [40].

6.2 – Previous work

Rhytmic metrics. Different proposals have been made to find quantitative bases for the division of languages according to speech rhythm [104]. The recent proposals are no longer based on the concept of *isochrony* but are derived from Dauer's assertion that rhythm reflects phonological differences between languages. In 1999, Ramus et al. showed that simple statistics (mean, variance) of the durations of vocalic and consonantal intervals separate stress-timed languages and syllable-timed languages [131]. These measures were based on a consonant-vowel (CV) segmentation of sentences (if two consonants or vowels follow, they are grouped together). Two prominent features of stress-timing are believed to be vowel reduction and complex syllabic structure. Ramus et al. showed that, consistently, the (time) ratio $\%V$ of vocalic intervals, and the standard deviation of consonantal intervals ΔC are lower (or higher, respectively) for stress-timed languages. Grabe and Low proposed comparable estimates but with a local measure of segmental duration variability. They also introduced an estimate that includes a local normalization of the interval durations [64]. Normalization is necessary to take into account the speech rate variability (that can occur even inside a sentence) but there is not a unique way to do it [130]. Despite the success of these rhythmic metrics, at least when considering prototypical stress- or syllable-timed languages, their relevance has been questioned [9, 160]. The main criticism is that they only reflect a few phonological properties of languages, whereas the perception of speech rhythm is believed to depend on a wide diversity of factors. The original work by Ramus et al. suggested that languages cluster into a few rhytm classes, but later studies showed that $\%V$ and ΔC describe rather a continuum when considering a larger number of languages [9]. Another limitation of the rhythmic metrics are that they rely on a CV segmentation, which was done manually, and can suffer from several biases [130]. In addition of being time consuming, this process extracts abstract information that may not correspond to

the information most relevant to the perception of speech rhythm – especially considering the perceptual abilities of infants. Galves et al. made a similar analysis with a measure of sonority instead of a CV segmentation [55]. This proposal was based directly on the acoustic signal instead of abstract phonetic categories.
In the work presented in this chapter, rhythm is considered as a high-level characteristic that is exploited in a task of language identification based on the temporal evolutions of simple signal statistics (e.g. intensity patterns). This task is reminiscent of a 20-year-old study by Cummins et al. who used the newly invented LSTM neural network (presented in next section) to investigate the role of prosodic cues (intensity, pitch F_0) in language identification [36]. However, this work was carried out at the early stages of the LSTM, without the availability of large speech databases. In 2005, another work by JL Rouas et al. investigated a classification task with a Gaussian mixture model and measures of CV durations at the syllable level [138].

Acoustic correlates and efficient coding of speech. Prior to this thesis, R.G. Erra and J. Gervain conducted a cross-linguistic comparison of representations learned with Independent Component Analysis. One of the discoveries was that the β parameter (Q_{10}/f_c slope, and the main focus of the previous chapters) is negatively correlated with the vocalic percentage $\%V$. This finding is not a surprise in the light of the fine-grained description of the statistical structure of speech (chap. 4 & 5). The authors therefore made a link between the rhythm classes and the cross-linguistics differences for the optimal representation of speech. This work also explains the genealogy of the research presented in this thesis, and the connection between the efficient coding of speech and the study of speech rhythm, which seem to be separate topics at first glance. The initial concern was whether the time course of β in sentences of different languages could be an account of speech rhythm. However, it appeared that the β parameter was too unreliable when considering single occurrences (of around 20 ms), and less important than other parameters (especially intensity). As shown in the previous chapters, the parameters β and h are also largely redundant with sound intensity level.

6.3 – Methods

I propose a novel method for the investigation of the acoustic correlates at the basis of speech rhythm perception. It involves a recurrent neural network trained for a language identification task. The aim is to see whether classification errors are consistent with the empirical evidence of infant capacities for language discrimination, but also to find out if the neural network develops a sense of speech rhythm during training. The abstract knowledge that the classifier acquires can be visualized from the activations of artificial neurons when examples are passed to the neural network.

In the Methods section, I introduce first the techniques that will be used (recurrent neural networks, methods of data visualization), then I present the specificities of experiments (data, model architecture & training).

6.3.1 Recurrent neural networks

Recurrent neural networks are powerful models for the detection of temporal patterns specific to a category of inputs. Long short-term memory networks (LSTM) is a special case of recurrent neural networks. In a LSTM, a cell state is associated with each computation unit. The cell states play the role of a memory for the neural network, which can then remember past events on longer time scales than standard reccurent neural networks [74]. LSTM networks can find patterns in formal languages [58] or temporal structure in music [45], they have been used for automatic speech recognition [20].

a) Standard Recurrent Neural Networks (RNN)

Recurrent neural networks emerged during the 1980s. They are similar to standard neural networks but can have recurrent connections between neural units, so that the network can handle inputs of varying size. Hence they are well adapted for the analysis of time series. They appear in particular in automatic speech recognition (ASR) [20], handwriting recognition [66] and many other pattern recognition problems. They are also used for machine translation [157].

When « unfold », a recurrent neural network is equivalent to a standard neural network but with equality constraints on the weights (fig. 6.1). It makes recurrent neural networks similar to convolutional neural networks which are also defined with equality constraints on weights. Learning works as for a convolutional neural network, through backward propagation of errors on the unfold neural network. To prevent from a too large computational cost the computation of the gradient is usually truncated at a number of steps backward (in my experiments, this number is typically 32 or 64).

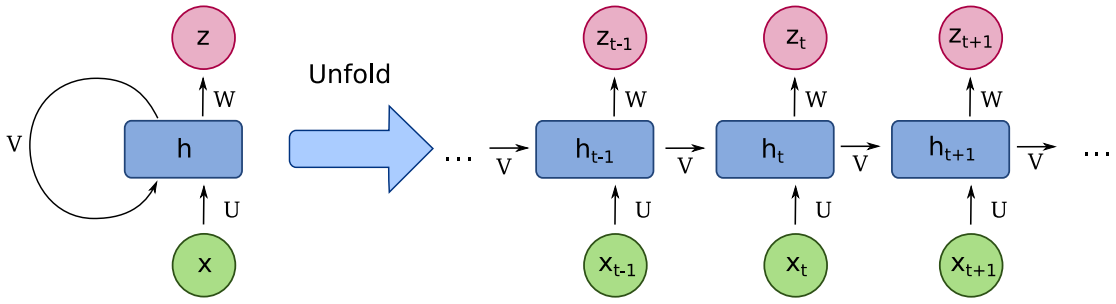


Figure 6.1 – Diagram of a 1-unit recurrent neural network connecting an input vector with the output and the equivalent « unfold » diagram of the structure.

An example of RNN model with one recurrent unit associated with a (possibly multidimensional) state h is shown in fig. 6.1. It connects an input vector x with an output vector z. The equations that define the temporal evolution are :

$$h_t = \sigma_h(Ux_t + Vh_{t-1} + b_h)$$
$$z_t = \sigma_z(Wh_t + b_z)$$

where σ_h and σ_z are activation functions (defined element-wise for multidimensional vectors). Classically the activation functions are the sigmoid function or the hyperbolic tangent. For the output layer it can be the linear function (regression problem) or the softmax function (classification function).

The vanishing gradient problem. Recurrent neural networks have to deal with the so called *vanishing gradient problem*. This issue prevents recurrent networks from modifying their weights in function of past events [73]. During the training, one has to minimize a functional $E(z)$. The dependence on the weight U_{mn} is:

$$\begin{aligned}\frac{\partial E(z_t)}{\partial U_{mn}} &= \cdots + \frac{\partial E(z_t)}{\partial z_t}\frac{\partial z_t}{\partial h_t}\frac{\partial h_t}{\partial h_{t-1}}\cdots\frac{\partial h_{t-k+1}}{\partial h_{t-k}}\frac{\partial h_{t-k}}{\partial U_{mn}} + \cdots \\ &= \cdots + \frac{\partial E(z_t)}{\partial z_t}\frac{\partial z_t}{\partial h_t}\frac{\partial h_t}{\partial h_{t-1}}\cdots\frac{\partial h_{t-k+1}}{\partial h_{t-k}}\sigma_h'(Ux_{t-k} + Vh_{t-k-1})E_{m,n}x_{t-k} + \cdots\end{aligned}$$

where E_{mn} is the matrix where the only nonzero term is is of value 1 and at position (i, j). Only the term that is related to the event x_{t-k} is displayed. For normal values of the activation function and network weights the product $\frac{\partial h_t}{\partial h_{t-1}}...\frac{\partial h_{t-k+1}}{\partial h_{t-k}}$ decreases exponentially in function of k. If the neural network must consider a past event, the correction of the weights will decrease exponentially with the delay between the past event and present. As a consequence, the network will not be able to take into consideration past events.

b) Long short-term memory networks (LSTM)

Long short-term memory network is the most popular recurrent architecture that proposes a solution against the vanishing gradient problem. It was introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997 [74]. The idea behind LSTM is that each unit is bound to a hidden state h but also to a cell state c that plays the role of memory. c_t is obtained from c_{t-1} with a constant gain of value 1. This way errors propagate backwards without the phenomenon of vanishing gradient and the network can learn to memorize past events (up to 1000 steps backward [74] or even more). The cell state can be modified through a gate that allows or not the update (*input gate*). Another gate controls the output of the unit (*output gate*). The most common version of LSTM unit also has a *forget gate* that can reset the cell state [59]. The basic architecture is represented in figure 6.2.

Equations.
Initial values are $c_0 = 0$, $h_0 = 0$. The operator $\circ$ symbolizes the Hadamard product (element-wise product). The activation functions are generally the sigmoid function and the hyperbolic tangent. The temporal evolution is governed by the following equations :

$$\begin{aligned}F_t &= \sigma(W_F x_t + U_F h_{t-1} + b_F) && \text{(forget gate)}\\ I_t &= \sigma(W_I x_t + U_I h_{t-1} + b_I) && \text{(input gate)}\\ O_t &= \sigma(W_O x_t + U_O h_{t-1} + b_O) && \text{(output gate)}\\ c_t &= F_t \circ c_{t-1} + I_t \circ \tanh(W_c x_t + U_c h_{t-1} + b_c)\\ h_t &= O_t \circ \tanh(c_t)\\ z_t &= f(W_z h_t + b_z).\end{aligned}$$

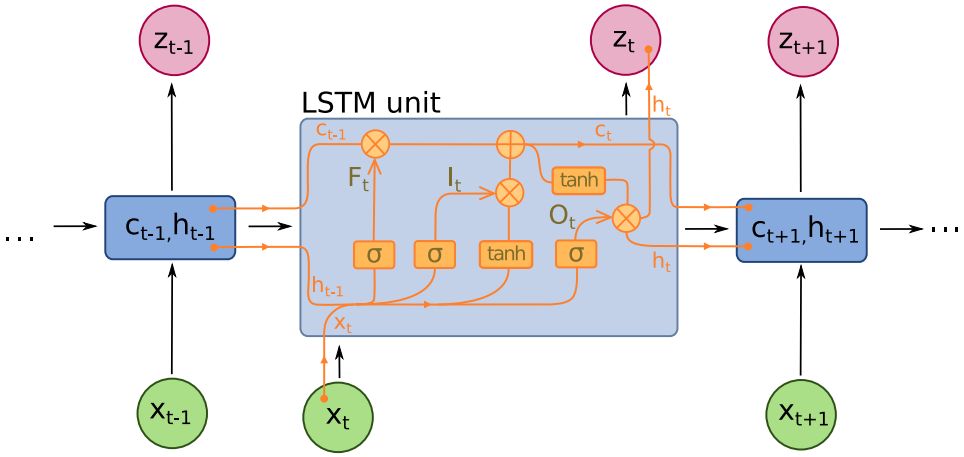


Figure 6.2 – A simplified diagram for a one-unit LSTM network. From bottom to top : input state, hidden state and cell state, output state. Gate functions are sigmoids or hyperbolic tangents. The other operators are element-wise plus and multiplication. Weights are not displayed.

c) Variant: Gated Recurrent Unit (GRU)

Gated Recurrent Unit (GRU) networks are an alternative to LSTMs introduced in 2014. They have comparable performance on prediction of times series such as polyphonic music scores and speech data [32]. A Gated Recurrent Unit has fewer parameters than a LSTM unit. Compared to LSTM, GRUs have no cell state, the forget gate and input gate are merged into a unique gate (*update gate*) and the output gate is replaced with a *reset gate* (fig. 6.3).

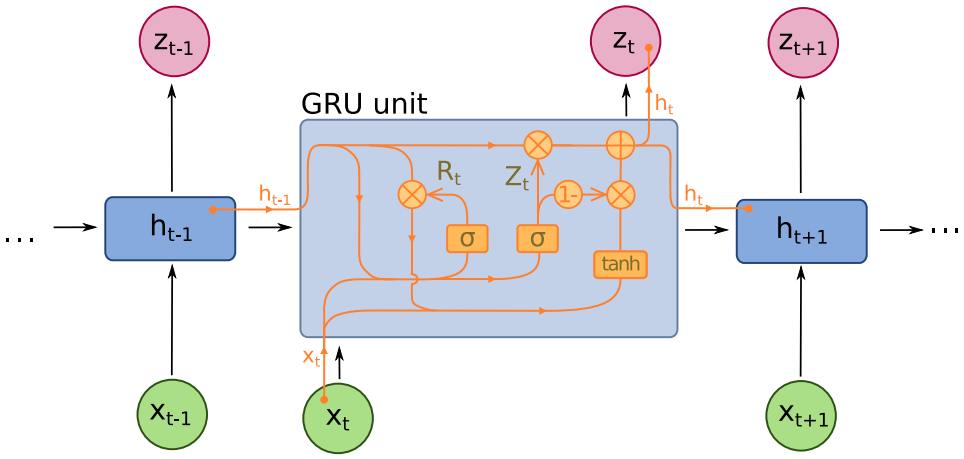


Figure 6.3 – A simplified diagram for une-unit GRU network.

Equations:
Initial value : $h_0 = 0$.

$$Z_t = \sigma(W_Z x_t + U_Z h_{t-1} + b_Z) \quad \text{(update gate)}$$
$$R_t = \sigma(W_R x_t + U_R h_{t-1} + b_R) \quad \text{(reset gate)}$$
$$h_t = Z_t \circ h_{t-1} + (1 - Z_t) \circ \tanh(W_h x_t + U_h(R_t \circ h_{t-1}) + b_h)$$

6.3.2 Data visualization

Techniques of visualization are dimensionality reduction methods that aim to facilitate the analysis and interpretation of data by positioning data points in a low dimensional space (usually two or three dimensions). The scatter plots that are obtained by the techniques of visualization can help to identify clusters or similarities in data. Two techniques are introduced here : *multidimensional scaling (MDS)* and *t-distributed stochastic neighbor embedding (t-SNE).*

Context. Let $x_1, x_2, \cdots, x_N$ be N points living in a high dimensional space (dimension p). We seek to place points in a space of low dimension $m < p$ with N new points $y_1, y_2, \cdots, y_N$ keeping the similarities : two points that were closed in the initial space are closed in the created space (and inversely). A distance matrix D is given. It can be defined with the euclidian distance $d_{ij} = ||x_i - x_j||_2$ or with another dissimilarity measure.

a) Multidimensional scaling

Multidimensional scaling (MDS) is the name for several techniques that seek to position points in a low dimension space ($m = 2$ or $m = 3$) from a distance matrix defined from high dimensional data [71]. These techniques seek to minimize a cost function $S(y_1, y_2, \cdots, y_N)$ called *stress*. There are two kinds of multidimensional scaling :

— *metric multidimensional scaling* tries to have the euclidian distances in the low dimensional space as close as possible to the actual distance matrix.

— *non-metric multidimensional scaling* puts stress on the ranks of similarities rather than the actual value of distances. It considers than the order of similarities is more important than the extent of these similarities.

Metric multidimensional scaling. Metric MDS includes least-squares MDS that correspond to the least-squares error :

$$S(y_1, y_2, ..., y_N) = \sum_{i \neq j} (d_{ij} - ||y_i - y_j||)^2 .$$

However this formulation does not have an explicit solution in general. The stress function is replaced in classical multidimensional scaling by

$$S(y_1, y_2, ..., y_N) = \sum_{i \neq j} (b_{ij} - \langle y_i, y_j \rangle)^2$$

where b_{ij} is defined by $b_{ij} =< x_i - \overline{x}, x_j - \overline{x} >$ with $\overline{x} = \frac{1}{N} \sum_{i=1 \cdots N} x_i$. More generally, B is called the similarity matrix and can be obtained from the distance matrix D by double centering :

$$B = (I - \frac{1}{N} J) D^2 (I - \frac{1}{N} J)$$

where J is the matrix of ones of size $N \times N$. MDS benefits from this formulation as it has an explicit solution that is derived by the eigenvalue decomposition of B. Let $\lambda_1, \lambda_2, \cdots, \lambda_m$ be the m largest eigenvalues of B and $e_1, e_2, \cdots, e_m$ the corresponding eigenvectors. Then a solution for classical MDS is to take the columns of $Y = \Lambda_m^{1/2} E_m^T$ where E_m is the matrix of the m eigenvectors of B and Λ_m the diagonal matrix of eigenvalues.

Non-metric multidimensional scaling. Non-metric multidimensional scalings give more importance to rank of distances than to the actual values of these distances. Their objective is to minimize the stress

$$S(y_1, y_2, ..., y_N) = \sum_{i \neq j} \left(d_{ij} - f(||y_i - y_j||)\right)^2 ,$$

with the particularity that the function f can adapt to the simulation during the optimization. The function f is chosen as a monotonic regression of the points $(||y_i - y_j||, d_{ij})$.

b) Stochastic neighbor embedding (SNE, t-SNE)

The goal of multidimensional scaling is to preserve distances of high-dimensional points in a low-dimensional space. However, since high-dimensional spaces by nature have different topological behaviors, the result can suffer from several distortions, and the local or global structure of the data can in fact be misrepresented. Classical MDS, which is linear, also tends to favor global structure instead of local structure. Close points in a high-dimensional space often belong (or are close) to a lower-dimensional manifold, and the proximity of points will be better displayed with a non-linear mapping. Distributed stochastic neighbor embedding (SNE), takes another point of view: it is not *distance*s but rather *proximities* that have to be preserved. This means that visualization methods should not be concerned with distorting long-range distances, but with maintaining short and medium-range structure (similarities between close points).

Stochastic neighbor embedding (SNE) is based on a probabilistic interpretation of proximity. SNE defines a probability distribution over pairs of points so that close points will be drawn more often than distant points. The goal of the SNE algorithm is to build a low-dimensional map such that the probability distribution defined in the same way, with the created points, is similar to the original distribution. More explicitly, let $p_{j|i}$ be defined by:

$$p_{j|i} = \frac{\exp\left(-d_{ij}^2/(2\sigma_i^2)\right)}{\sum_{k \neq i} \exp\left(-d_{ik}^2/(2\sigma_i^2)\right)} ,$$

where σ_i is set with the method described further below, and let p_{ij} be:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \text{ for } i \neq j, \; p_{ii} = 0.$$

The probability distribution q is defined in the same way given the new points, with:

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)} . \tag{6.1}$$

The (symmetric) SNE algorithm seeks to minimize the cost function defined by the KL-divergence between the two distributions:

$$S(y_1, \cdots, y_n) = D_{KL}(p||q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} . \tag{6.2}$$

Perplexity. The method has free parameters σ_i that must be set by the user. The standard method is not to set directly σ_i, which depends on the generally unknown geometry of the configuration around the point x_i, but to chose a unique parameter for the entire point distribution, called perplexity, that can be interpreted as a smooth measure of the number of neighbors [162, 169]. Perplexity is defined by

$$\mathrm{perp}(p) = 2^{-H_2(p)} \tag{6.3}$$

where $H_2(p)$ is the entropy associated with the probability distribution p in base 2:

$$H_2(p) = -\sum_{i,j} p_{ij} \log_2(p_{ij}) \ .$$

In particular, if we assume that $p_{ij} = 1/m$ inside a cluster of equidistant points, then $\mathrm{perp}(p) = m$. From a given value of perplexity, σ_i is found using a root-finding method. It can vary from one point to another depending on the geometry (therefore, one implicit assumption of SNE is that the clusters have approximately the same number of points).

t-distributed stochastic neighbor embedding (t-SNE). Standard SNE still suffer from a geometric issue, when going from a high dimensional space to a lower dimensional space, known as the 'crowding problem'. The reason behind the crowding problem is that the points at a fixed distance around a given point represent a wider volume in high dimensional spaces than in low-dimensional spaces for medium and long-range distances (fig. 6.4). This difference results in an unwanted concentration of points, potentially belonging to different clusters, in the SNE mapping, since the volume available to display points at a same distance is reduced. L. Van der Maaten and H. Hinton proposed in 2008 to alleviate this problem with the t-SNE algorithm [162]. t-SNE replaces the Gaussian distribution in eq. 6.1 with a heavy-tail distribution, the Student distribution with one degree of freedom:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}} \ . \tag{6.4}$$

This trick provides more space for points at moderate distance to accommodate between them.

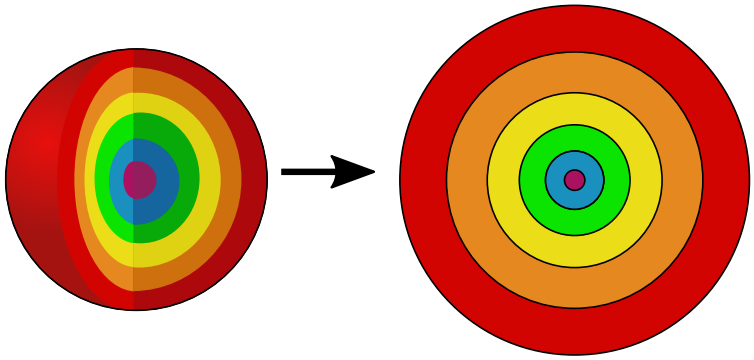


Figure 6.4 – Illustration of the phenomenon at the origin of the 'crowding problem': layers of constant radius increments in dimension 3 (*left*) are mapped by increasing radius increments in dimension 2 (*right*) if the transformation preserves the same volume ratios between layers.

The minimization of the cost function (eq. 6.2) is done with a gradient descent algorithm. However, since the cost is a non-convex function, the result depends strongly on the initialization. It is necessary to try the algorithm with different initial configurations to check the robustness of the representation learned.
The t-SNE algorithm has been popular in recent years for the visualization of labeled samples, using feature vectors learned by an artificial neural network as input points.

6.3.3 Experimental settings

The hypothesis at the basis of this work is that the sense of speech rhythm arises from the search of regular temporal patterns in the variations of intensity and structure. The most typical examples of structure changes are the CV (consonant-vowel) alternations. The knowledge of the rhythmic characteristics of languages is assumed to be gained by training a recurrent neural network (LSTM or GRU) for a language identification task. The neural classifier is relevant to the modeling of perceptual processes and able to detect complex time patterns involving precise timing [59]. When trained on a large database in a supervised fashion, a LSTM newtork can achieve high performance and capture high-order regularities while being robust to context changes (e.g. speech rate). In particular, it has been successfully used for automatic speech recognition [65, 6].

a) Data

Inputs. The input of the recurrent neural network is a three dimensional vector that evolves in time. The three dimensions are:

1. *Intensity* (root mean square value) in dB
2. The h *score*, measuring the lack of structure
3. The β parameter (estimated by a softmin function), indicating if the sound has more frequency or time structure.

An example of input is shown in fig. 6.5. The intensity changes indicate the main prosodic boundaries (silences between groups of words or syllables) and the alternations between low-energy and high-energy phonemes (especially consonant-vowel transitions). A rise in intensity can also indicate a stressed syllable. The h score and the β parameter were presented in the previous chapters, and are summary statistics of signal structure in the high frequency range. A value of β close to 0 indicates temporal structure, while a value close to 1 indicates frequency structure. All inputs were normalized so that the values were typically between 0 and 1.

Each sample lasts 8 seconds. If the duration of the sound examples did not match this duration, they were concatenated (with a short pause) or truncated accordingly. Different sampling frequencies were tested for the three-dimensional input vector. Most of the times, the sampling frequency is 30 Hz.

Datasets. Speech data was an aggregate of three collaborating on-line databases, namely *Tatoeba*, *LibriVox* and *VoxForge*. The databases present a diversity of recording conditions. The Tatoeba and VoxForge databases are made up of short sentences (~2s) of conversational speech. On the other hand, the LibriVox database contains longer files of texts read by the users. The reason I used several databases is not the number of sound examples, which can be high, especially for the most widespread languages, but because the number of different

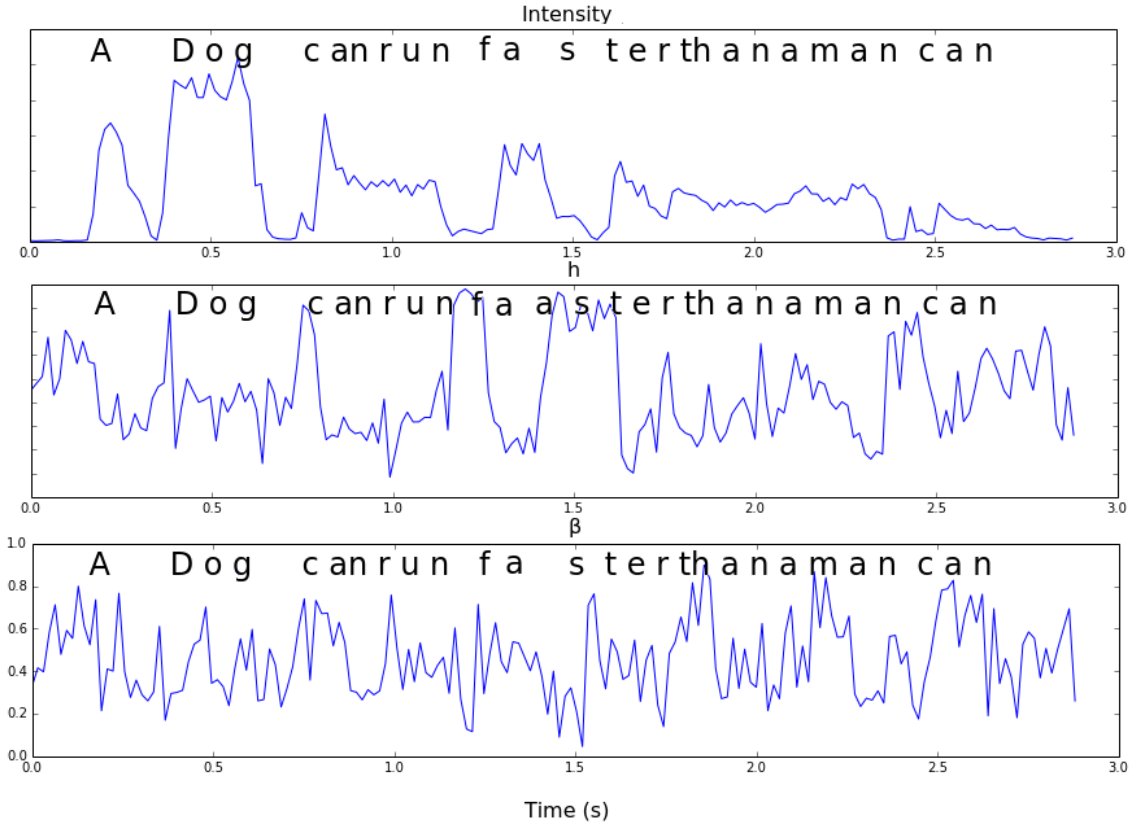


Figure 6.5 – Input for a sentence of the TIMIT database.

speakers for a language can be low. The reduced number of speakers can prevent the model from generalizing, since the inputs depend on the speaker and on the recording conditions. When the available data was too large, the examples were selected so that the number of different speakers was maximal. Other limitations are that the recordings can be of poor quality and that some speakers are nonnative (VoxForge was considered as poorer quality database on this criterion). LibriVox files very occasionally contain non-speech sounds like music. The databases were cleaned as much as possible before processing.
More information on the databases: Tatoeba, launched in 2006, is a tool for foreign language learners. It collects example sentences translated in various languages by native speakers. Some of these sentences come with an audio file. LibriVox started in 2005. It collects thousands of free public domain audiobooks. Librivox has an API that can be used to retrieve information on the audiobooks. The VoxForge database contains examples along with text transcriptions. Its aim is to provide resource for the training of open source automatic speech recognitions systems.

Languages. Only the languages of interest and with sufficient data were selected. There is a total of 15 languages: English, German, Italian, Spanish, Japanese, French, Swedish, Danish, Polish, Mandarin, Finnish, Hungarian, Russian, Portuguese, Dutch (in the inverse order of the number of examples). Most of them are European languages (13), two of them are East Asian languages (2). Table 6.1 shows the number of examples by languages in the training set.

Table 6.1 – **Number of samples by languages in the training set for the identificiation task**. The percentages indicate the proportion it represents in relation to the total number of examples (over 120 000).

English	14080 (11.42 %)	Finnish	7310 (5.93 %)
Spanish	13270 (10.76 %)	Polish	5500 (4.46 %)
Italian	11900 (9.65 %)	Russian	5510 (4.47 %)
Korean	10740 (8.71 %)	Danish	4930 (4.00 %)
German	10560 (8.57 %)	Mandarin	4910 (3.98 %)
Portuguese	9340 (7.58 %)	Dutch	3220 (2.61 %)
French	8820 (7.16 %)	Japanese	2850 (2.31 %)
Swedish	8280 (6.72 %)	Hungarian	2050 (1.66 %)

b) Model & training

Architecture and error function. The model was a recurrent neural network with 3 layers of 128 GRU units (variants of LSTM) each. A linear followed by a softmax layer were placed on top of the GRU layers with $N = 15$ outputs corresponding to the 15 languages. The output vector is written

$$z_t = f(W_z h_t + b_z)$$

where h_t is the last hidden layer, (W_z, b_z) are the matrix of the weights and the biases for the linear layer, and f is the softmax activation function, with:

$$[f(h)]_i = \frac{\exp^{h_i}}{\sum_{j=1}^{N} \exp^{h_j}},\ i = 1 \cdots N\ .$$

The output vector z_t can be interpreted as a probability vector at instant t (in the Bayesian sense): $z_{t,i}$ is the estimated probability that the sound example is an utterance of the language i. The error function (loss) is the cross-entropy between the output vector z_t and the real input y. Let denote by $y_i = \delta_{i,c(x)}$ the vector such that $y_i = 1$ if and only if $i = c(x)$, the class to which the sound belongs. Then the error function is:

$$E(z_t) = -\sum_{i=1}^{N} y_i \log(z_{t,i})\ , \tag{6.5}$$

corresponding to the gradient:

$$\frac{\partial E}{\partial h_t} = z_t - y\ .$$

The objective of the neural network is to minimize the expected loss approximated by the empirical loss, that is the error function averaged over all the samples. To offset the imbalance of the dataset – the uneven number of samples by languages or by speaker – I incorporated weights in the cost function (eq. 6.5) to give more importance to the classes with rare examples. The normalization was done with a weight on the form $1/(K + m)$ where K is a constant and m is the number of samples for the speaker or the number or speakers for the language considered.

Training. The network was trained with truncated backpropagation of the gradient with 32 or 64 steps (1 or 2 s) depending on the experiment. The initial points for the backpropagation

algorithm were considered every one second (meaning that only the outputs at intervals of one second were considered). The training required between 30 and 60 epochs [1]. The meta-parameters (architecture, gradient descent) were set to minimize the error on a test set. The performance of the classifier was then evaluated separately on a validation set. The three sets (training, test, validation set) had no speaker in common. The test set and validation set typically had 10 000 examples each. The model was implemented with TensorFlow [1]. The optimization was done with a stochastic gradient descent algorithm (Nesterov accelerated gradient scheme). To add more robustness to the network, the data was augmented by applying distortion functions to the inputs, in particular intensity. The training implied other standard techniques (e.g. dropout).

6.4 – Results

Performance. Two performance measures are the classification error (the proportion of wrong classifications) and the *top3* error, which evaluates the proportion of examples whose class did not appear in the three languages with the highest probability scores. On the test set, the classification error was 60% (chance level: 93%) and the *top3* error was 33% (chance level: 60%) with a sampling frequency of 30 Hz. Overall, the results showed a good level of performance, considering that the inputs contained poor segmental information, and the fact that performance was not the first objective of the experiments. The number of successes for the identification task was high below chance level. The measured perplexity (eq. 6.3), corresponding roughly to the number of equiprobable classes as seen from the classifier, was 3.75. The classification error dropped to 50% (*top3* error: 22%) when the sampling frequency was twice higher (60Hz). In the same settings, the classification on the training set was 35% (*top3* error: 14%), showing that overfitting was limited. However, with a high sampling frequency, the classifier could exploit phonological differences that may not be attributed to 'rhythm'. Therefore, even though the performance was higher, the classifier learned with a lower frequency sampling was generally preferred. I found that the most important feature for the classifier is intensity level. As discussed earlier, the β parameter is not a measure reliable enough, and the h score is redundant with intensity: sounds with a high h score (mainly consonants) are most often associated with a low intensity value, and low h scores (mainly vowels) are associated with a high intensity value. The addition of the h score to intensity only improved the success rate by a few percent.

Proximity between languages. An example of confusion matrix is shown in fig. 6.6. The confusion matrix is complex to interpret as such, and some errors are more frequent than others without any apparent reason (e.g. Spanish is often chosen as a predicted label). To have a more robust measure of proximity, I computed the divergence between probability vectors considering a high number of samples. Specifically, I considered 512 examples (x_n) at random and computed the associated output probability vectors (z_n). These vectors were used to compute new probability vectors $\omega_{i,n}$ for each language i whose coordinate n is proportional to $[z_n]_i$. The "distance" between language i and language j was computed

1. Epoch: number of iterations required for the neural network to go through all the samples of the training set.

with the KL divergence:

$$d_{i,j} = D_{KL}(\omega_i||\omega_j) = \sum_n \omega_{i,n} \log\left(\frac{\omega_{i,n}}{\omega_{j,n}}\right) \ .$$

An example of result is represented in fig. 6.7. The matrix obtained is not a distance matrix strictly speaking because the KL-divergence is not symmetric. By symmetrization, or using another divergence function like the Hellinger distance, the matrix can be used for multidimensional scaling (fig. 6.8). As an alternative, the output probability vectors (z_n) can be used as an input for the t-SNE algorithm (fig. 6.9). (remark: contrary to maps in previous work, the axes do not have any particular signification here).

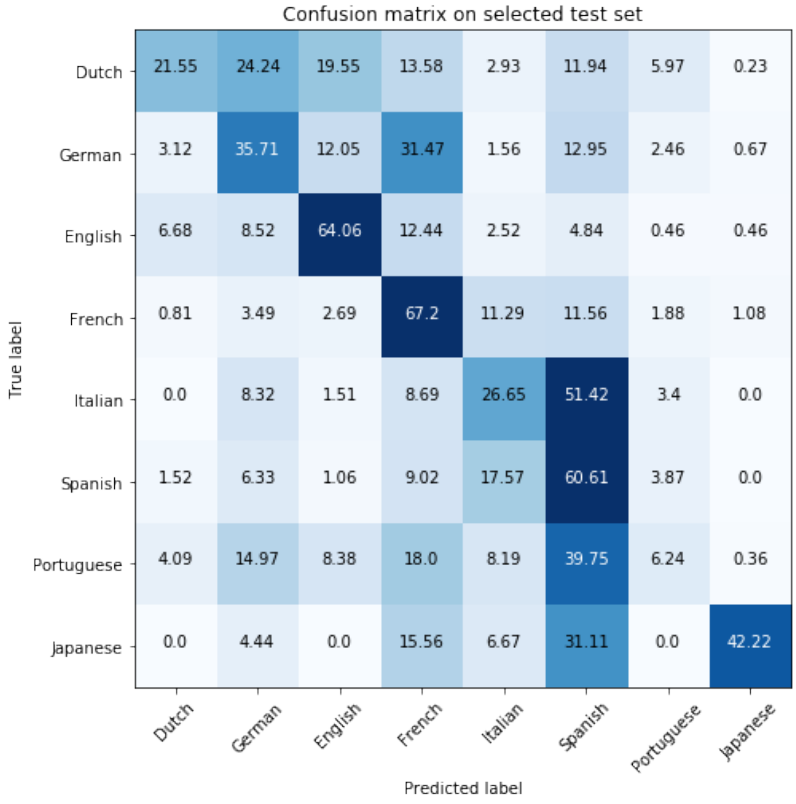


Figure 6.6 – Example of a confusion matrix with a reduced number of languages (sampling freq. at 60Hz).

6.5 – Discussion

Interpretation of the results. The purpose of the experiments described in this chapter was to establish a 'proof of concept' that a recurrent neural network trained for a language identification task is able to gain a sense of speech rhythm, which is made visible with visualization methods (MDS, t-SNE). The results are not final and the interpretations I give are only partial. Overall, the map built by the visualization methods show proximity relationships between languages that are consistent with previous accounts of speech rhythm. Especially, we find groups of stress-timed languages in fig. 6.9, (English, German,

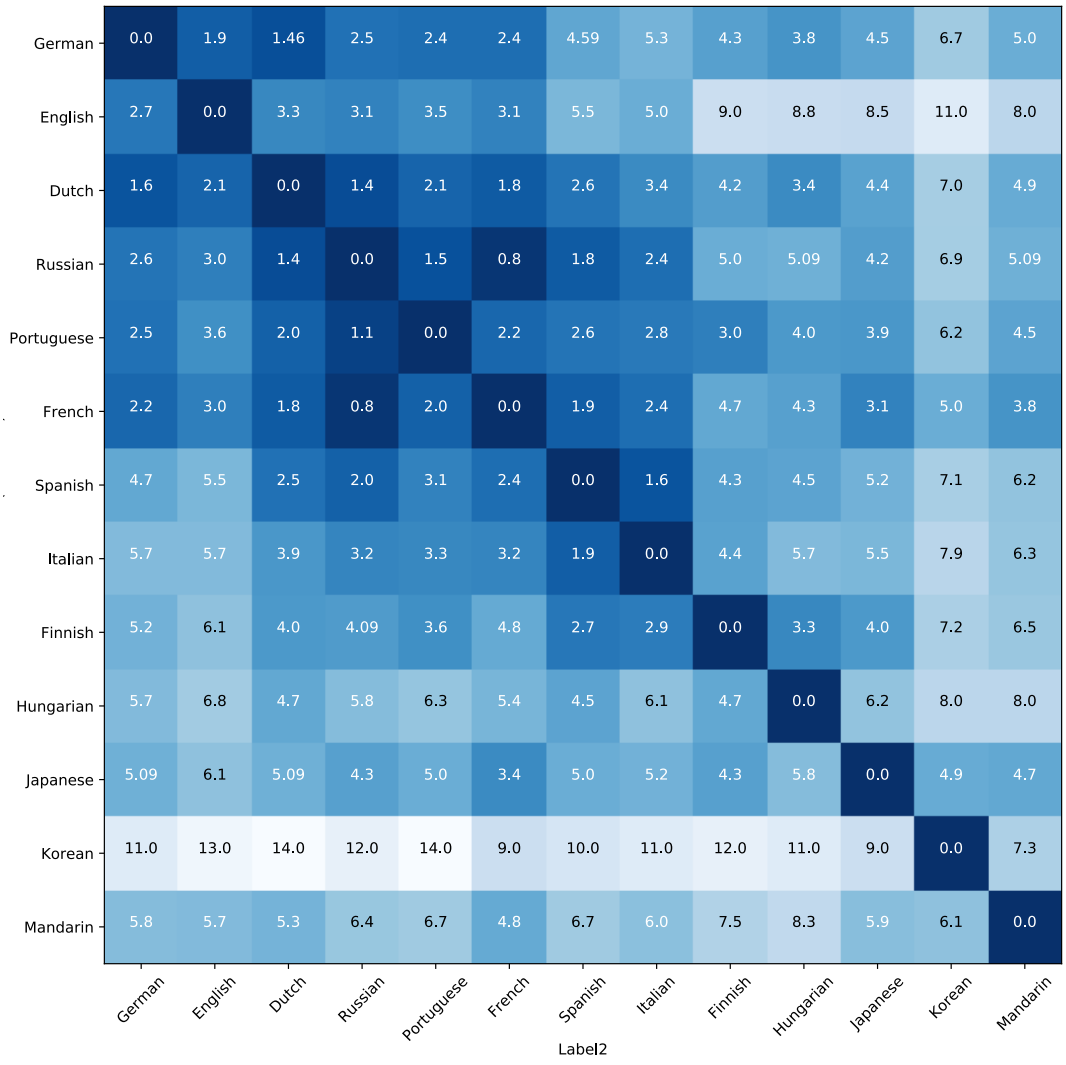


Figure 6.7 – Distance matrix based on the KL divergence between vector probabilities estimated by the network (sampling freq. at 30Hz).

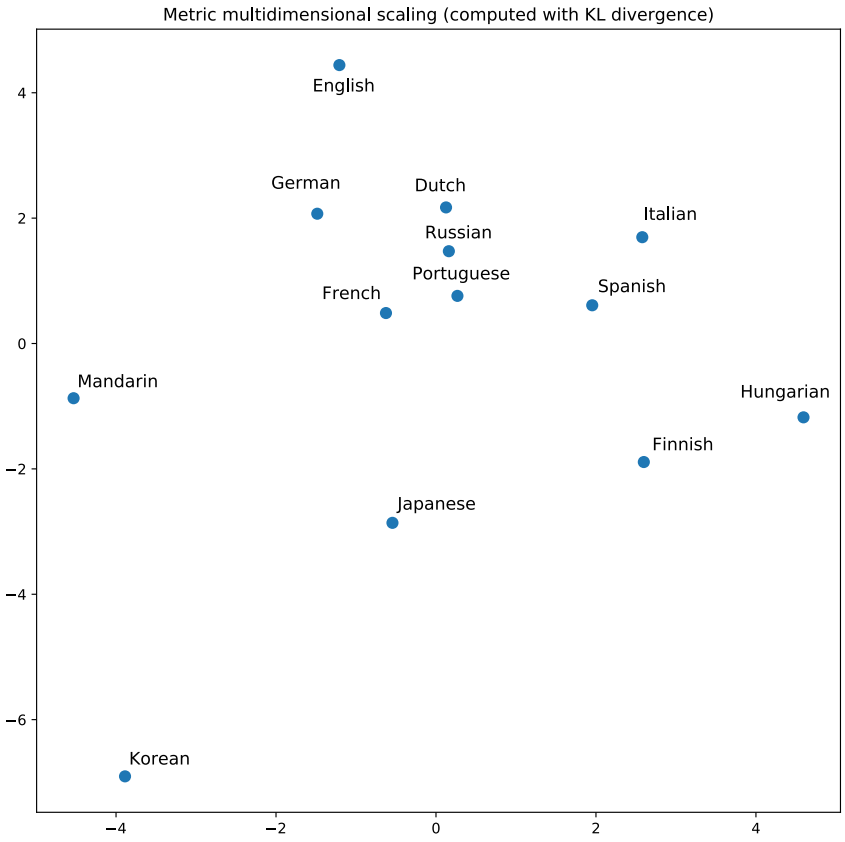


Figure 6.8 – Visualization of the proximity between languages as seen by the neural network based on metric Multidimensional Scaling (MDS).

Figure 6.9 – Visualization of samples represented by their probability vectors at the output of the classifier with the t-SNE algorithm. The labels are the languages as predicted by the classifier (sampling frequency at 30 Hz).

Dutch) or syllable-time languages (Italian, Spanish, Finnish). However, neither the distance matrix (fig. 6.7) nor the maps present obvious clusters, although the t-SNE algorithm fosters the formation of groups of languages. Among these groups, we find geographical groups (e.g. Danish-Swedish, Asian languages) or languages that belong to the same language family (notably Finnish-Hungarian), but it is difficult to find out whether the classifier groups these languages based on broad rhythm characteristics or isolated common phonological patterns. Other links between languages are consistent with known rhythm properties. In particular, Portuguese is known to belong to the two different classes of rhythm, depending notably on whether the speaker is European (stress-timed) or Brazilian (syllable-timed). Coherently, Portuguese bridges the gap between a group of syllable-timed languages (especially Spanish) and a group of stress-timed languages (Russian, Polish). Finnish, which is known to have some features of mora-timed languages, is connected to Japanese. A possible pitfall is that French is in regions marked by stress-timing. The distance matrix (fig. 6.7) reveals that French is the language with the minimum distance from other languages. It is possible that the network failed to generalize well on French examples, or that French has less rhythm cues (compared to stress-timed languages) that make it less distinguishable from other languages (a similar case is Polish, which is also considered as a 'mixed'-timing languages). I also noticed that the French examples were often pronounced by French Canadians, or by foreign speakers. The database will be cleaned of these examples in future experiments.

Future research. These firsts results are promising and encourage further explorations of the quantitative bases of speech rhythm perception using the recurrent neural network approach. In the short term, the model can benefit from several improvements that will clarify the results and their interpretations. The modeling of speech rhythm using artificial neural networks and prosodic acoustic features opens up promising research opportunities in the longer term.
Immediate improvements of the model and experimental design are:

- *Dropping/adding acoustic features at the input of the neural network:* the β and h scores do not give much additional information when considering already the intensity. Inversely, new features that are known to contribute to the perception of speech rhythm [87] can be considered:
  - *Fundamental frequency F_0.* F_0 is a predominant prosodic characteristic related to intonation, often considered as a prosodic cue in its own right (then separated from speech rhythm). However, it also plays a role in speech rhythm both because it marks the alternations between voiced and unvoiced phonemes, and because it contributes to the perception of syllable prominence and grouping [19, 9]. This information will have to be included in future models, but by making sure it does not supersede the other cues.
  - *Spectral balance between low and medium/high frequencies*: in addition to provide information for the detection of CV boundaries (spectral power is dominated by low frequencies for vowels and by high frequencies for consonants), the spectral balance or spectral tilt is known to be an acoustic correlate of the most stressed vowels more informative than changes in intensity [148]. This characteristic has been implemented by computing the intensity for a high-passed version of the signal in a study of the Swedish prosody by Fant et al. [50].
- *The case of Portuguese, and other improvements of the database:* As already mentioned, Portuguese is traditionally considered as a syllable-timed language when

the speaker is Brazilian or stress-timed when the speaker is European. This view is consistent with the location of Portuguese in the generated maps. We could further challenge the model by separating the utterances of Portuguese according to the speaker's origin. Beyond the case of Portuguese, the corpora would be improved by separating the different accents for every language (e.g. European French vs Canadian French), but doing that for all possible pairs would require other data and a lot of effort to build the database properly. The datasets could be augmented with new languages and new databases, such as the *Wide Language Index* (radio podcasts in many languages), and *Common Voice*, an open source collaborative database recently launched by Mozilla. A particular feature of the latter is information on the speaker's accent.

This chapter outlined a new approach for the search of the acoustic correlates that account for the perception of speech rhythm. It paves the way for new investigations that would place the neural network, trained for a classification task, at the center of research. The neural network classifier is capable of detecting complex regularities in the acoustic signal, without explicit modeling steps, simulating neural processes underlying perception to some degree. The difficulty of this approach is that the model also behaves as a "black box". The gain in accuracy is at the expense of a simple and explicit description of speech rhythm. In a way, we find ourselves in the same situation as the behavioral psychology examiner who observes tangible effects in subjects without necessarily having clear explanations for it, and who seeks to design experiments to verify his hypotheses. Nevertheless, by having a baseline computational model, we gain in controllability to set up these experiments. This control is primarily at the input level, since the model can be trained for a specific task with limited features, in contrast to adult subjects that can use a wider range of phonological characteristics. The outputs of chosen artificial examples or misclassified examples can give us information on the classifier's behavior. We can also inspect the inner functioning of the model more easily: for example, we could find out whether the neural network exploits trained features that are correlated with previous rhythmic metrics. Well-designed experiments will eventually help to establish the bases of speech rhythm perception.

Conclusion

This thesis investigated the power laws characterizing the frequency selectivity of efficient speech decompositions, on short time-scales (~10 ms) and in high frequencies (1-8 kHz). I proposed to do this analysis with a simple method, based on a constrained representation model and a sparsity score. The score reflecting the lack of sparsity has been justified following two approaches: an information-theoretic approach, the efficient coding theory (chap. 1), and a signal analysis approach (chap. 2). The procedure was to evaluate the sparsity of decompositions in a family of dictionaries of Gabor filters that were characterized by different power laws for their quality factor (Q factor). This family of dictionaries, called flexible-Gabor wavelets, makes the transition from standard wavelets (multiresolution analysis, constant Q factor) to Gabor frames (uniresolution analysis, linear Q factor). The Gabor dictionaries were motivated both by empirical and theoretical arguments. The dictionaries are in line with the decompositions that were found in previous works based on Independent Component Analysis (ICA), in particular when ICA was applied to phonetic categories instead of speech as a whole. I explained why the parametric method is similar to ICA, with however a strong prior on the representation motivated by the earlier analyses. On the theoretical side, the choice of Gabor filters is supported by the fact that they provide the sparsest responses for time-frequency energy distributions. This property was recalled with the explanation of Lieb's uncertainty principle (chap. 2).

I showed that the power-law exponent, the β parameter, when examined at a fine level of speech, provides a rich interpretation of the statistical structure of speech. The analyses, based on simulated (chap. 3) and real (chap. 4) data, made explicit the relationships between the exponent and the acoustic features of speech. The key acoustic factors were enumerated according to the dichotomy between structured and non-structured sounds. For non-structured sounds, which can be modeled as noise (obstruents), the exponent is related to the localization of the intensity or the spectral power. Low β values – poor quality factor, time decomposition – are found where there is a jump in intensity (e.g. stop onsets). On the contrary, high β values – good quality factor, frequency decomposition – are found when there is a sudden increase/decrease in spectral power, such as in sibilant fricatives ([s], [z]...). Among non-structured sounds, stops and affricates have been shown to be biphasic after the closure: the transient part (burst) is better captured by a time representation, but the rest of the release is a fricative-like sound better captured by a frequency representation. For structured sounds, mainly vowels, the power laws are related to formant bandwidths, partly determined by the degree of acoustic radiation at the lips, then lip opening.

The description of the fine-grained statistical structure of speech could motivate advanced coding schemes based on a nonlinear representation of the acoustic input. I discussed the case where the representation has to adjust according to sound intensity level (chap. 5). The analysis predicted that the β exponent should be negatively correlated with sound intensity to be adapted to speech statistics. Cochlear frequency selectivity, in first

approximation, also follows a power law whose exponent decreases with sound intensity level as a result of cochlear compressive nonlinearities. Hence, the present study suggests a connection between nonlinear cochlear filtering and the fine-grained statistical structure of speech. Further analyses, along with new experimental investigations, will have to be carried out to determine whether the efficient coding hypothesis can be extended for peripheral auditory coding.

The recent success of machine learning algorithms for automatic speech recognition and other applications has reinforced the motivation of data-driven approaches for the study of sensory systems and perception. Statistical learning becomes an indispensable tool in the modeling of complex high-dimensional signals, including sensory signals that are at the center of computational neuroscience. The combination of information theory, statistical learning and signal processing makes it possible to address classic problems of speech modeling with a high level of abstraction, while previous attempts were based on explicit modeling. This thesis used this approach for two specific issues: the efficient coding of speech on short time scales and the perception of speech rhythm. Statistical analyses based on increasingly complex representation models provide an opportunity to expand our knowledge of how speech conveys information, and eventually answer Liberman's question – "what is specific to the speech code?".

Bibliography

[1] Abadi, M., and Others. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, mar 2016.

[2] Abbott, L. F., and Peter, D. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, nov 2001.

[3] Abdallah, S. A., and Plumbley, M. D. If the Independent Components of Natural Images are Edges, What are the Independent Components of Natural Sounds? *Proceedings of ICA2001* (2001), 534–539.

[4] Ablin, P., Cardoso, J.-F., Gramfort, A., and Cardoso, J.-F. Faster ICA under orthogonal constraint. Tech. rep., 2017.

[5] Amari, S., Chickocki, A., and Yang, H. A new learning algorithm for blind source separation. Tech. rep., 1996.

[6] Amodei, D., and Others. Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin. Tech. rep., 2016.

[7] Arnela, M., and Guasch, O. Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method. *The Journal of the Acoustical Society of America 133*, 6 (jun 2013), 4197–4209.

[8] Arnela, M., Guasch, O., and Alías, F. Effects of head geometry simplifications on acoustic radiation of vowel sounds based on time-domain finite-element simulations. *The Journal of the Acoustical Society of America 134*, 4 (oct 2013), 2946–2954.

[9] Arvaniti, A. Rhythm, Timing and the Timing of Rhythm. *Phonetica 66* (2009), 46–63.

[10] Atick, J. J. Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems 3*, 2 (1992), 213–251.

[11] Attias, H., and Schreiner, C. E. Coding of Naturalistic Stimuli by Auditory Midbrain Neurons 1 Natural Scene Statistics and the Neural Code. In *NIPS Proceedings, Advances in Neural Information Processing Systems 10* (1997).

[12] Attneave, F. Some informational aspects of visual perception. *Psychological Review 61*, 3 (1954), 183–193.

[13] Balazs, P., Dörfler, M., Jaillet, F., Holighaus, N., and Velasco, G. Theory, implementation and applications of nonstationary Gabor frames. *Journal of Computational and Applied Mathematics 236*, 6 (oct 2011), 1481–1496.

[14] Balazs, P., Dörfler, M., Kowalski, M., and Torrésani, B. Adapted and Adaptive Linear Time-Frequency Representations. *IEEE Signal Processing Magazine 30*, November (2013), 20–31.

[15] Ball, K. M., and Böröczky, K. J. Stability of the Prékopa-Leindler inequality. *Mathematika 56*, 2 (2010), 339–356.

[16] Barlow, H. Possible principles underlying the transformations of sensory messages. In *Sensory Communication*, W. A. Rosenblith, Ed. MIT Press, Cambridge, MA, 1961, pp. 217–234.

[17] Barlow, H. Finding minimum entropy codes. *Neural Computation 1*, 3 (1989), 412–423.

[18] Barlow, H. Redundancy reduction revisited. *Network: Computation in Neural Systems 12*, 3 (2001), 241–253.

[19] Barry, W., Andreeva, B., and Koreman, J. Do rhythm measures reflect perceived rhythm? *Phonetica 66*, 1-2 (2009), 78–94.

[20] Beaufays, F., Sak, H., and Senior, A. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling Has. *Interspeech*, September (2014), 338–342.

[21] Bell, A. J., and Sejnowski, T. J. An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation 7*, 6 (1995), 1129–1159.

[22] Bharadwaj, H. M., Verhulst, S., Shaheen, L., Liberman, M. C., and Shinn-Cunningham, B. G. Cochlear neuropathy and the coding of supra-threshold sound. *Frontiers in systems neuroscience 8* (2014), 26.

[23] Bonnasse-Gahot, L., and Nadal, J.-P. Neural coding of categories: information efficiency and optimal population codes. *Journal of Computational Neuroscience 25*, 1 (2008), 169–187.

[24] Boutillon, X. *Elements d'acoustique lineaire*. Ecole polytechnique, 2006.

[25] Boysson-Bardies, B. *Comment la parole vient aux enfants : de la naissance jusqu'à deux ans*. O. Jacob, 1996.

[26] Brascamp, H. J., and Lieb, E. H. Best constants in Young's inequality, its converse, and its generalization to more than three functions. *Advances in Mathematics 20*, 2 (1976), 151–173.

[27] Brette, R. Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences* (feb 2019), 1–44.

[28] Bruna, J., and Mallat, S. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*, 8 (2013), 1872–1886.

[29] Cardoso, J. F. Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters 4*, 4 (1997), 112–114.

[30] Carlson, N. L., Ming, V. L., and DeWeese, M. R. Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS Comput. Biol. 8*, 7 (2012), 1002594.

[31] Carney, L. H., and Yin, T. C. T. Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model. *Journal of neurophysiology 60*, 5 (nov 1988), 1653–1677.

[32] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *NIPS 2014 Workshop on Deep Learning* (2014).

[33] Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports 6*, 1 (sep 2016), 27755.

[34] Comon, P. Independent component analysis, A new concept? *Signal Processing 36*, 3 (1994), 287–314.

[35] Cover, T. M., and Thomas, J. A. *Elements of information theory*. Wiley-Interscience, 2006.

[36] Cummins, F., Gers, F., and Schmidhuber, J. Language Identification From Prosody Without Explicit Features. In *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary,* (1999).

[37] Daubechies, I., Roussos, E., Takerkart, S., Benharrosh, M., Golden, C., D'Ardenne, K., Richter, W., Cohen, J. D., and Haxby, J. Independent component analysis for brain fMRI does not select for independence. *Proceedings of the National Academy of Sciences of the United States of America 106*, 26 (jun 2009), 10415–22.

[38] Dauer, R. M. Stress-timing and Syllable-timing Reanalyzed. *Journal of Phonetics 11* (1983), 51–62.

[39] de Boer, E., and Nuttall, A. L. The mechanical waveform of the basilar membrane. III. Intensity effects. *The Journal of the Acoustical Society of America 107*, 3 (2002), 1497–1507.

[40] DeCasper, A. J., and Spence, M. J. Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior and Development 9*, 2 (1986), 133–150.

[41] Deller, J. R., Proakis, J. G., and Hansen, J. H. L. *Discrete-time processing of speech signals*. Macmillan Pub. Co., 1993.

[42] Donoho, D. L. Compressed Sensing. *IEEE Transactions on Information Theory 52*, 4 (2006), 1289.

[43] Donoho, D. L. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics 59*, 6 (jun 2006), 797–829.

[44] Dubuc, S. Critères de convexité et inégalités intégrales. *Annales de l'institut Fourier 27*, 1 (2011), 135–165.

[45] Eck, D., and Schmidhuber, J. Finding temporal structure in music: blues improvisation with LSTM recurrent networks. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing* (2002), IEEE, pp. 747–756.

[46] Eimas, P., Siqueland, E., Juscyk, P., and Vigorito, J. Speech perception in infants. *Science* (1971).

[47] Erra, R. G., and Gervain, J. The efficient coding of speech: Cross-linguistic differences. *PLoS ONE 11*, 2 (feb 2016), e0148861.

[48] Fant, G. Vocal tract wall effects, losses, and resonance bandwidths. *STL-QPSR 13* (1972), 28–52.

[49] Fant, G. The LF-model revisited. Transformations and frequency domain analysis. *STL-QPSR 36*, 2-3 (1995), 121–156.

[50] Fant, G., Kruckenberg, A., and Liljencrants, J. Acoustic-phonetic Analysis of Prominence in Swedish. In *Intonation*. Springer, Dordrecht, 2000, pp. 55–86.

[51] Fant, G., Liljencrants, J., and Lin, Q. A four-parameter model of glottal flow. *Stlqpsr 4* (1985), 1–13.

[52] Feichtinger, H. G., and Fornasier, M. Flexible Gabor-wavelet atomic decompositions for L2-Sobolev spaces. *Annali di Matematica Pura ed Applicata 185*, 1 (2006), 105–131.

[53] Fleischer, M., Pinkert, S., Mattheus, W., Mainka, A., and Mürbe, D. Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall. *Biomechanics and Modeling in Mechanobiology 14*, 4 (aug 2015), 719–733.

[54] Folland, G. B., and Sitaram, A. The uncertainty principle: A mathematical survey. *The Journal of Fourier Analysis and Applications 3*, 3 (may 1997), 207–238.

[55] Galves, A., Garcia, J., Duarte, D., and Galves, C. Sonority as a basis for rhythmic class discrimination. *Speech Prosody*, 2001 (2002), 323–326.

[56] Garofolo, J. S., Lamel, L. F., Fischer, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. *NIST* (1986), 1–94.

[57] Gazor, S., and Zhang, W. Speech probability distribution. *IEEE Signal Processing Letters 10*, 7 (jul 2003), 204–207.

[58] Gers, F. A., and Schmidhuber, J. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks 12*, 6 (2001), 1333–1340.

[59] Gers, F. A., Schraudolph, N. N., and Schmidhuber, J. Learning Precise Timing with LSTM Recurrent Networks. *Journal of Machine Learning Research 3* (2002), 115–143.

[60] Giraud, A.-L., and Poeppel, D. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience 15*, 4 (apr 2012), 511–517.

[61] Glorot, X., Bordes, A., and Bengio, Y. Deep Sparse Rectifier Neural Networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (Fort Lauderdale, FL, USA, 2011), G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15 of *Proceedings of Machine Learning Research*, PMLR, pp. 315–323.

[62] Gold, B., Morgan, N., and Ellis, D. *Speech and Audio Signal Processing*. John Wiley & Sons, Inc., Hoboken, NJ, USA, aug 2011.

[63] Gold, T. Hearing. II. The Physical Basis of the Action of the Cochlea. *Proceedings of the Royal Society of London. Series B - Biological Sciences 135*, 881 (dec 1948), 492–498.

[64] Grabe, E., and Low, E. L. Durational variability in speech and the Rhythm Class Hypothesis. In *Laboratory Phonology 7*. 2008, pp. 515–546.

[65] Graves, A., and Jaitly, N. Towards End-To-End Speech Recognition with Recurrent Neural Networks. *JMLR Workshop and Conference Proceedings 32*, 1 (2014), 1764–1772.

[66] Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., and Schmidhuber, J. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 5 (2009), 855–868.

[67] Gribonval, R., and Lesage, S. A survey of Sparse Component Analysis for blind source separation: principles, perspectives, and new challenges. Tech. rep., 2006.

[68] Grochenig, K. *Foundations of Time-Frequency Analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, Boston, MA, 2014.

[69] Hanna, N., Smith, J., and Wolfe, J. Frequencies, bandwidths and magnitudes of vocal tract and surrounding tissue resonances, measured through the lips during phonation. *The J. Acoust. Soc. Am. 139*, 5 (2016), 2924–2936.

[70] Hanson, H. M., and Chuang, E. S. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *The J. Acoust. Soc. Am. 106*, 2 (1999), 1064–1077.

[71] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer-Verlag New York, 2011.

[72] Heil, C. History and Evolution of the Density Theorem for Gabor Frames. *The Journal of Fourier Analysis and Applications 13*, 2 (2007).

[73] Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies. In *A Field Guide to Dynamical Recurrent Networks*. 2001.

[74] Hochreiter, S., and Schmidhuber, J. Long Short-Term Memory. *Neural Computation 9*, 8 (nov 1997), 1735–1780.

[75] Hsu, A., Woolley, S. M. N., Fremouw, T. E., and Theunissen, F. E. Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *The Journal of neuroscience : the official journal of the Society for Neuroscience 24*, 41 (oct 2004), 9201–11.

[76] Hudson, R. L. When is the wigner quasi-probability density non-negative? *Reports on Mathematical Physics 6*, 2 (1974), 249–252.

[77] Hyvärinen, A. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. Tech. Rep. 3, 1999.

[78] Hyvärinen, A. Survey on Independent Component Analysis. *Neural Comp. Surveys 2* (1999), 94–128.

[79] Hyvärinen, A., and Inki, M. Estimating overcomplete independent component bases for image windows. *Journal of Mathematical Imaging and Vision 17*, 2 (2002), 139–152.

[80] Irino, T., and Patterson, R. D. A time-domain, level-dependent auditory filter: The gammachirp. *The Journal of the Acoustical Society of America 101*, 1 (2002), 412–419.

[81] Irino, T., and Patterson, R. D. A dynamic compressive gammachirp auditory filterbank. *IEEE Transactions on Audio, Speech and Language Processing 14*, 6 (nov 2006), 2222–2232.

[82] Jaitly, N., and Hinton, G. Learning a better representation of speech soundwaves using restricted boltzmann machines. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (may 2011), IEEE, pp. 5884–5887.

[83] Johnson, K. *Acoustic and Auditory Phonetics*. Wiley, 2011.

[84] Jong-Hwan Lee, Ho-Young Jung, Te-Won Lee, and Soo-Young Lee. Speech feature extraction using independent component analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings* (2000), vol. 3, IEEE, pp. 1631–1634.

[85] Jutten, C., and Herault, J. Une solution neuromimétique au problème de séparation de sources. *Traitement du signal 5*, 6 (1988), 389–403.

[86] Kemp, D. T. Stimulated acoustic emissions from within the human auditory system. *The Journal of the Acoustical Society of America 64*, 5 (nov 1978), 1386–1391.

[87] Kohler, K. J. Rhythm in Speech and Language. *Phonetica 66*, 1-2 (2009), 29–45.

[88] Kuhl, P., Williams, K., Lacerda, F., Stevens, K., and Lindblom, B. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science 255*, 5044 (jan 1992), 606–608.

[89] Kuhl, P. K. Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development 6*, 2-3 (jan 1983), 263–285.

[90] Lachambre, H., Ricaud, B., Stempfel, G., Torresani, B., Wiesmeyr, C., and Onchis-Moaca, D. Optimal Window and Lattice in Gabor Transform. Application to Audio Analysis. In *Proceedings - 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2015* (2016), IEEE Comp Soc, pp. 109–112.

[91] Lee, J.-H., Lee, T.-W., Jung, H.-Y., and Lee, S.-Y. On the Efficient Speech Feature Extraction Based on Independent Component Analysis. *Neural Processing Letters* (2002), 235–245.

[92] Lesica, N. A., and Grothe, B. Efficient temporal processing of naturalistic sounds. *PLoS ONE 3*, 2 (feb 2008), e1655.

[93] Levy, W. B., and Baxter, R. A. Energy Efficient Neural Codes. *Neural Computation 8*, 3 (apr 1996), 531–543.

[94] Lewicki, M. S. Efficient coding of natural sounds. *Nature neuroscience 5*, 4 (2002), 356–363.

[95] Lewicki, M. S., and Olshausen, B. A. Probabilistic framework for the adaptation and comparison of image codes. *JOSA A 16*, 7 (1999), 1587–1601.

[96] Liberman, A. M. *Speech : a special code*. MIT Press, 1996.

[97] Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. Perception of the speech code. *Psychological review 74*, 6 (1967), 431.

[98] Lieb, E. H. Integral bounds for radar ambiguity functions and Wigner distributions. *Journal of Mathematical Physics* (1990), 625–630.

[99] Linsker, R. Self-organization in a perceptual network. *Computer 21*, 3 (1988), 105–117.

[100] Lopez-Poveda, E. A., and Meddis, R. A human nonlinear cochlear filterbank. *The Journal of the Acoustical Society of America 110*, 6 (2001), 3107–3118.

[101] Lyon, R. F. *Human and machine hearing: Extracting meaning from sound*. Cambridge University Press, may 2017.

[102] MacKay, D. J. C. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.

[103] Mairal, J., Bach, F., Ponce, J., Normale Supérieure, E., and Sapiro, G. Online Dictionary Learning for Sparse Coding. In *Proceedings of the 26th International Conference on Machine Learning* (2009).

[104] Mairano, P. *Rhythm typology: acoustic and perceptive studies.* PhD thesis, Universita degli studi di Torino, mar 2011.

[105] Mallat, S. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way.* Academic Press, 2008.

[106] Mallat, S., and Zhifeng Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing 41*, 12 (1993), 3397–3415.

[107] Manley, G. A. Comparative Auditory Neuroscience: Understanding the Evolution and Function of Ears. *Journal of the Association for Research in Otolaryngology : JARO 18*, 1 (feb 2017), 1–24.

[108] Marr, D. C., and Poggio, T. From Understanding Computation to Understanding Neural Circuitry. Tech. rep., Massachusetts Institute of Technology, may 1976.

[109] Mehler, J., Christophe, A., and Ramus, F. What we know about the initial state for language. *Image, Language, Brain: Papers from the First Mind-Brain Articulation Project Symposium*, 33 1 (2000), 51–75.

[110] Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., and Amiel-Tison, C. A precursor of language acquisition in young infants. *Cognition 29*, 2 (1988), 143–178.

[111] Miller, R. L., Schilling, J. R., Franck, K. R., and Young, E. D. Effects of acoustic trauma on the representation of the vowel /ϵ/ in cat auditory nerve fibers. *The Journal of the Acoustical Society of America 101*, 6 (1997), 3602–3616.

[112] Ming, V. L., and Holt, L. L. Efficient coding in human auditory perception. *The Journal of the Acoustical Society of America 126*, 3 (sep 2009), 1312–20.

[113] Młynarski, W., and McDermott, J. H. Learning Midlevel Auditory Codes from Natural Sound Statistics. *Neural Computation 30*, 3 (mar 2018), 631–669.

[114] Mullen, J. *Physical modelling of the vocal tract with the 2D digital waveguide mesh.* PhD thesis, University of York, 2006.

[115] Nadal, J.-P., and Parga, N. Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. *Network: Computation in neural systems 5*, 4 (1994), 565–581.

[116] Nadarajah, S. A generalized normal distribution. *Journal of Applied Statistics 32*, 7 (sep 2005), 685–694.

[117] Nair, V., and Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (2010), pp. 807–814.

[118] Nazzi, T., Bertoncini, J., and Mehler, J. Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of experimental psychology. Human perception and performance 24*, 3 (jun 1998), 756–66.

[119] Nolan, F., and Jeon, H.-S. Speech rhythm: a metaphor? *Philosophical Transactions of the Royal Society B: Biological Sciences 369*, 1658 (dec 2014), 20130396.

[120] Olshausen, B. A., and Field, D. Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature 381* (1996), 607–609.

[121] Olshausen, B. A., and Field, D. J. Sparse coding of sensory inputs. *Current Opinion in Neurobiology 14*, 4 (2004), 481–487.

[122] Olshausen, B. A., Sallee, P., and Lewicki, M. S. Learning Sparse Image Codes using a Wavelet Pyramid Architecture. In *Advances in Neural Information Processing Systems 13* (2001), pp. 887–893.

[123] Oxenham, A. J., and Shera, C. A. Estimates of human cochlear tuning at low levels using forward and simultaneous masking. *Journal of the Association for Research in Otolaryngology : JARO 4*, 4 (dec 2003), 541–54.

[124] Oxenham, A. J., and Simonson, A. M. Level dependence of auditory filters in nonsimultaneous masking as a function of frequency. *The Journal of the Acoustical Society of America 119*, 1 (jan 2006), 444–453.

[125] Pérez-González, D., and Malmierca, M. S. Adaptation in the auditory system: an overview. *Frontiers in Integrative Neuroscience 8* (2014).

[126] Polka, L., and Werker, J. F. Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance 20*, 2 (1994), 421–435.

[127] Pollard, T., Reimer, M., San, D., Mul, A., Thomas, M. G., Nowosad, J., Weissmann, D., and McDaniel, W. C. Template for writing a PhD thesis in Markdown, jul 2016.

[128] Prékopa, A. Logarithmic concave measures with application to stochastic programming. *Acta Scientiarum Mathematicarum 32* (1971), 301–316.

[129] Ramus, F. La discrimination des langues par la prosodie : Modélisation linguistique et études comportementales. *De la caractérisation à l'identification des langues, Actes de la 1ère Journée d'Étude sur l'identification automatique des langues* (1999), 186–201.

[130] Ramus, F. Acoustic correlates of linguistic rhythm: Perspectives. In *Proceedings of Speech Prosody 2002* (2002), pp. 115–120.

[131] Ramus, F., Nespor, M., and Mehler, J. Correlates of linguistic rhythm in the speech signal, 1999.

[132] Rhode, W. S. Observations of the Vibration of the Basilar Membrane in Squirrel Monkeys using the Mössbauer Technique. *Citation: The Journal of the Acoustical Society of America 49* (1971), 1218.

[133] Rhode, W. S., and Smith, P. H. Characteristics of tone-pip response patterns in relationship to spontaneous rate in cat auditory nerve fibers. *Hearing Research 18*, 2 (1985), 159–168.

[134] Ricaud, B., and Torresani, B. A survey of uncertainty principles and some signal processing applications. *Advances in Computational Mathematics 40*, 3 (2014), 629–650.

[135] Rieke, F., Bodnar, D. A., and Bialek, W. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society B: Biological Sciences 262*, 1365 (dec 1995), 259–265.

[136] Roach, P. On the distinction between 'stress-timed' and 'syllable-timed' languages. *Linguistics Controversies* (1982), 73–79.

[137] Rodriguez, F. A., Chen, C., Read, H. L., and Escabi, M. A. Neural Modulation Tuning Characteristics Scale to Efficiently Encode Natural Sound Statistics. *Journal of Neuroscience 30*, 47 (nov 2010), 15969–15980.

[138] Rouas, J. L., Farinas, J., Pellegrino, F., and André-Obrecht, R. Rhythmic unit extraction and modelling for automatic language identification. *Speech Communication 47*, 4 (2005), 436–456.

[139] Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., and Robles, L. Basilar-membrane responses to tones at the base of the chinchilla cochlea. *The Journal of the Acoustical Society of America 101*, 4 (apr 1997), 2151–2163.

[140] Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., and Robles, L. Basilar-membrane responses to clicks at the base of the chinchilla cochlea. *The Journal of the Acoustical Society of America 103*, 4 (apr 1998), 1972–89.

[141] Saremi, A., Beutelmann, R., Dietz, M., Ashida, G., Kretzberg, J., and Verhulst, S. A comparative study of seven human cochlear filter models. *The Journal of the Acoustical Society of America 140*, 3 (sep 2016), 1618–1634.

[142] Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal 27*, July 1928 (1948), 379–423.

[143] Shannon, C. E. Prediction and Entropy of Printed English. *Bell System Technical Journal 30*, 1 (jan 1951), 50–64.

[144] Shera, C. A., Guinan, J. J., and Oxenham, A. J. Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proceedings of the National Academy of Sciences 99*, 5 (mar 2002), 3318–3323.

[145] Simoncelli, E. P., and Olshausen, B. A. Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience 24*, 1 (2001), 1193–1216.

[146] Siqueland, E. R., and DeLucia, C. A. Visual reinforcement of nonnutritive sucking in human infants. *Science (New York, N.Y.) 165*, 3898 (sep 1969), 1144–6.

[147] Sjölander, K., and Beskow, J. WAVESURFER- An open source speech tool. In *Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing* (2000), pp. 464–467.

[148] Sluijter, A. M. C., and Van Heuven, V. J. Spectral balance as an acoustic correlate of linguistic stress. Tech. rep., 1996.

[149] Smith, E. C., and Lewicki, M. S. Efficient auditory coding. *Nature 439*, 7079 (2006), 978–82.

[150] Sondhi, M. Model for wave propagation in a lossy vocal tract. *The Journal of the Acoustical Society of America 55*, 5 (may 1974), 1070.

[151] Stevens, K. N. On the quantal nature of speech. *Journal ol Phonetics 17* (1989), 3–45.

[152] Stevens, K. N. *Acoustic phonetics*. MIT Press, 1998.

[153] Stilp, C. E., and Lewicki, M. S. Statistical structure of speech sound classes is congruent with cochlear nucleus response properties. In *Proceedings of Meetings on Acoustics 166ASA* (nov 2013), vol. 20, p. 050001.

[154] Stone, J. V. *Principles of neural information theory : computational neuroscience and metabolic efficiency.* Sebtel Press, 2018.

[155] Story, B. H., Titze, I. R., and Hoffman, E. A. Vocal tract area functions from magnetic resonance imaging. *The Journal of the Acoustical Society of America 100*, 1 (1996), 537–54.

[156] Strahl, S., and Mertins, A. Sparse gammatone signal model optimized for English speech does not match the human auditory filters. *Brain Research 1220* (2008), 224–233.

[157] Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to Sequence Learning with Neural Networks. In *NIPS* (2014), pp. 3104–3112.

[158] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

[159] Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method.

[160] Turk, A., and Shattuck-Hufnagel, S. What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. *Laboratory Phonology 4*, 1 (2013), 93–118.

[161] Tüske, Z., Golik, P., Schlüter, R., and Ney, H. Acoustic modeling with deep neural networks using raw time signal for LVCSR. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (2014), pp. 890–894.

[162] van der Maaten, L., and Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research 1* (2008), 1–48.

[163] van Hateren, J. H. A theory of maximizing sensory information. *Biological Cybernetics 68*, 1 (nov 1992), 23–29.

[164] van Hateren, J. H., and Ruderman, D. L. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London B: Biological Sciences 265*, 1412 (1998), 2315–2320.

[165] Verschooten, E., Desloovere, C., and Joris, P. X. High-resolution frequency tuning but not temporal coding in the human cochlea. *PLOS Biology 16*, 10 (oct 2018), e2005164.

[166] Verschooten, E., Robles, L., Kovačić, D., and Joris, P. X. Auditory nerve frequency tuning measured with forward-masked compound action potentials. *JARO - Journal of the Association for Research in Otolaryngology 13*, 6 (2012), 799–817.

[167] Vincent, E., Bertin, N., Gribonval, R., and Bimbot, F. From blind to guided audio source separation: How models and side information can improve the separation of sound. Tech. Rep. 3, 2014.

[168] Voss, R. F., and Clarke, J. "1/f noise" in music and speech. *Nature 258*, 5533 (1975), 317–318.

[169] Wattenberg, M., Viégas, F., and Johnson, I. How to Use t-SNE Effectively. *Distill 1*, 10 (oct 2016), e2.

[170] Yu, D., and Deng, L. *Automatic Speech Recognition : A Deep Learning Approach*. Signals and Communication Technology. Springer London, London, 2015.

[171] Zhang, X., Heinz, M. G., Bruce, I. C., and Carney, L. H. A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. *The Journal of the Acoustical Society of America 109*, 2 (feb 2001), 648–670.

[172] Zhang, Z., Xu, Y., Yang, J., Li, X., and Zhang, D. A Survey of Sparse Representation: Algorithms and Applications. *IEEE Access 3* (feb 2015), 490–530.

[173] Zilany, M. S. A., Bruce, I. C., and Carney, L. H. Updated parameters and expanded simulation options for a model of the auditory periphery. *The Journal of the Acoustical Society of America 135*, 1 (2014), 283–286.

[174] Zilany, M. S. A., Bruce, I. C., Nelson, P. C., and Carney, L. H. A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. *The Journal of the Acoustical Society of America 126*, 5 (2009), 2390–2412.

[175] Zilany, M. S. A., and Carney, L. H. Power-Law Dynamics in an Auditory-Nerve Model Can Account for Neural Adaptation to Sound-Level Statistics. *Journal of Neuroscience 30*, 31 (2010), 10380–10390.