# L'efficacité du système auditif humain pour la reconnaissance de sons naturels

Vincent Isnard

▶ **To cite this version:**

Vincent Isnard. L'efficacité du système auditif humain pour la reconnaissance de sons naturels. Sciences cognitives. Université Pierre et Marie Curie - Paris VI, 2016. Français. NNT : 2016PA066458 . tel-01444576v2

**HAL Id: tel-01444576**
**https://hal.science/tel-01444576v2**

Submitted on 12 Apr 2017

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

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

Université Paris 6 Pierre et Marie Curie
Ecole Doctorale n°158 : Cerveau, Cognition, Comportement

# L'efficacité du système auditif humain pour la reconnaissance de sons naturels

Vincent Isnard

Thèse de Doctorat de Sciences Cognitives

Présentée et soutenue publiquement le 25 novembre 2016

Devant le jury composé de :

| | | |
|---|---|---|
| Pascal BELIN | Professeur des Universités, Université d'Aix-Marseille | Rapporteur |
| Catherine SEMAL | Professeure des Universités, Institut National Polytechnique de Bordeaux | Rapporteure |
| Anne CACLIN | Chargée de Recherche, INSERM | Examinatrice |
| Bruno GAS | Professeur des Universités, Université Paris 6 | Examinateur |
| Christophe MICHEYL | Chercheur Principal, Starkey France | Examinateur |
| Isabelle VIAUD-DELMON | Directrice de Recherche, CNRS | Directrice |
| Clara SUIED | Chercheure, IRBA | Co-encadrante |

Institut de Recherche et Coordination Acoustique/Musique (IRCAM)
Laboratoire Sciences et Technologies de la Musique et du Son (STMS)
UMR 9912 IRCAM CNRS UPMC
Equipe Espaces Acoustiques et Cognitifs
1 place Igor Stravinsky, 75004 Paris, France

Institut de Recherche Biomédicale des Armées (IRBA)
Département Action et Cognition en Situation Opérationnelle (ACSO)
Unité Perception
1 place général Valérie André, 91220 Brétigny-sur-Orge, France

Ce travail de thèse a été financé par la Direction Générale de l'Armement (DGA).

# Résumé

Dans l'environnement sonore quotidien, les sons naturels sont en général facilement reconnaissables. Cette efficacité de la reconnaissance auditive peut être décrite et quantifiée suivant deux aspects différents : la quantité d'information nécessaire pour y parvenir et sa rapidité. L'objectif de cette thèse est d'évaluer expérimentalement ces deux aspects. Dans une première partie expérimentale, nous nous sommes intéressés à la quantité d'information en créant des représentations parcimonieuses de sons naturels originaux pour constituer ce qui est appelé des *esquisses auditives*. Nous avons montré qu'une esquisse auditive est reconnue malgré la quantité très limitée d'information auditive présente dans les stimuli. Pour parvenir à ces résultats, nous avons consacré une partie importante de notre travail à l'élaboration d'outils d'analyse adéquats en fonction des catégories sonores testées. Ainsi, pour l'analyse des stimuli auditifs, nous avons développé un modèle de distance auditive entre catégories sonores. Pour l'analyse des performances des participants, nous avons développé un modèle pour le calcul de la sensibilité par catégorie sonore et tenant compte du biais, qui s'intègre dans la théorie de détection du signal. Ces analyses nous ont permis de montrer qu'en réalité les résultats ne sont pas équivalents entre les différentes catégories sonores. En particulier, la voix se démarque des autres catégories testées (e.g. instruments de musique) : la technique de sélection de l'information parcimonieuse ne semble pas adaptée aux indices de la voix. Dans une seconde partie expérimentale, nous avons étudié le décours temporel de la reconnaissance auditive. Afin d'estimer le temps nécessaire au système auditif pour reconnaître un son, nous avons utilisé un récent paradigme de présentation audio séquentielle rapide (RASP, pour *Rapid Audio Sequential Presentation*). Nous avons montré que moins de 50 ms suffisent pour reconnaître un son naturel court, avec une meilleure reconnaissance pour la voix humaine. L'ensemble de nos résultats suggère un traitement efficace des sons naturels par le système auditif, et en particulier pour la voix humaine.

*Mots-clés :* reconnaissance auditive ; modélisation auditive ; sons naturels ; codage parcimonieux ; rapidité de traitement auditif ; timbre ; théorie de détection du signal.

# Abstract

In the daily soundscape, natural sounds are generally easy to recognize. Auditory recognition relies on two different aspects for such efficacy : the quantity of information necessary and the processing speed. The objective of this thesis was to experimentally evaluate these two aspects. In a first experimental part, we explored the amount of information by creating sparse representations of original natural sounds to form what is called *auditory sketches*. We showed that an auditory sketch is recognizable despite the very limited quantity of auditory information in the stimuli. To achieve these results, we dedicated an important part of our work on the elaboration of adequate tools in function of the tested sound categories. Thus, for the analysis of auditory stimuli, we have developed an auditory distance model between sound categories. For the analysis of the performances of the participants, we have developed a model to calculate the sensitivity by sound category and taking into account the bias, which falls within the signal detection theory. These analyses allowed us to show that, actually, the results are not equivalent between the different sound categories. Voices stand out from the other categories tested (e.g. musical instruments) : the technique of selection of the sparse information does not seem adapted to the voice features. In a second experimental part, we investigated the temporal course of auditory recognition. To estimate the time necessary for the auditory system to recognize a sound, we used a recent paradigm of *Rapid Audio Sequential Presentation* (RASP). We showed that less than 50 ms are enough to recognize a short natural sound, with a better recognition for the human voice. Altogether, our results suggest an efficient treatment of natural sounds by the auditory system, and in particular for the human voice.

*Keywords :* auditory recognition ; auditory modeling ; natural sounds ; sparse coding ; speed of auditory processing ; timbre ; signal detection theory.

# Remerciements

Je remercie chaleureusement ma directrice de thèse, Isabelle Viaud-Delmon, ainsi que ma co-encadrante, Clara Suied, pour leurs conseils et leur confiance durant ces trois années de thèse. J'ai pu bénéficier de votre complémentarité sur tous les points de la démarche scientifique théorique et expérimentale, et ce document de thèse n'en est qu'une mince illustration.

Merci aux membres de mon jury : à Pascal Belin et Catherine Semal pour avoir accepté d'être rapporteurs de cette thèse, et à Anne Caclin, Bruno Gas, et Christophe Micheyl pour avoir pris le temps d'évaluer ce travail.

Merci également à Corinne Roumes et Shihab Shamma pour leurs encouragements prodigués à différentes étapes de cette thèse.

J'ai eu la chance d'effectuer ma thèse dans deux instituts de recherche où la perception auditive est abordée suivant des thématiques variées et approfondies. Je remercie tout le personnel de l'IRCAM et de l'IRBA pour leur accueil et leur contribution à créer un environnement propice à la recherche, et pour tous les échanges enrichissants dont j'ai pu bénéficier au cours de cette thèse.

Merci aux copains et aux copines de Strasbourg, de Brest, de Paris, des labos, du conservatoire, pour toutes les aventures parisiennes pendant ces années de thèse.

Enfin, merci à ma famille pour leur soutien sans faille depuis le début du commencement.

# Table des matières

# Liste des figures

# Liste des tableaux

# Lexique

**m-AFC** : *m-Alternative Forced Choice* (choix forcé à m alternatives)
**AIM :** *Auditory Image Model* (modèle d'image auditive)
**ANOVA :** *ANalysis Of VAriance* (analyse de la variance)
**A1 :** cortex auditif primaire
**DTW :** *Dynamic Time Warping* (déformation temporelle dynamique)
**CGS :** Centre de Gravité Spectrale
**F0 :** fréquence fondamentale
**HF :** Haute-Fréquence
**HNR :** *Harmonic-to-Noise Ratio* (rapport harmonique-sur-bruit)
**IRMf :** Imagerie par Résonance Magnétique fonctionnelle
**ISI :** *Inter-Stimulus Interval* (intervalle inter-stimuli)
**MDS :** *MultiDimensional Scaling* (échelle multidimentionnelle)
**PE :** Potentiel Evoqué
**RASP :** *Rapid Audio Sequential Presentation* (présentation séquentielle audio rapide)
**RDM :** *Representational Dissimilarity Matrix* (matrice de dissemblance représentationnelle)
**RMS :** *Root Mean Square* (racine carrée de la moyenne des carrés)
**RSA :** *Representational Similarity Analysis* (analyse de similarité représentationnelle)
**RSVP :** *Rapid Serial Visual Presentation* (présentation visuelle sérielle rapide)
**SDM :** *Stimulus-feature Dissimilarity Matrices* (matrices de dissemblance de caractéristiques de stimuli)
**SNR :** *Signal-to-Noise Ratio* (rapport signal-sur-bruit)
**STEP :** *Spectro-Temporal Activity Pattern* (pattern d'activité spectro-temporelle)
**STRF :** *Spectro-Temporal Receptive Field* (champ récepteur spectro-temporel)
**TDS :** Théorie de la Détection du Signal
**TR :** Temps de Réaction

# Introduction générale

Lorsqu'on réalise le croquis d'un objet visuel, on reproduit certaines caractéristiques de l'objet original qui permettent de le reconnaître. Les caractéristiques sélectionnées suffisent à le reconnaître, tandis qu'une grande quantité d'information est supprimée. Peut-on en faire de même avec des sons ? Peut-on sélectionner et reproduire certaines caractéristiques sonores et reconnaître l'objet sonore original avec la même facilité et avec la même rapidité ? Des recherches en perception auditive chez l'humain montrent que l'on peut reconnaître un stimulus dont on ne conserve qu'une petite quantité d'information (e.g. un son tronqué temporellement ou fréquentiellement ; Gray, 1942 ; Remez et al., 1981). Cela signifie que toute l'information sonore originale n'est pas nécessaire à sa reconnaissance.

Le premier aspect de notre travail de recherche vise à déterminer s'il existe des indices acoustiques ou auditifs [1] à sélectionner en priorité pour favoriser la reconnaissance auditive de l'objet sonore original, et comment les sélectionner. Des études expérimentales et computationnelles montrent que les mécanismes perceptifs codent l'information de l'environnement naturel de façon parcimonieuse (e.g. Smith & Lewicki, 2006 ; Hromadka et al., 2008). Nous avons donc cherché à sélectionner cette information parcimonieuse dans des sons naturels et à la restituer sous forme d'*esquisses auditives*.

Le deuxième aspect de notre travail est complémentaire au premier et concerne le temps de reconnaissance auditive (Massaro, 1972a). Lorsqu'on écoute un son, il peut nous sembler qu'on le reconnaît immédiatement. Ce n'est pas à proprement parler le cas du fait de certaines constantes cérébrales, bien que les latences du traitement auditif soient très réduites (e.g. Liegeois-Chauvel et al., 1994). Nous avons cherché à évaluer l'influence de la quantité et de la nature de l'information sonore transmise sur la rapidité du traitement de reconnaissance auditive.

La partie théorique de cette thèse comprend trois sections : la première fait un état de l'art de la littérature sur le traitement auditif des sons naturels en s'intéressant aux premières étapes du traitement auditif humain et à leur modélisation,

1. Pour traduire le terme anglais *"features"* utilisé dans le contexte de la reconnaissance auditive, qu'on pourrait aussi traduire par "traits" pour poursuivre l'analogie avec la modalité visuelle.

aux enjeux de l'utilisation de sons naturels pour comprendre le fonctionnement efficace du traitement auditif, à la manière de caractériser les sons naturels par les corrélats acoustiques et auditifs du timbre, et enfin aux corrélats cérébraux de sons naturels observés par imagerie cérébrale. Suivent deux sections qui argumentent, d'après des résultats de la littérature, sur l'efficacité du traitement auditif de sons naturels selon les deux aspects que nous avons introduit plus haut : la parcimonie du traitement auditif et la rapidité de la reconnaissance auditive.

La partie expérimentale vise ensuite à tester ces deux aspects de l'efficacité du traitement auditif de sons naturels dans trois études qui portent d'abord sur la parcimonie auditive puis sur la rapidité de la reconnaissance auditive. La parcimonie auditive est étudiée à l'aide d'esquisses auditives, i.e. des sons très simplifiés qui ne doivent contenir que les indices parcimonieux conduisant à la reconnaissance auditive. Cette première étude a donné lieu à une publication (en anglais) qui est intégrée au manuscrit. La rapidité de la reconnaissance auditive est ensuite étudiée à l'aide d'un paradigme de présentation séquentielle rapide de sons naturels courts permettant d'évaluer le temps de traitement nécessaire pour la reconnaissance d'un son. Un article en préparation (en anglais) est également intégré dans ce manuscrit.

Dans leur globalité, nos résultats montrent que le système auditif est effectivement très performant, à la fois en termes de quantité d'information nécessaire pour reconnaître un son, et en termes de temps de traitement de sons contenant peu d'information. Toutefois, pour ces deux aspects du traitement auditif, des disparités sont apparues entre les différentes catégories sonores testées. Ces disparités ont pu être en grande partie saisies à l'aide d'un modèle de distances auditives, construit sur la base de représentations auditives temps-fréquence.

# I Contexte théorique

## 1 Traitement auditif des sons naturels

Le système auditif humain est capable d'isoler et d'identifier de façon robuste les indices acoustiques conduisant à la reconnaissance d'un très grand nombre de sources sonores. La description et la modélisation du système auditif permettent d'approcher le signal auditif traité par le cerveau pour effectuer ce type de tâches auditives. Ainsi, une représentation basée sur un modèle auditif met en avant l'information acoustique susceptible d'être utilisée pour la reconnaissance auditive. De telles représentations auditives seront reprises dans les parties expérimentales de cette thèse.

Dans cette section, nous verrons dans un premier temps que le système auditif est bien caractérisé au moins au niveau périphérique, et fait l'objet de modélisations performantes pour tenter d'expliquer plusieurs phénomènes perceptifs (e.g. sonie, masquage ; paragraphe I.1.1). On argumentera ensuite en faveur de l'utilisation de sons naturels dans le contexte expérimental, pour affiner la compréhension du système auditif (paragraphe I.1.2). Puis, on verra comment des sons naturels peuvent être caractérisés par le timbre, qui est l'une des premières approches importantes pour discerner les indices acoustiques utilisés en reconnaissance auditive (paragraphe I.1.3). Nous verrons enfin que l'encodage cérébral de l'information auditive peut dépasser le traitement des indices acoustiques avec des aires cérébrales spécialisées pour le processus d'abstraction de l'information auditive sous forme de catégories auditives (paragraphe I.1.4).

### 1.1 Description et modélisation du traitement auditif

De plus en plus de données de la physiologie auditive permettent de suivre le traitement du signal auditif à travers les voies auditives, et donc de repérer les indices utiles à sa reconnaissance. Les fonctions de chaque étape du système auditif sont aussi mieux identifiées. La compatibilité des résultats obtenus selon ces deux points de vue (physiologique et fonctionnel) permet de fiabiliser la modélisation auditive pour, par exemple, améliorer les performances d'analyse de signaux

acoustiques (e.g. en reconnaissance automatique de la parole ; Yang et al., 1992 ; Wang & Shamma, 1994).

### 1.1.1 Transduction de l'onde acoustique en activité cérébrale

Le système auditif périphérique humain transmet le signal acoustique à l'oreille interne via une transformation complexe, avant que le signal résultant ne soit encodé en potentiels d'action dans le nerf auditif. Les ondes acoustiques passent d'abord à travers l'oreille externe (pavillon de l'oreille, conduit auditif, et tympan) et moyenne (chaîne des osselets), avant d'entrer dans la cochlée, où l'onde se propage dans le fluide cochléaire et entraîne le déplacement de la membrane basilaire (Moore, 2012). L'onde se propage le long de la membrane basilaire en fonction de ses propriétés mécaniques : chaque point résonne avec un maximum d'amplitude pour une fréquence dite "caractéristique" (Lyon & Shamma, 1996). Les aigus sont répartis vers la base et les graves vers l'apex, suivant un axe tonotopique.

Puis, les cellules ciliées internes, réparties le long de la membrane basilaire, se déplacent avec celle-ci avec une amplitude qui est convertie en signal électrique par des neurones reliés à chaque cellule ciliée. Cette transduction du signal acoustique en signal électrique est donc effectuée en fonction de la fréquence et de l'amplitude grâce à la tonotopie cochléaire. Des cellules ciliées externes permettent également d'amplifier l'excitation de la membrane basilaire pour des niveaux faibles, par un mécanisme descendant, modifiant aussi le filtrage auditif en fonction du niveau sonore (Moore, 2012). Le signal électrique induit est ensuite transmis par le nerf auditif au cerveau via les voies auditives ascendantes, en passant par plusieurs relais auditifs successifs (noyau cochléaire, olive supérieure, colliculus inférieur, corps genouillé médian du thalamus ; cf. Figure 1) jusqu'au cortex auditif (temporal).

Le passage par ces relais auditifs affine la sélectivité fréquentielle grâce à des neurones accordés aux différentes fréquences (voir plus loin une illustration de courbe d'accord, cf. Figure 6). Des mécanismes d'inhibition progressifs, issus du système auditif central, sont responsables de cet affûtage de l'accord en fréquence (Suga et al., 1997 ; Zhang et al., 1997)[2]. Par ailleurs, les relations de voisinage

2. L'importance de l'inhibition latérale pour sculpter les courbes d'accord a notamment été montrée chez la chauve-souris pour le traitement des échos par Zhang et al. (1997), avec l'im-

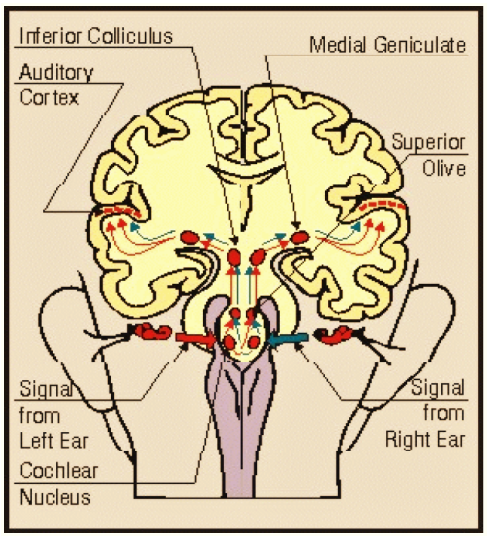


FIGURE 1 – Voies auditives primaires.

fréquentiel sont préservées à chaque étape du traitement auditif jusque dans la représentation tonotopique du cortex auditif primaire (A1), de façon analogue aux cartes topographiques d'autres systèmes sensoriels (rétinotopie dans le cortex visuel, somatotopie dans le cortex somatosensoriel ; Lyon & Shamma, 1996 ; Rauschecker, 1998 ; King & Nelken, 2009). Cependant, la structure précise du système auditif reste incertaine car elle présente davantage de relais avant d'atteindre le cortex, et donc de pré-traitements (King & Nelken, 2009). Ainsi, certaines caractéristiques auditives sont décodées tôt dans les voies auditives, comme l'intensité, la fréquence, ou les indices de localisation, tandis que c'est seulement au niveau du cortex auditif que se situent les fonctions plus complexes, en particulier de la discrimination auditive (e.g. Bathellier et al., 2012).

plication de projections corticofugales modulant l'activité de neurones primaires. Dans leur expérience, un anesthésiant était appliqué à des neurones corticaux accordés à environ 61 kHz, et n'avait pas d'effet sur des neurones accordés à d'autres fréquences (35 ou 84 kHz). Cette inactivation corticale diminuait la réponse des neurones sous-corticaux correspondants, sans décaler leur courbe de réponse, et augmentait à la fois la réponse d'autres neurones sous-corticaux (deux neurones thalamiques à 0.3 kHz en-dessous et 0.68 kHz au-dessus de la fréquence cible) tout en décalant leur fréquence d'accord vers celle des neurones inactivés. Ces résultats montrent que des neurones corticaux augmentent les réponses de neurones sous-corticaux correspondants, mais suppriment également celles d'autres neurones sous-corticaux tout en décalant leur courbe d'accord de celle des neurones cibles, améliorant ainsi le contraste de la représentation cérébrale de l'information auditive.

### 1.1.2 Modélisation du traitement auditif humain

La visualisation temps-fréquence de l'énergie sur un spectrogramme acoustique classique a été développée pour donner accès au contenu de la vibration de la source sonore qui n'est pas directement visible dans l'onde de pression. De la même façon, la modélisation du système auditif doit rendre compte de la transformation du signal acoustique en signal auditif, dont justement les premières étapes consistent en une analyse en fréquence similaire à une transformation de Fourier utilisée pour obtenir un spectrogramme acoustique (Moore, 2012). Il est donc courant de donner une approximation des transformations de l'oreille sur une représentation temps-fréquence, que l'on pourra directement comparer à des représentations acoustiques. Pour cela, les études peuvent reprendre des résultats issus de la psychoacoustique (e.g. détection de tons purs dans des bruits masquants) comme de la physiologie (e.g. réponses électrophysiologiques du nerf auditif), afin de vérifier que leur modélisation est en accord avec les phénomènes perceptifs et physiologiques (Lyon et al., 2010). On présente ici les principaux modèles auditifs existants dont certains seront utilisés dans la partie expérimentale.

#### Modélisation du système auditif périphérique [3].

**Périphérie et transduction auditives par les cellules ciliées.** Meddis & Hewitt (1991a) ont proposé un modèle du système auditif périphérique sur la base de résultats physiologiques. Ce modèle a notamment été appliqué par les auteurs pour expliquer les mécanismes d'identification de la hauteur et de sensibilité à la phase (Meddis & Hewitt, 1991a,b).

Le traitement auditif périphérique est décrit selon sept étapes principales (Figure 2) : (1) un filtrage passe-bande simulant le traitement de l'oreille externe, (2) une atténuation des basses et des hautes-fréquences de l'oreille moyenne, (3) un

3. Slaney & Lyon (1991) ont créé un "Musée du corrélogramme" sur internet, présentant plusieurs modèles de représentations temporelles de sons par le système auditif avec un grand nombre d'exemples : https ://ccrma.stanford.edu/~malcolm/correlograms/

Slaney (1998) a également proposé l'*Auditory Toolbox*, qui permet de réaliser en tout six représentations temps-fréquences de sons en incluant plusieurs modèles cochléaires, dont celui de Lyon & Mead (1988), celui de Patterson et al. (1995) combiné avec le modèle de cellule ciliée de Meddis & Hewitt (1991a), ou encore celui de Seneff (1988) utilisé en reconnaissance de la parole.

filtrage mécanique de la membrane basilaire, (4) la transduction du signal mécanique en signal cérébral par les cellules ciliées, (5) une inhibition de la décharge des fibres du nerf auditif, (6) une estimation de la distribution des intervalles des décharges dans des fibres d'un même canal, (7) l'addition de ces estimations à travers les canaux.

Le filtrage par la membrane basilaire est effectué à l'aide d'un banc de filtres gammatones. Ces filtres sont une bonne approximation des filtres roex (*rounded exponential*), qui permettent de suivre la réponse impulsionnelle de neurones auditifs primaires chez le chat, et qui sont toujours communément utilisés pour représenter les filtres auditifs humains (Patterson et al., 1995). La Figure 2a représente la sortie du banc de filtres gammatones pour un stimulus composé de dix harmoniques de 200 Hz de même amplitude.

La sortie de chaque filtre passe dans un simulateur de cellule ciliée. Ce dernier, réalisé par Meddis (1988), et dont une implémentation est présentée dans un autre article (Meddis et al., 1990), sera repris dans d'autres modèles ultérieurs (e.g. Patterson et al., 1995 ; Slaney, 1998). Il donne la probabilité d'occurrence d'un potentiel d'action d'après le mouvement de la membrane basilaire. La Figure 2b représente l'activité du banc de cellules ciliées par canal. On peut remarquer la rectification demi-onde et la compression des amplitudes importantes en comparaison aux amplitudes plus faibles, ce qui est caractéristique d'enregistrements du nerf auditif selon Meddis & Hewitt (1991a). Finalement, les intervalles temporels entre les potentiels d'action sont calculés séparément pour chaque canal, avec des histogrammes obtenus par autocorrélation (Figure 2c).

**Auditory Image Model (AIM).** Patterson et al. (1992) ont proposé un modèle fonctionnel de la cochlée simulant l'activité produite par des sons complexes dans le nerf auditif. Ce modèle permet de prendre en compte l'information temporelle fine jouant un rôle dans la perception de la parole notamment. Ce modèle a été repris et développé par Patterson et al. (1995) dans un programme modulaire modélisant les transformations successives opérées sur l'information auditive spectro-temporelle jusqu'aux décharges neuronales ordonnées tonotopiquement dans le nerf auditif. Des "images auditives" illustrent la réponse dynamique du système auditif à des sons naturels (Figure 3).

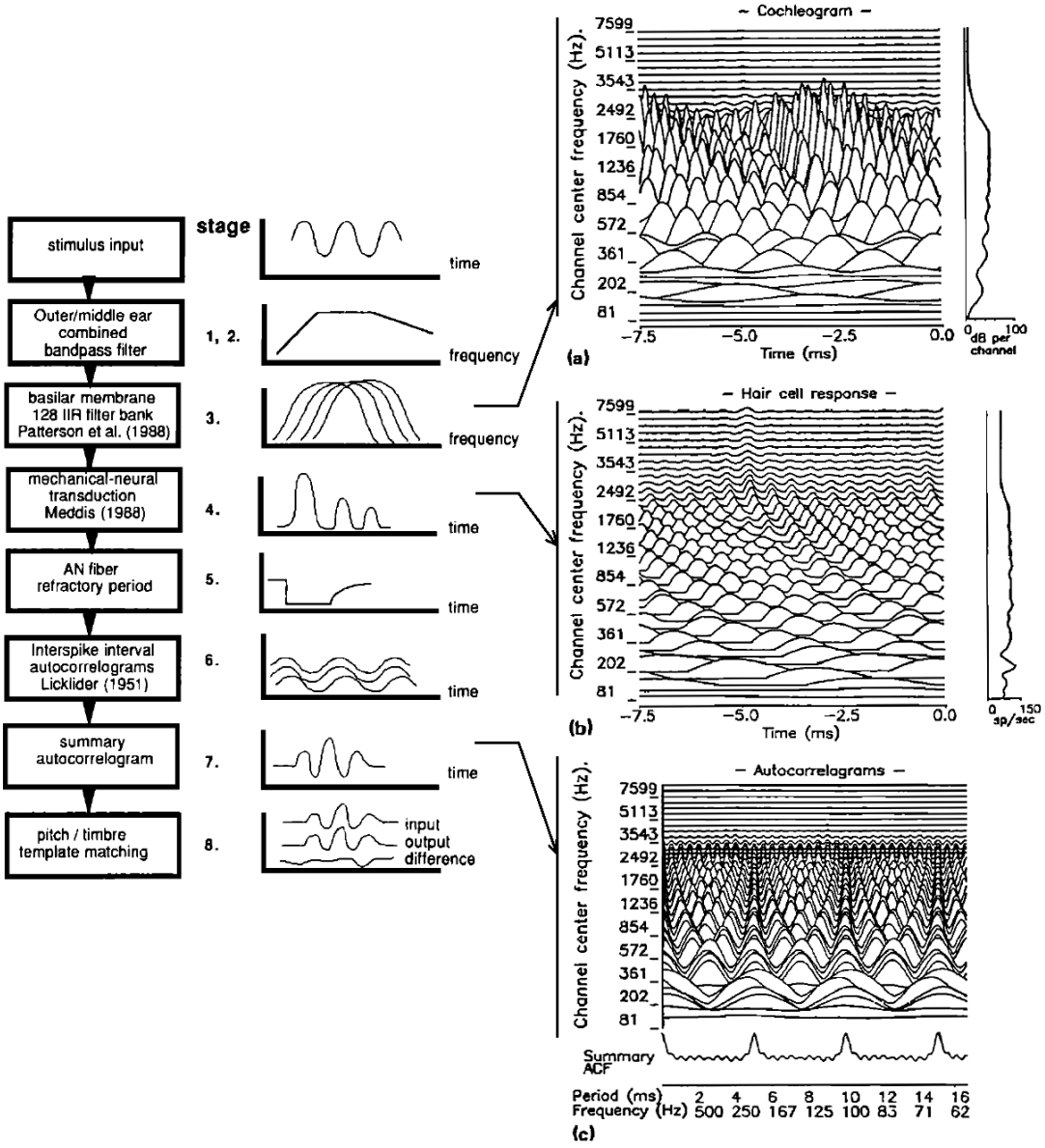


FIGURE 2 – Séquence de traitement du modèle auditif de Meddis & Hewitt (1991a). La sortie du modèle est représentée en réponse à un stimulus composé des dix premiers harmoniques de la fréquence 200 Hz : (a) cochléogramme, (b) réponse des cellules ciliées, (c) histogrammes individuels par intervalle de potentiels d'action. Figure adaptée de Meddis & Hewitt (1991).

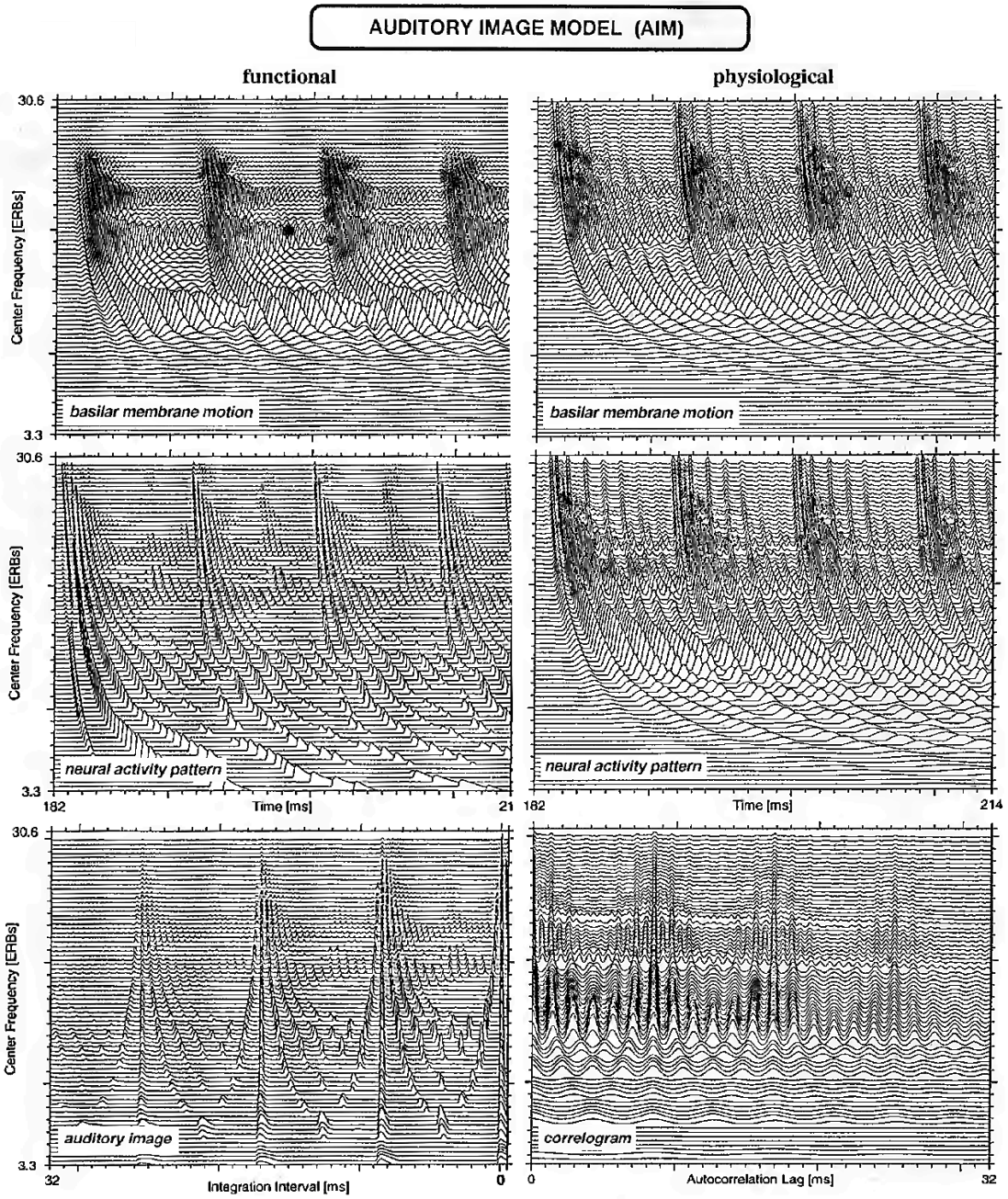


FIGURE 3 – La représentation des trois couches structurant le programme AIM. Colonne de gauche : voie fonctionnelle ; colonne de droite : voie physiologique. Le modèle comprend trois modules correspondant aux trois couches pour chacune des voies fonctionnelle et physiologique. La fonction de ces modules sont, de bas en haut : l'analyse spectrale, l'encodage cérébral, et la stabilisation par intervalle temporel. Figure adaptée de Patterson et al. (1995).

Le modèle d'image auditive (AIM, pour *Auditory Image Model*) de Patterson et al. (1995) est structuré en trois étapes, après un filtrage simulant celui de l'oreille moyenne, qui décrivent successivement l'analyse spectrale, l'encodage cérébral, et la stabilisation par intervalle temporel effectués par le système auditif (Figure 3 ; pour une implémentation, voir Bleeck et al., 2004). Deux points de vue sont comparés au cours de ces trois étapes : fonctionnel (indépendant du niveau sonore) et physiologique (dépendant du niveau sonore). La version physiologique inclut le modèle de cellule ciliée interne de Meddis (1986).

Lors de la première étape, l'analyse spectrale de la cochlée est implémentée par un banc de filtres de 75 canaux répartis entre 100 et 6000 Hz, linéaires gammatones dans le cas du modèle fonctionnel, et caractérisé par un filtrage non-linéaire, un affûtage spectral, et une compression dans le cas du modèle physiologique. Le décalage en phase des canaux basses-fréquences est causé par les filtres correspondants, plus étroits, et qui répondent plus lentement à l'entrée. A la deuxième étape, le pattern d'activité cérébrale est encodé par une rectification, une compression logarithmique, et un seuillage adaptatif en deux dimensions dans le modèle fonctionnel, et par une simulation de cellules ciliées internes dans le modèle physiologique. Enfin, lors de la troisième étape, l'image auditive du modèle fonctionnel est obtenue par intégration temporelle hachée, et le corrélogramme du modèle physiologique est obtenu par autocorrélation.

La version fonctionnelle de l'AIM a été utilisée pour modéliser la perception de la phase, de l'octave, ou du timbre, tandis que la version physiologique a été utilisée pour simuler des pertes auditives cochléaires (Patterson et al., 1995). Patterson (2000) donne d'autres exemples d'applications de la version fonctionnelle de l'AIM : pour des bruits, des transitoires, et des tons purs. Selon l'auteur, le pattern d'activité cérébrale est une représentation permettant de visualiser explicitement certains phénomènes de la perception des sons, en particulier avec l'étape d'intégration temporelle qui va permettre d'expliquer la perception fixe que produit un son stable dans le temps. Des temps d'intégration trop longs détruiraient la structure fine qu'on peut observer dans le nerf auditif, information par ailleurs nécessaire pour expliquer la perception de la hauteur, de certaines qualités sonores, et de la phase. Le problème de la stabilité est résolu avec l'intégration temporelle sur des fenêtres d'environ 30 ms, qui est une durée cohérente

avec le seuil inférieur de la perception de la hauteur chez l'homme (autour de 30 Hz ; cf. Patterson, 2000).

**Spectro-Temporal Excitation Pattern (STEP).** En partant du principe que des processus temporels semblent impliqués dans différents phénomènes perceptifs et que le système auditif est sensible aux changements temporels du spectre, Moore (2003) a cherché à prendre en compte ces effets temporels dans sa modélisation de la représentation interne de stimuli auditifs. Il calcule celle-ci sous forme de pattern d'excitation spectro-temporel (STEP, pour *Spectro-Temporal Excitation Pattern*). Selon l'auteur, un auditeur réalise une tâche auditive en comparant la représentation interne du stimulus, dont le STEP donne une approximation, avec une référence, par exemple pour la reconnaissance de phonèmes. L'objectif des STEPs est donc d'approcher ces représentations auditives internes en extrayant les parties utiles de l'information acoustique pour une tâche auditive donnée.

Le modèle est constitué de quatre étapes : (1) des filtres atténuent les fréquences inférieures à 500 Hz et supérieures à 5000 Hz, de façon similaire à la fonction de transfert globale entre les oreilles externe et moyenne ; (2) un banc de filtres passe-bandes avec des largeurs de bande dépendant de la fréquence simule les filtres cochléaires ; (3) chaque filtre auditif est suivi d'une compression non-linéaire dépendant du niveau sonore simulant la vibration de la membrane basilaire ; (4) un lissage est appliqué sur chaque sortie avec un filtre passe-bas pour refléter le processus central qui limite la résolution temporelle du système auditif. Ces quatre étapes successives permettent d'obtenir des STEPs plus proche de la représentation interne, en fonction de la fréquence, du temps, et de l'amplitude (Figure 4). De plus, cette représentation est facilement implémentable et a par exemple été utilisée pour simuler l'écart perceptif entre différents stimuli (Agus et al., 2012).

**Traitement auditif jusqu'au noyau cochléaire.** Le signal acoustique est temporel et les premières étapes du traitement auditif doivent donc traiter le flux d'information auditive dans la dimension temporelle. Cependant, on a vu aussi que le traitement auditif devient rapidement spatial : la tonotopie cochléaire répartit le traitement temporel par bandes fréquentielles. Les auteurs des modèles présen-

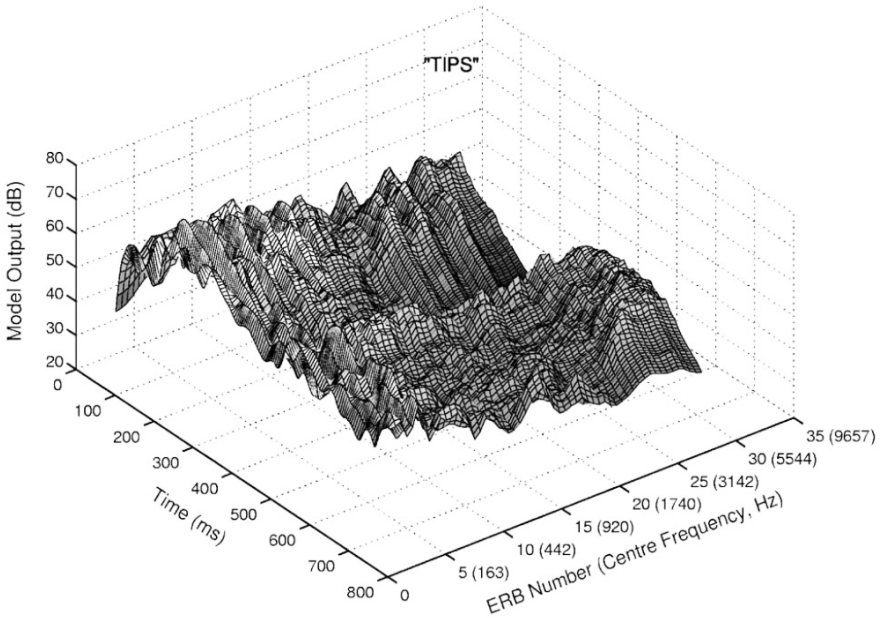


FIGURE 4 – STEP du mot "TIPS". Source : Moore (2003).

tés précédemment attribuaient un traitement séparé pour chaque sortie de filtre auditif, nécessitant des compromis d'intégration temporelle. Pourtant, on observe véritablement deux mécanismes d'encodage des fréquences distincts et complémentaires : temporel, en fonction des patterns temporels de décharges neuronales dans le nerf auditif qui se synchronisent par *phase locking* aux vibrations de la membrane basilaire jusqu'à des fréquences élevées (4 kHz ; Shamma, 2001), et spatial, avec des fréquences caractéristiques dépendant de la position de l'excitation le long de la membrane basilaire puis dans le nerf auditif.

De plus, les mécanismes d'inhibition latérale, dont l'existence est postulée dans le noyau cochléaire, reflètent la jonction des deux dimensions spatiale et temporelle, car ils jouent sur la synchronisation temporelle du contenu spectral. C'est pourquoi, plutôt que de modéliser des sorties de filtres cochléaires purement temporellement et indépendamment les unes des autres, Shamma (2001) a proposé une représentation en patterns spatio-temporels dans le nerf auditif, qui seraient en outre plus appropriés au traitement auditif des étapes centrales et mieux compatibles avec les distributions spatiales cérébrales dans les aires sensorielles primaires d'autres modalités (e.g. visuelle, somatosensorielle).

Dans des travaux antérieurs, Shamma (1985a,b) présente des représentations physiologiques du nerf auditif montrant le codage des paramètres du stimulus dans des profils temporels et spatiaux. L'auteur utilise des enregistrements de l'activité cérébrale dans de grandes populations de fibres de nerf auditif chez le chat en réponse à deux stimuli de voix et à un ton dans du bruit large bande. Les profils de décharges neuronales reflètent dans une certaine mesure le spectre du stimulus, avec par exemple des représentations adéquates de la position des formants. Cependant, les profils moyens obtenus ne donnent pas une bonne estimation de composantes spectrales individuelles qui pourraient être détectées perceptivement. La phase de l'activité synchrone dans le nerf auditif, en créant des patterns caractéristiques en fonction des régions cochléaires, serait donc nécessaire pour compléter la description des paramètres du stimulus (Shamma, 1985a). Enfin, dans son article complémentaire, Shamma (1985b) montre d'une part que les réseaux d'inhibition latérale sont biologiquement réalistes, et d'autre part que la synchronisation des décharges neuronales permet d'affiner la représentation du spectre, deux résultats qui ne sont pas vérifiés avec des algorithmes soit strictement spatiaux, soit strictement temporels.

Les modèles présentés précédemment n'appuient pas autant l'aspect spatio-temporel de la représentation auditive, notamment redevable aux mécanismes d'inhibition latérale. Jusque-là, leur importance avait été mentionnée au niveau cochléaire seulement (Lyon, 1982 ; Lyon & Mead, 1988). Le modèle plus récent de Chi et al. (2005) (voir aussi Yang et al., 1992 ; Wang & Shamma, 1994) synthétise les résultats de Shamma (1985a,b). Ce modèle est constitué d'une étape primaire d'analyse fréquentielle qui simule le traitement de la cochlée jusqu'au mésencéphale en transformant un signal acoustique en une représentation temps-fréquence auditive, et complété par une étape corticale reflétant l'analyse spectro-temporelle plus complexe de A1 (décrite dans le paragraphe suivant). Dans l'étape primaire du traitement, le signal acoustique est d'abord analysé par un banc de filtres cochléaires (cf. Figure 5). Ensuite, la sortie de chaque filtre est traitée par un modèle de cellule ciliée (filtre passe-haut, compression non-linéaire, filtre passe-bas pour la baisse de *phase locking* dans le nerf auditif au-delà de 2 kHz). Puis, un réseau d'inhibition latérale permet d'améliorer la sélectivité fréquentielle (dérivation suivie d'une rectification demi-onde). Le signal est finalement intégré sur de courtes

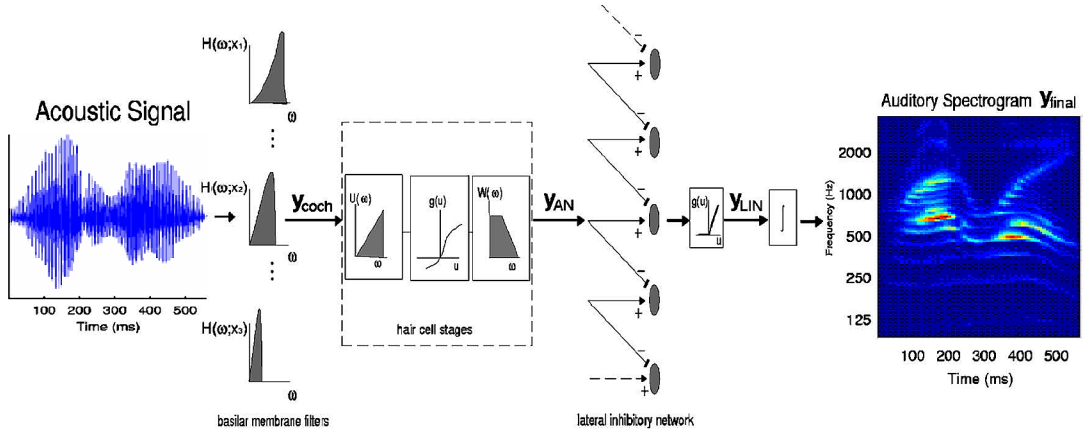


FIGURE 5 – Schéma des étapes auditives primaires du modèle de Chi et al. (2005). Le signal acoustique est analysé par un banc de filtres dont chaque sortie ($y_{coch}$) est traitée par un modèle de cellule ciliée ($y_{AN}$), suivie par un réseau d'inhibition latérale, et finalement rectifiée ($y_{LIN}$) et intégrée pour produire le spectrogramme auditif ($y_{final}$). Source : Chi et al. (2005).

fenêtres temporelles pour produire le spectrogramme auditif (Chi et al., 2005).

**Modélisation du système auditif central.** La représentation auditive au niveau du nerf auditif est bien établie par des résultats concordants en psychoacoustique et en physiologie. C'est beaucoup moins le cas au niveau cortical. Néanmoins, on peut mentionner le modèle complet de Chi et al. (2005), basé sur les connaissances récentes de l'organisation de ces structures centrales, et validé par des évaluations antérieures en perception de la parole ou pour expliquer la sensibilité à la phase par exemple. Deux observations physiologiques importantes sont prises en compte dans ce modèle : (1) la perte progressive des dynamiques temporelles de la périphérie vers le cortex, d'un *phase locking* rapide dans le nerf auditif à des taux de modulation beaucoup plus lents dans le cortex ; (2) l'émergence de la sélectivité dans les réponses cérébrales pour combiner des caractéristiques spectrales et temporelles, sélectivité plus complexe que les courbes d'accord et que les dynamiques des réponses des fibres du nerf auditif. Cette sélectivité, mesurée dans A1 avec des bruits de bandes de paramètres d'enveloppe différents (*ripples*), est décrite par les champs récepteurs spectro-temporels (STRFs ; Shamma, 2001 ; Nelken, 2004 ; Theunissen & Elie, 2014). Ceux-ci peuvent en retour modéliser

la réponse cérébrale, par convolution avec la description spectro-temporelle d'un son.

Concrètement, la modélisation de Chi et al. (2005) consiste à estimer le contenu de modulations spectrales et temporelles du spectrogramme auditif. Un banc de filtres sélectifs décompose les taux de modulations sur un intervalle temporel (taux lents à rapides) et spectral (échelles étroites à larges) pour obtenir une estimation des STRFs répartis sur l'axe tonotopique. Les auteurs apportent également les outils pour la reconstruction du signal original à partir de la transformation finale, sachant que les transformations non-linéaires ne permettent pas une reconstruction directe et parfaite mais seulement une approximation.

Ces représentations de l'encodage cortical sur différentes résolutions spectrales et temporelles seraient pertinentes pour le traitement flexible de sons naturels, et ont notamment permis de modéliser des données d'IRMf et cartographier des caractéristiques bas-niveaux (Santoro et al., 2014). De telles représentations, cohérentes avec celles d'autres aires sensorielles (e.g. visuelle, somatosensorielle), pourraient permettre de dériver les principes computationnels d'autres modalités, et par suite d'expliquer différents percepts auditifs (hauteur, localisation, etc.). Ainsi par exemple, l'analyse du profil spectro-temporel d'un son s'apparenterait à l'analyse de la forme visuelle (e.g. sélectivité à l'orientation et à la direction du mouvement), qui elle-même s'expliquerait par une analyse multi-résolution (cf. Shamma, 2001).

**Bilan sur les modèles auditifs.** On a présenté ici des modèles visant à reproduire le fonctionnement du système auditif humain, bien que d'autres modèles plus spécialisés existent, par exemple en traitement de la parole (e.g. Seneff, 1988) ou en sonie (e.g. Glasberg & Moore, 2002). En général, ces modèles ont été construits de façon indépendante par les différentes équipes de recherche. Malgré cela, on y retrouve les mêmes transformations auditives principales :

1. un filtrage passe-bande modélise la fonction de transfert de l'oreille externe et moyenne ;
2. la réponse tonotopique de la cochlée est modélisée sous la forme d'un banc de filtres passe-bandes, dont la largeur est relativement étroite en basses-fréquences puis augmente avec la fréquence ;

3. la transduction des cellules ciliées internes est modélisée par une rectification non-linéaire et une compression en fonction de l'intensité du stimulus, avec une réponse plutôt linéaire pour des niveaux intermédiaires, et une compression des niveaux forts, jusqu'à saturation au-delà d'un seuil ;
4. l'adaptation temporelle à court-terme doit permettre de rendre compte de la perception de la phase, notamment pour la localisation auditive, avec un taux de décharge neuronale dépendant de la fréquence (*phase locking*).

A ces quatre étapes, on pourrait ajouter celle d'inhibition latérale qui a été prise en compte plus récemment par Chi et al. (2005), en comparant les contenus des canaux fréquentiels pour améliorer le contraste fréquentiel.

Ces étapes successives du traitement auditif peuvent être implémentées sous la forme de modules compatibles entre eux et pouvant s'intégrer à d'autres modèles d'audition (e.g. Lyon, 1984 ; Yang et al., 1992 ; Wang & Shamma, 1994 ; Patterson et al., 1995 ; Moore, 2003 ; Chi et al., 2005 ; Simpson et al., 2013).

Prendre en compte les étapes de traitement de la cochlée, en passant par le nerf auditif, jusqu'au système central, ainsi que le niveau de détail de ces transformations permettra d'affiner la compréhension des phénomènes auditifs plus ou moins complexes (e.g. la sonie des sons non-stationnaires est un phénomène perceptif complexe en partie expliqué à l'aide des seuils d'audition dépendant de la fréquence, de la compression non-linéaire en fonction du niveau sonore, et du masquage fréquentiel ; Glasberg & Moore, 2002 ; Simpson et al., 2013). Les indices de reconnaissance auditive que nous tâchons d'identifier seront susceptibles d'être mis en évidence à l'aide de ces modèles, ou au minimum, ceux-ci auront permis d'écarter les indices non-pertinents, particulièrement denses dans des sons naturels. Néanmoins, nous allons voir maintenant que la complexité de l'information spectro-temporelle dans les sons naturels semble nécessaire à l'optimisation du traitement auditif humain.

## 1.2 Utiliser des sons naturels pour comprendre le traitement auditif

### 1.2.1 Sélectivité auditive à des caractéristiques acoustiques complexes

Le système auditif humain est capable de reconnaître tous types de sons, et en particulier des sons naturels complexes. Mais pendant longtemps, les chercheurs ont utilisé essentiellement des stimuli de synthèse pour évaluer les performances des mécanismes perceptifs. Les stimuli de synthèse présentent en effet le double avantage d'être parfaitement contrôlés et facilement manipulables. Ils isolent et reproduisent des caractéristiques de stimuli naturels, et peuvent de cette façon permettre d'expliquer les bases du traitement neuronal comme les champs récepteurs ou la sélectivité des neurones (e.g. Shamma et al., 1986 ; Suga et al., 1997 ; Zhang et al., 1997). En audition, des différences perceptives peuvent être interprétées facilement à partir des différences acoustiques entre des stimuli simples, comme par exemple des différences de sonie (e.g. Ponsot et al., 2015), de localisation (e.g. Hari, 1995), etc.

Cependant, le système auditif humain permet de reconnaître facilement des sons naturels complexes, et cette reconnaissance ne semble pas se baser uniquement sur des paramètres isolés tels que leur contenu fréquentiel, leur amplitude, ou leur durée. La variabilité de l'ensemble des propriétés acoustiques des sons naturels semble déterminante pour réaliser ce type de tâche auditive. En l'occurrence, le traitement auditif de ces propriétés acoustiques est non-linéaire et ces effets non-linéaires sont plus marqués avec des stimuli naturels (Theunissen et al., 2000 ; Schwartz & Simoncelli, 2001 ; Lewicki, 2002 ; O'Connor et al., 2005 ; Woolley et al., 2005). Par ailleurs, le traitement des sons naturels se distingue par une sélectivité croissante de caractéristiques acoustiques plus complexes le long des voies auditives, d'après des comparaisons de mesures de STRFs entre le colliculus inférieur et A1 (cf. Theunissen & Elie, 2014).

En somme, le fonctionnement du système auditif pour traiter des stimuli naturels complexes risque d'être mal prédit par de simples combinaisons linéaires des effets observés avec des stimuli de synthèse isolés, bien que la somme des descriptions acoustiques de stimuli simples puisse équivaloir aux descriptions acoustiques de stimuli complexes (voir aussi Simoncelli & Olshausen, 2001). Par exemple, les

études avec des tons purs ne peuvent pas expliquer comment la composition tonale des sons complexes est rassemblée, ni les performances cognitives comme en reconnaissance auditive (Rauschecker, 1998). De plus, même avec des modèles linéaires, les propriétés d'accords neuronaux sont mieux prédites à l'aide de stimuli naturels qu'avec des stimuli de synthèse (Santoro et al., 2014). C'est donc la généralisation des conclusions à des stimuli naturels à partir des résultats obtenus avec des stimuli de synthèse qui pourra poser problème.

Pour ces raisons, plusieurs auteurs promeuvent l'utilisation de stimuli naturels dans le contexte expérimental (e.g. Nelken et al., 1999 ; Kording et al., 2002 ; Felsen & Dan, 2005 ; Smith & Lewicki, 2006 ; King & Nelken, 2009 ; Giordano et al., 2013 ; Theunissen & Elie, 2014). Selon Smith & Lewicki (2006), l'efficacité du traitement cérébral issue de la phylogénie et de l'ontogénie est optimisée en premier lieu pour répondre à des stimuli naturels. Concernant la perception auditive, Nelken et al. (1999) affirment que la compréhension du système auditif doit passer par un élargissement des variétés de sons naturels employés expérimentalement, de sorte à être représentatifs de l'ensemble du biotope acoustique, et donc d'être cohérent avec l'objet étudié. Le risque, avec des stimuli naturels, reste cependant de ne pas pouvoir interpréter correctement les résultats obtenus, à cause justement de la complexité des stimuli et du lien avec celle des résultats expérimentaux.

### 1.2.2 Contrôle de stimuli naturels

**La complexité des stimuli naturels.** Dans leur revue de la littérature, Theunissen & Elie (2014) opposent l'approche de la neurophysiologie classique utilisant des sons de synthèse simples, et l'approche de l'éthologie auditive qui utilise historiquement des sons naturels pertinents pour décrire le comportement animal (Figure 6).

La première approche décrit les neurones auditifs périphériques par des courbes d'accord obtenues avec des tons purs, pourtant jamais entendus tels quels dans la nature. Par contraste, pour choisir et contrôler la variabilité de stimuli de synthèse plus complexes, Suga (1992) proposait d'étoffer cette démarche analytique : les faire varier en fréquence, en amplitude, et en temps, soit en synthétisant des éléments isolés supports d'information, soit en variant leurs paramètres dans le signal original par synthèse. Les paramètres restaient limités et déterminés d'avance,

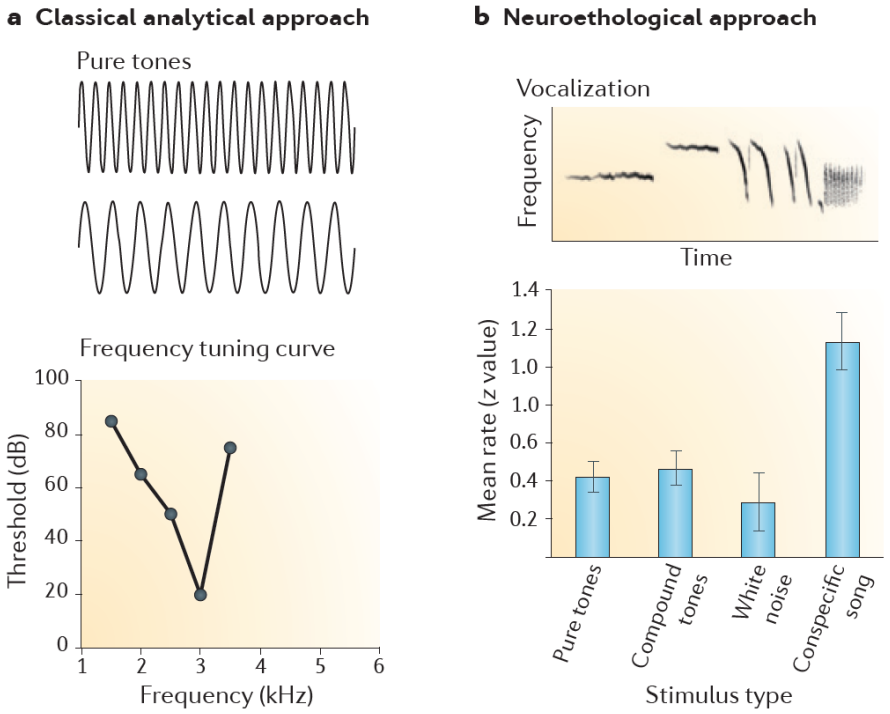


FIGURE 6 – Approches analytique et éthologique en neurosciences auditives. Ces deux approches sont basées sur l'analyse de réponses cérébrales à des sons produits par des sources sonores particulières. (a) Dans l'approche classique, les sources sonores sont souvent des synthétiseurs de tons purs (en haut), et les réponses cérébrales sont souvent décrites comme courbes d'accord en fréquence d'un neurone (en bas). La courbe d'accord en fréquence indique le niveau sonore minimum de tons purs nécessaire pour générer des réponses seuils (ici : courbe d'accord étroite d'un neurone aviaire dans le colliculus inférieur, accordé pour détecter une fréquence de 3 kHz, avec une augmentation brusque du seuil de réponse de chaque côté de cette fréquence). (b) Dans l'approche neuroéthologique, les sources sonores sont souvent des vocalisations animales ou d'autres sons de communication. Ces sons naturels sont des signaux complexes mieux représentés en temps-fréquence, comme ici le spectrogramme d'un son d'oiseau (en haut). Les réponses à ces sons complexes sont comparées aux réponses à des sons de synthèse (tons purs, combinaisons de tons purs, bruit blanc). Les données cérébrales présentées en bas proviennent de neurones isolés dans les aires auditives primaires aviaires. Les taux de décharge moyens de ces neurones montrent que le son naturel est le stimulus qui excite le plus les neurones mesurés. Figure adaptée de Theunissen & Elie (2014).

même si les techniques de synthèse sonore se sont diversifiées pour gagner en complexité et s'apparenter à des sons naturels. Malgré cela, l'utilisation de stimuli de synthèse reste controversée. En effet, elle implique d'introduire artificiellement un nombre limité de différences acoustiques et risque donc d'introduire un biais important dans l'expérience, c'est-à-dire expliquer les résultats uniquement par la différence introduite entre les stimuli sans envisager des combinaisons plus complexes avec d'autres composantes des stimuli.

Des chercheurs en éthologie auditive ont observé que des neurones répondent très fortement à des sons naturels et pas à leurs composantes séparées, ni à des stimuli de synthèse simples (e.g. bruit blanc, sinusoïdes ; cf. Theunissen & Elie, 2014). Cet accord neuronal résulte de la sélectivité à une information auditive spécifique dépendant du contexte naturel (e.g. un pattern spectro-temporel naturel caractéristique d'un cri d'alarme). L'enjeu, décrit par Theunissen & Elie (2014), consiste à améliorer les analyses des composantes statistiques des sons naturels, qui semblent induire ces réponses cérébrales optimisées. Les auteurs mentionnent également l'alternative de sons de synthèse créés avec des statistiques choisies dans des sons naturels, comme dans le cas des textures sonores (présenté plus bas, paragraphe I.2.1.2 ; McDermott & Simoncelli, 2011 ; McDermott et al., 2013). C'est vers ce type de procédures qu'on s'orientera dans le cadre de nos méthodes expérimentales.

**Maîtriser la complexité naturelle.** En réalité, à l'heure actuelle, l'utilisation de stimuli naturels peut être bien contrôlée. D'abord les banques de sons sont riches et variées (e.g. Goto et al., 2003) et des outils proposent un large panel d'analyses possibles, comme Praat (Boersma & Weenink, 2015) ou Straight (Kawahara & Matsui, 2003). Pour tester certaines caractéristiques spécifiques des sons naturels originaux, des chercheurs ont aussi créé une autre classe de sons conservant une partie de la complexité spectro-temporelle initiale : il s'agit des sons hybrides. Ils résultent de la combinaison de plusieurs sons, et présentent donc la caractéristique d'être ambigus d'un point de vue perceptif. Ainsi, dès les années 70, Grey & Gordon (1978) ont testé la perception du timbre de sons hybrides créés à partir de paires de sons dont les enveloppes spectrales avaient été interchangées (voir aussi Krumhansl, 1989). Plus récemment, Smith et al. (2002) décrivent dans

leur étude ce qu'ils ont appelé des "chimères auditives" comme étant la combinaison de l'enveloppe d'un son avec la structure fine d'un autre son. Leur procédé est presque équivalent à celui proposé par Grey & Gordon (1978), bien qu'utilisant des techniques de traitement du signal plus récentes. Les auteurs sont ainsi parvenus à montrer que la compréhension de la parole est d'abord basée sur l'information d'enveloppe, tandis que la structure fine permet de déterminer sa localisation, la séparation des deux types d'informations (le "quoi" et le "où") s'effectuant dans le cortex auditif (Rauschecker & Tian, 2000).

Agus et al. (2012) se sont également inspirés du procédé de Grey & Gordon (1978) pour créer des chimères auditives combinant des voix et des instruments. La mesure de temps de réaction permet de mettre en évidence que la reconnaissance rapide de la voix est due à une combinaison de ses caractéristiques spectrales et temporelles, en comparaison à des chimères ne retenant que les unes ou les autres de ces caractéristiques. Pour prendre un autre exemple, Suied et al. (2010) se sont servi de la dissociation enveloppe/structure fine pour montrer que la réaction rapide à des sons naturels de félins n'est pas due à un traitement sémantique mais strictement acoustique. En effet, les temps de réaction à des sons de félins sont équivalents à ceux obtenus avec des bruits modulés avec l'amplitude des sons naturels originaux.

Enfin, il est aussi possible, grâce à la technique de *morphing*, de combiner des stimuli pour doser dans différentes proportions des traits caractéristiques, voire de les exagérer. En vision, Leopold et al. (2006) ont réalisé des enregistrements neuronaux chez le singe en réponse à des interpolations de visages, depuis un visage moyen d'identité ambigüe jusqu'à des caricatures. L'augmentation de l'aspect caricaturé augmentait le taux de décharges neuronales, indiquant que le codage des caractéristiques du visage dépendait de traits saillants révélés par la technique de morphing. Cette méthode a été reprise avec succès en audition pour réaliser et tester des combinaisons d'émotions dans la voix (e.g. colère et peur ; Bestelmeyer et al., 2010, 2014), ou d'identités de locuteurs (Latinus et al., 2013).

L'ensemble de ces techniques atteste que la manipulation de la complexité des sons est devenue un enjeu prépondérant pour la compréhension du système auditif. L'étape suivante est de faire le lien entre cette complexité acoustique et la variété des jugements perceptifs à l'écoute de ces stimuli.

## 1.3 Caractériser des sons naturels et complexes par le timbre

Les études du timbre de sons complexes ont permis de cerner les patterns acoustiques spectro-temporels utiles à la reconnaissance auditive, et restent à l'heure actuelle un point de référence pour la compréhension de ces mécanismes. Le timbre s'applique en théorie seulement aux sons musicaux, c'est-à-dire avec une hauteur déterminée. Cependant, cette limitation permet justement de cerner des indices acoustiques fiables et interprétables pour une tâche de reconnaissance auditive. De plus, son formalisme est bien établi et a été repris dans de nombreuses études de la littérature sur le sujet, voire étendu à d'autres types de sons (environnementaux). Les indices de reconnaissance auditive identifiés grâce au timbre seront eux-mêmes susceptibles d'être transposés à d'autres types de sons.

### 1.3.1 Définir le timbre

Le timbre est l'attribut perceptif permettant de différencier deux sons ayant même durée perçue, hauteur, et niveau sonore perçu (ANSI, 1973). C'est de cette définition que sont parties la grande majorité des études qui ont cherché à décrire le timbre autrement que par la négative. En l'occurrence, les auteurs se sont accordés pour décrire le timbre comme un attribut perceptif multidimensionnel (cf. Wessel, 1973 ; Grey, 1977), c'est-à-dire un ensemble de qualités auditives, en-dehors de la durée perçue, de la hauteur, et du niveau sonore perçu, dont le nombre et les définitions ne sont pas fixés à l'avance. Ces qualités sont assimilées aux axes de l'espace de timbre multidimensionnel.

D'après une analyse du nombre de bandes critiques, Plomp (1970) estimait que le nombre de dimensions perceptives du timbre pouvait s'élever à 15, et qu'il dépend des stimuli utilisés. Les études sur le timbre ont depuis proposé des interprétations acoustiques et auditives des dimensions perceptives qui auraient motivé les jugements perceptifs, et se réfèrent à des qualités perçues telles que la brillance ou la rapidité d'attaque. Nous décrivons plus bas les méthodes conduisant à de telles descriptions du timbre.

Avant cela, il faut mentionner une autre définition du timbre plus flexible, celle de Handel & Erickson (2001). En effet, avec la définition de l'ANSI on peut remarquer que le timbre de chaque note d'un même instrument sera différent.

Pourtant, comme cela a été confirmé par la suite (Marozeau et al., 2003), des variations de timbre en fonction de la hauteur ne contredisent en général pas l'identité de timbre d'un instrument (bien que pour certains instruments comme la trompette, la qualité sonore puisse dépendre de la note jouée, de son intensité, ou de sa durée). C'est pourquoi Handel & Erickson (2001) ont proposé de considérer le timbre comme une qualité invariante basée sur des transformations à travers la hauteur et/ou l'intensité perçue permettant d'identifier un instrument ou une voix.

Bien entendu, des définitions plus larges du timbre ont aussi été employées dans un cadre compositionnel, avec notamment des timbres composés de mélanges de timbres, pour les articulations et fusions orchestrales par exemple (Boulez, 1987).

### 1.3.2 La méthode d'échelle multidimensionnelle

**Stimuli.** Historiquement, les premiers sons étudiés en acoustique étaient des sons d'instruments de musique (cf. Helmholtz, 1895). Les instruments de musique présentent l'avantage de pouvoir produire naturellement des sons contrôlés dans plusieurs dimensions, principalement en durée, hauteur, et niveau sonore. Ce n'est donc pas un hasard si la définition du timbre contraint les stimuli à être égalisés dans ces trois dimensions. Néanmoins, il est possible dans une certaine mesure de parler du timbre de sons de l'environnement en faisant la distinction entre la reconnaissance de la source sonore de celle de ses qualités auditives (Rasch & Plomp, 1982), cas que l'on évoquera plus bas (paragraphe I.1.3.5).

En fonction des études, les auteurs emploient des sons naturels (e.g. Wedin & Goude, 1972 ; Wessel, 1973 ; Elliott et al., 2013) ou de synthèse (e.g. Miller & Carterette, 1975 ; Grey, 1975, 1977 ; Grey & Gordon, 1978 ; Ehresman & Wessel, 1978 ; Wessel, 1979 ; Krumhansl, 1989 ; McAdams et al., 1995 ; Samson et al., 1997 ; Caclin et al., 2005), en veillant à ce que ces derniers soient équivalents à des sons naturels tout en bénéficiant de leur égalisation facilitée dans plusieurs dimensions acoustiques. A noter que certaines études partagent un même ensemble de stimuli (e.g. Krumhansl (1989), Krimphoff et al. (1994), et McAdams et al. (1995)). L'influence de l'ensemble de stimuli utilisés est discuté plus bas (paragraphe I.1.3.5).

**Procédure.** Plomp (1970) a proposé la méthode expérimentale permettant de représenter le timbre dans un espace euclidien sur la base de jugements perceptifs de dissemblances entre paires de sons, et qui a été majoritairement utilisée dans la littérature sur le timbre [4].

Plus précisément, avec cette méthode d'échelle multidimensionnelle (MDS, pour *Multidimensional Scaling*), la tâche des participants consiste à donner une valeur de dissemblance sur une échelle définie à l'avance, afin de décrire la distance perçue séparant des paires de sons [5]. Il est ensuite possible, à partir de ces mesures de distances, d'établir un espace multidimensionnel où chaque dimension correspond à une qualité perceptive plus ou moins importante dans chaque son. Cette méthode a été reprise par un grand nombre d'études sur le timbre car elle présente plusieurs avantages, comme relevés par McAdams et al. (1995) :

1. les jugements sont faciles à faire pour les participants ;
2. la technique ne génère pas d'hypothèses a priori sur la nature des dimensions ;
3. la représentation géométrique peut être visualisée sur un modèle spatial ;
4. le modèle spatial a un pouvoir prédictif.

On nuance toutefois ces différents points dans les analyses qui suivent : l'interprétation des dimensions perceptives est nécessairement biaisée a posteriori par l'expérimentateur ; la visualisation de l'espace de timbre vaut pour un petit nombre de dimensions ; enfin, le pouvoir prédictif du modèle spatial est limité d'après les études sur les clusters et les intervalles de timbres (paragraphe I.1.3.3).

**Calcul de l'espace perceptif multidimensionnel.** Trois modèles principaux permettent de générer des espaces multidimensionnels à partir de jugements per-

4. D'autres méthodes d'estimation du timbre existent, comme l'identification ou les descriptions verbales de timbres, cf. Caclin (2004) (voir aussi Gygi et al., 2007).

5. A noter que la méthode MDS a également été appliquée en vision dans études comportementales (e.g. Cutzu & Edelman, 1996) ou physiologiques (e.g. Young & Yamane, 1992 ; Rolls & Tovee, 1995). Cutzu & Edelman (1996) ont utilisé des stimuli visuels définis paramétriquement, et ont obtenu des représentations perceptives MDS proches de l'espace MDS paramétrique. De plus, un modèle computationnel était capable de simuler la représentation des similarités entre les objets, sans se baser sur la géométrie de chaque objet individuel. Selon les auteurs, ces résultats montrent que la nature des représentations internes relève de la similarité entre objets, accordés à une forme de référence, plutôt que d'un encodage de propriétés structurelles individuelles.

ceptifs de dissemblance entre paires de sons. On pourra se reporter à l'article de McAdams et al. (1995) pour une revue plus détaillée de ces modèles et des programmes permettant de les mettre en œuvre. Les algorithmes utilisent les jugements de dissemblance des participants pour générer une représentation géométrique dans un espace de faible dimensionnalité, généralement euclidien, afin de pouvoir interpréter a posteriori ces distances sur des bases acoustiques.

Le modèle original calcule la distance euclidienne entre tous les stimuli sur chaque dimension, dont le nombre est à déterminer en négligeant une certaine variabilité. Par la suite, Krumhansl (1989) a proposé un modèle étendu, initialement créé par Winsberg & Carroll (1989), répartissant les sons dans un espace de faible dimensionnalité mais tenant aussi compte de spécificités dans les sons. Ces spécificités sont des caractéristiques uniques, propres à certains instruments (e.g. l'attaque d'un clavecin), et qui prennent la forme d'une constante ajoutée pour chaque son dans le calcul de la distance.

McAdams et al. (1995) ont quant à eux utilisé la version étendue du modèle avec classes latentes proposé par Winsberg & De Soete (1993), qui permet de regrouper a posteriori les participants en sous-groupes, toujours sur la base de leurs jugements de dissemblance. En effet, la reconnaissance des timbres ne dépend pas uniquement des propriétés acoustiques des sons mais également de leur perception, propre à chaque participant. L'un des objectifs de McAdams et al. (1995) était de tester par ce biais l'impact de l'apprentissage musical sur les jugements de dissemblance, avec un grand nombre de participants répartis dans trois groupes suivant leur niveau musical (musiciens professionnels, musiciens amateurs, et non-musiciens). Le modèle étendu avec classes latentes pondère le calcul des distances entre les sons en incluant les spécificités. Cette pondération provient initialement du modèle de distance euclidien proposé par Carroll & Chang (1970), qui leur permettait une meilleure interprétation de l'espace multidimensionnel car les dimensions tiennent compte de la manière dont les participants donnent plus ou moins d'importance à telle ou telle dimension.

Dans le cas de l'étude de McAdams et al. (1995), le nombre de classes latentes était fixé à 5. Toutefois, chaque classe contenait finalement des participants de chacune de ces catégories, indiquant que chacune d'elle pouvait donner tous les types de pondérations révélés par la structure de classe. Il n'y avait donc pas de

division claire dans les jugements de dissemblance selon l'apprentissage musical des participants, même si la variance dans les résultats était plus grande pour les non-musiciens et les musiciens amateurs que pour les musiciens professionnels, indiquant une plus grande précision et une plus grande cohérence pour les musiciens professionnels. Similairement, des études ultérieures utilisant des classes latentes n'ont pas trouvé que l'apprentissage était un facteur déterminant dans l'élaboration de l'espace de timbre multidimensionnel (Caclin et al., 2005 ; Chon & McAdams, 2012).

### 1.3.3 Clusters et intervalles de timbres

On présente sur la Figure 7 un exemple typique de solution d'espace multidimensionnel. Dans ce type d'espaces, la répartition des sons perçus permet de révéler des relations entre eux potentiellement informatives sur la signification du timbre (McAdams et al., 1995). Qualitativement, on peut observer à la fois des regroupements de sons (clusters) et des espacements (intervalles) qui nous permettent de mieux saisir la notion de timbre à l'aide d'un espace multidimensionnel.

**Clusters de timbres.** Une représentation multidimensionnelle du timbre permet d'observer la présence de groupes de sons, ou clusters, dans lesquels les sons sont plus proches les uns des autres par rapport aux sons d'autres clusters (cf. Figure 7). Des auteurs ont fait le rapprochement entre les clusters et les familles instrumentales (e.g. Wessel, 1973 ; Grey, 1977 ; Grey & Gordon, 1978), bien que les auteurs relèvent généralement des exceptions à ces familles. D'autres auteurs remettent directement en cause de tels regroupements (Wedin & Goude, 1972). En effet, la comparaison du timbre de différents instruments n'est pas toujours pertinente, puisque ceux-ci peuvent présenter des timbres variables en fonction du registre, du niveau sonore, et des modes de jeu employés (cf. Iverson & Krumhansl, 1993).

Toutefois, le fait de constater que des instruments ont pu être séparés de leur famille originale laisse au moins penser que les jugements de dissemblance sont bien basés sur la perception de caractéristiques auditives, plutôt que sur leur reconnaissance catégorielle. Autrement dit, les participants ne cherchent pas à

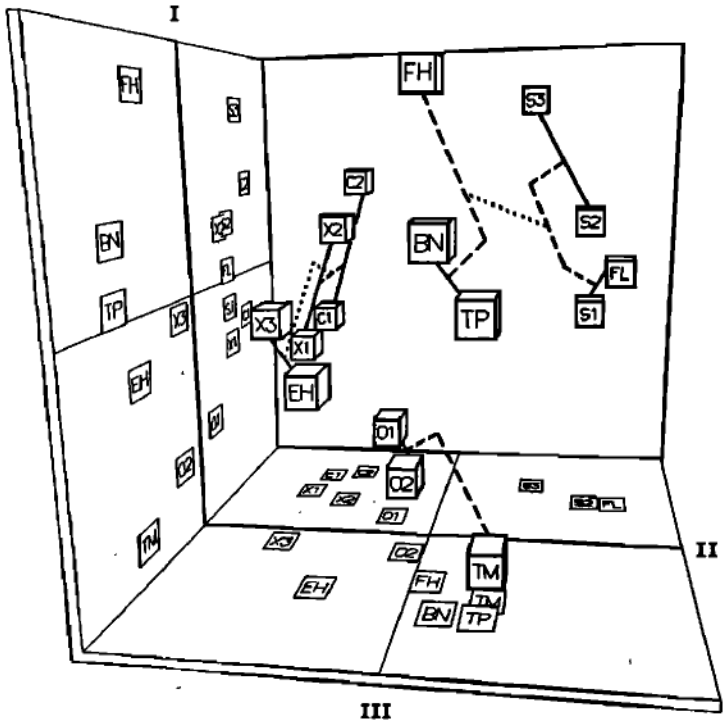


FIGURE 7 – Exemple de solution spatiale générée par un programme de MDS. Une analyse de clustering hiérarchique est représentée par des lignes connectant les instruments entre eux, dans l'ordre d'importance : lignes continues, tirets, pointillés. Des projections en deux dimensions apparaissent sur la face de gauche et du bas. Abréviations des instruments : O1, O2 = hautbois ; C1, C2 = clarinettes ; X1, X2, X3 = saxophones ; EH = cor anglais ; FH = cor ; S1, S2, S3 = violons ; TP = trompette ; TM = trombone ; FL = flûte ; BN = basson. Source : Grey (1977).

analyser et regrouper les sons suivant la facture instrumentale dont ils sont issus, mais effectuent bien un jugement de dissemblance uniquement sur la perception auditive du timbre.

**Intervalles de timbres.** En cherchant à tirer parti de la représentation multidimensionnelle et continue du timbre, Ehresman & Wessel (1978) ont mesuré la capacité des auditeurs à percevoir des relations abstraites d'analogies entre différents timbres. Il s'agissait d'évaluer si un intervalle de timbres peut être transposé, comme on transpose un intervalle de hauteurs par exemple. Krumhansl (1989) nuance cependant l'analogie entre des intervalles de hauteurs et de timbres, étant donné que le timbre est un attribut multidimensionnel, chaque dimension pouvant prendre plus ou moins d'importance de façon indépendante.

Pour réaliser cette évaluation, Ehresman & Wessel (1978) ont utilisé la représentation géométrique de l'espace de Grey (1977). Les participants effectuent une tâche d'analogie sur le modèle du parallélogramme : ils doivent déterminer le son D parmi {D1, D2, D3, D4} qui, par rapport au son C, donne un vecteur équivalent à un vecteur de référence AB (Figure 8). Les résultats ne se sont pas avérés très probants, sans doute car les calculs étaient effectués à partir de leur solution d'espace de timbre à deux dimensions d'une part, et que seule l'amplitude du vecteur était testée d'autre part, et pas sa direction. Cependant, comme l'ont noté McAdams & Cunible (1992), la notion d'intervalle de timbre était formalisée sous la forme de transpositions continues dans un espace perceptif multidimensionnel.

Par la suite, Wessel (1979) a montré dans un espace perceptif à deux dimensions (établi par l'auteur lui-même) que de telles propriétés géométriques de l'espace de timbre peuvent tout de même permettre de faire des prédictions sur les patterns de perception. Dans son expérience, les participants doivent ranger les quatre propositions {D1, D2, D3, D4}, couplées au son C, dans l'ordre du plus au moins équivalent par rapport à une référence AB (Figure 8). Les résultats des relations perceptives entre ces timbres montrent une tendance pour un classement cohérent avec les intervalles de timbres testés.

McAdams & Cunible (1992) ont également repris la notion d'intervalle de timbre pour la tester en amplitude et en direction dans l'espace obtenu par Krumhansl (1989). Ils présentent quatre séquences sonores correspondant à quatre com-

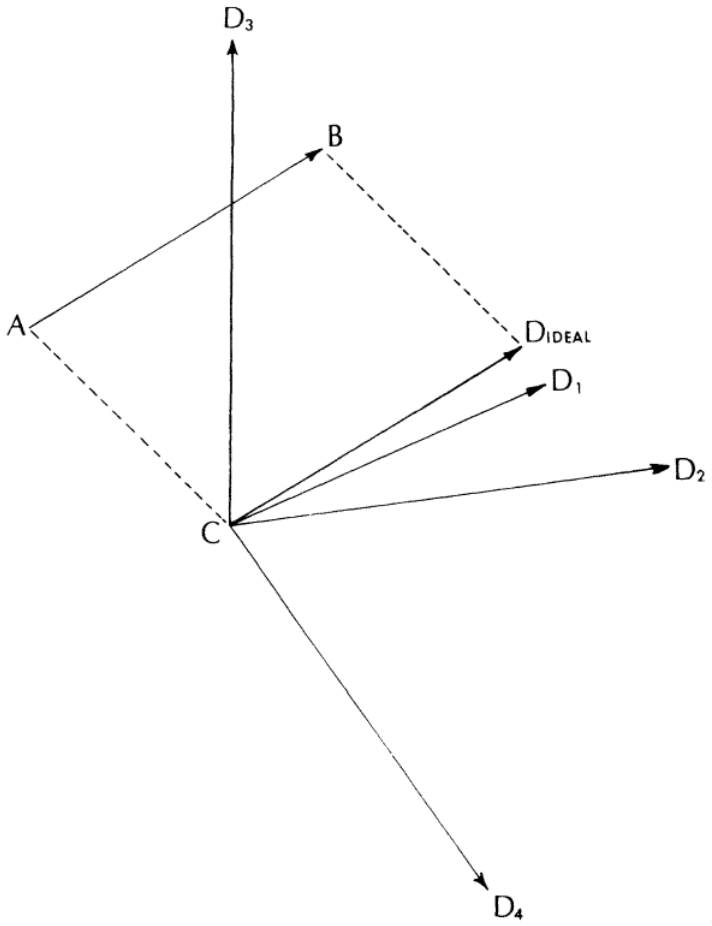


FIGURE 8 – Modèle du parallélogramme pour les analogies de timbres. Le vecteur AB représente un changement de timbre donné ; le vecteur CD représente l'analogie de timbre théorique, avec C donné. Le point D est la solution idéale. $D_1$, $D_2$, $D_3$, et $D_4$ sont les solutions proposées aux auditeurs. Source : Wessel (1979).

binaisons de normes et de directions de vecteurs de timbres, correctes ou fausses, en comparaison à un vecteur de timbre de référence (cf. Figure 8). La norme fausse correspond à une norme augmentée, tandis que la direction fausse correspond à la direction directement opposée à la direction correcte. Les résultats sont conformes à l'hypothèse faite sur le modèle du parallélogramme, avec une préférence pour une norme et une direction correctes. Cependant, cette préférence n'est pas plus forte que celle obtenue lorsque la norme ou la direction est fausse, sans prédominance de la norme ou de la direction. En général, les timbres proches du point idéal prédit par le modèle sont préférés de ceux situés à une plus grande distance. Néanmoins, ce modèle semble peu généralisable, d'autant moins pour des vecteurs de normes importantes, ou des petits vecteurs très espacés.

### 1.3.4 Interprétation des dimensions perceptives

**Corrélats acoustiques.** Plomp (1970) n'a pas essayé d'interpréter quantitativement les dimensions des espaces perceptifs qu'il a obtenues par des corrélats acoustiques, sauf par le calcul de distances entre des vecteurs d'énergie dans des bancs de filtres en tiers d'octave. La majorité des études suivantes, par contre, ont cherché à corréler ces dimensions perceptives avec des dimensions acoustiques, en se basant sur la simple observation des spectrogrammes (e.g. Grey, 1977), ou sur des calculs plus élaborés caractérisant les variations temporelles, spectrales, et spectro-temporelles des sons (e.g. Krimphoff et al., 1994). Avec cette méthode, les qualités sonores peuvent être interprétées grâce aux principaux paramètres acoustiques impliqués dans leur perception, et potentiellement généralisables à n'importe quel nouveau son.

On présente dans le Tableau 1 les résultats des analyses acoustiques du timbre issues d'une grande partie des études de la littérature sur le sujet, et employant la méthode MDS avec des sons d'instruments ou des tons complexes imitant des sons d'instruments. Les trois échelles temporelle, spectrale, et spectro-temporelle sont généralement représentées dans ces études, avec une ou plusieurs composantes pour chacune d'elles. On indique ici les définitions des corrélats acoustiques mentionnés dans le Tableau 1, bien qu'elles puissent varier en fonction des études et des observations qualitatives ou des opérations effectuées sur le signal :

– définitions temporelles :

  - temps d'attaque : intervalle temporel entre le début de la perception du son et la valeur maximale de l'enveloppe (Krimphoff et al., 1994) ;
  - enveloppe d'amplitude : de cor (avec une partie stationnaire à 40% du niveau maximum), de violon (avec une décroissance exponentielle), ou trapézoïdale (Miller & Carterette, 1975) ;
- définitions spectrales :
  - centre de gravité spectrale (CGS) : moyenne pondérée par les fréquences de l'amplitude de chacune d'elles (Krimphoff et al., 1994) ; Miller & Carterette (1975) font varier le nombre d'harmoniques par synthèse (3, 5, ou 7, avec une baisse de 3 dB à partir de la fréquence fondamentale), ce qui revient à décaler le CGS en fréquence ; de même pour Samson et al. (1997) (1, 4, ou 8 harmoniques de même amplitude) ;
  - structure fine du spectre (ou irrégularité spectrale) : différence entre l'amplitude des harmoniques et l'enveloppe globale ; ce corrélat spectral a été proposé par Krimphoff et al. (1994) en remplacement du corrélat spectro-temporel du flux spectral, proposée par Krumhansl (1989) ;
  - fréquence fondamentale (F0) : Miller & Carterette (1975) testent trois fréquences fondamentales de 200, 400, ou 800 Hz, à laquelle s'ajoutent 7 harmoniques, tandis que Marozeau et al. (2003) testent aussi l'influence de la fréquence fondamentale avec des sons séparés de 0, 2, ou 11 demi-tons ;
- définitions spectro-temporelles :
  - flux spectral : évolution temporelle des composantes spectrales (Krumhansl, 1989) ;
  - synchronisation des transitoires : ce corrélat est proche du flux spectral et dépend notamment du comportement des harmoniques élevés à l'attaque (Grey, 1977) ;
  - énergie haute-fréquence (HF) à l'attaque : comme le précédent, ce corrélat résulte de l'observation de patterns spectro-temporels différenciés, avec la présence ou non d'un burst d'énergie haute-fréquence à l'attaque et précédant le régime stationnaire harmonique (Grey, 1977).

Comme on peut l'observer dans le Tableau 1, le nombre de dimensions perceptives obtenues avec la méthode MDS varie suivant les études. La plupart des études obtiennent des espaces de timbre à deux ou trois dimensions. Tandis que selon

Elliott et al. (2013), cinq dimensions sont nécessaires et suffisantes.

Les définitions des corrélats acoustiques varient également suivant les études. Cependant, dès les premières études sur le timbre, certains ont été identifiés de façon récurrente et semblent faire consensus, en particulier le temps d'attaque dans la dimension temporelle et le CGS dans la dimension spectrale. A noter que d'autres études montrent la robustesse de ces deux corrélats chez des participants avec des implants cochléaires (e.g. Macherey & Delpierre, 2013). Ainsi, Macherey & Delpierre (2013) ont montré dans leur étude MDS que des instruments filtrés par un implant cochléaire peuvent être différenciés (et donc probablement identifiés) par des participants implantés. Et les deux principales dimensions de l'espace de timbre identifiées, pour les participants contrôles comme pour les participants implantés, sont le temps d'attaque et le CGS. Les performances réduites de ces derniers dans l'identification d'instruments pourraient être en fait dues à un manque d'apprentissage des qualités sonores (McDermott, 2004).

Dans la dimension spectro-temporelle par contre, les études utilisent des définitions plus variées. Mais toutes semblent rapporter l'évolution temporelle des composantes spectrales, avec des différences entre les sons plus marquées au niveau de l'attaque. McAdams et al. (1995) notent que les études semblent s'accorder sur une dimension correspondant au CGS, une autre correspondant soit à l'attaque temporelle du son, soit à l'enveloppe d'amplitude globale, et une troisième correspondant soit aux variations temporelles de l'enveloppe spectrale (propriété spectro-temporelle), soit à la structure fine du spectre (propriété spectrale). Enfin, selon Elliott et al. (2013), les caractéristiques acoustiques véritablement impliquées dans la perception du timbre relèveraient de représentations plus complexes (e.g. spectre de modulation spectro-temporelle) et donc plus difficilement interprétables que celles calculées sur des représentations simples comme l'enveloppe temporelle ou spectrale.

Ce qui ressort surtout de l'analyse qui suit, c'est la contribution de chacune des trois dimensions sonores dans la perception du timbre : temporelle, spectrale, et spectro-temporelle. En effet, ces trois dimensions semblent nécessaires pour caractériser les sons à partir de jugements de dissemblance. Le fait que la dimension spectro-temporelle puisse tenir compte de variations purement temporelle ou spectrale reste à discuter (Elliott et al., 2013).

A noter enfin qu'il existe des listes plus fournies respectivement de qualités sonores (voir par exemple le dictionnaire de 35 termes réalisé par Carron (2016)) et de descripteurs acoustiques (dont les applications peuvent être circonscrites à l'apprentissage automatique ; voir par exemple la liste de descripteurs de Peeters (2004) pouvant extraire jusqu'à 166 caractéristiques temporelles, énergétiques, spectrales, harmoniques, perceptives, et autres). Etant donnés les points de vue respectifs de la perception auditive et de l'apprentissage automatique, les liens entre qualités sonores et descripteurs acoustiques sont encore assez limités malgré quelques essais pour en établir (Marozeau et al., 2003 ; Elliott et al., 2013).

| | Dimensions | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Temporelle | | Spectrales | | | Spectro-temporelles | | | |
| | Temps d'attaque | Enveloppe d'ampl. | CGS | Structure fine | F0 | Flux spectral | Synchro. des transitoires | Energie HF à l'attaque |
| Plomp (1970) | | | | –[1] | | | | |
| Wedin & Goude (1972) | | | $\times\times\times^{2}$ | | | | | |
| Wessel (1973) | × | | × | | | | $\times^{3}$ | |
| Miller & Carterette (1975) | | × | × | | × | | | |
| Grey (1975) | × | | | | | | $\times^{4}$ | |
| Grey (1977) | | | × | | | | × | × |
| Grey & Gordon (1978) | | | × | | | | × | × |
| Ehresman & Wessel (1978) | | | × | | | | $\times^{5}$ | |
| Wessel (1979) | $\times^{6}$ | | × | | | | | |
| Rasch & Plomp (1982) | | | × | $\times^{7}$ | | | | |
| Krumhansl (1989) | × | | × | | | × | | |
| Iverson & Krumhansl (1993)[8] | | × | × | | | | | |
| Krimphoff et al. (1994)[9] | × | | × | × | | | | |
| McAdams et al. (1995) | × | | × | | | $\times^{10}$ | | |
| Samson et al. (1997)[11] | × | | × | | | | | |
| Kendall et al. (1999) | | | × | | | × | | |
| Marozeau et al. (2003) | × | | $\times\times^{12}$ | | $\times^{13}$ | | | |
| Caclin et al. (2005) | × | | × | × | | $\times^{14}$ | | |
| Elliott et al. (2013) | | | × | | | | $\times\times\times^{15}$ | |
| Total par sous-dimension | 9 | 2 | 20 | 3 | 2 | 4 | 5 | 2 |
| Total par dimension | 11 | | 25 | | | 15 | | |

Tableau 1 – Liste des corrélats acoustiques des dimensions perceptives du timbre proposés dans 19 études utilisant la méthode MDS avec des sons d'instruments naturels ou de synthèse.

Notes du Tableau 1 :

[1] Dans cette première étude utilisant les jugements perceptifs de dissemblance entre des sons d'instruments, l'auteur fait directement des corrélations entre des jugements de dissemblance et des différences de spectres calculées sur des bandes critiques, sans passer par un programme MDS (après un essai avec des tons complexes, cf. Plomp & Steeneken (1969) ; voir aussi Plomp (1975) pour le cas de tons complexes imitant des voyelles). Selon l'auteur, la perception du timbre est donc limitée au nombre de bandes critiques requises pour différencier les sons. Dans le cas de voix ou d'instruments, 3 dimensions semblent suffire.

[2] Répartition de l'énergie des harmoniques, élevée ou non, et comparée à celle de la fondamentale.

[3] Ce corrélat est combiné avec celui de la rapidité d'attaque dans une solution à deux dimensions.

[4] Ce corrélat correspond plus généralement à la synchronisation de l'amplitude des composantes spectrales. Cette étude est citée par Krumhansl (1989).

[5] La définition que donnent les auteurs de ce corrélat est assez proche de celle du flux spectral, cependant, ils précisent que la fluctuation spectrale peut être dans leur cas particulièrement importante au niveau de l'attaque. De plus, ils indiquent la similarité de leur solution avec celle de Grey (1975).

[6] L'auteur suggère cependant que la perception du temps d'attaque dépend de ses propriétés spectro-temporelles.

[7] Les données de cette étude sont reprises de celle de Plomp (1979). La première dimension de la solution à deux dimensions différencie les sons en fonction du nombre d'harmoniques aigus, ce qui revient, comme pour Miller & Carterette (1975) et Samson et al. (1997), à décaler le CGS en fréquence ; la seconde dimension est mal définie mais semble associée à des anomalies locales dans le spectre (Plomp, 1979, traduit du suédois avec Google Traduction), et semble donc proche de l'irrégularité spectrale définie par Krimphoff et al. (1994).

[8] Dans cette étude, les auteurs testent l'effet de la présence de l'attaque dans les jugements de dissemblance. On retient ici les analyses effectuées sur les sons complets. L'effet de la présence ou de l'absence de l'attaque est discuté plus bas.

[9] Dans cette étude, les auteurs proposent une nouvelle interprétation acoustique des dimensions perceptives de la solution de Krumhansl (1989), en transformant la dimension spectro-temporelle en dimension purement spectrale.

[10] Ce corrélat semble dépendre de la population de participants et/ou des stimuli.

[11] Dans cette étude, de façon similaire à Miller & Carterette (1975), les corrélats obtenus correspondent aux paramètres acoustiques manipulés par synthèse sonore, permettant simplement de confirmer leur effet sur les jugements de dissemblance.

[12] L'attaque, le CGS et l'étalement spectral (étalement du spectre autour de sa valeur moyenne) sont caractérisés de façon auditive.

[13] L'influence de la F0 est faible dans les cas de faibles écarts en hauteur tonale, tandis que les autres corrélats sont stables avec les changements de F0.

[14] L'importance perceptive de la structure fine du spectre ou du flux spectral dépend de leur saillance et du contexte, sachant qu'il s'agit de paramètres contrôlés dans des stimuli de synthèse.
[15] L'analyse d'Elliott et al. (2013) est basée sur des spectres de modulation spectro-temporelle (i.e. analyse spectrale en deux dimensions du spectrogramme acoustique), c'est pourquoi leurs résultats mettent en jeu cinq corrélats plus complexes que dans les études précédentes. Cependant, il est possible de répertorier ces corrélats de la façon suivante : un corrélat spectral indiquant un élargissement de largeur de bande avec un contenu haute-fréquence plus important, et quatre corrélats spectro-temporels (modulations temporelles rapides et lentes dépendant de la structure harmonique et non-harmonique, alternance d'harmoniques, et un dernier corrélat moins représenté et plus difficilement identifiable).

**Perception du temps d'attaque.** L'attaque est jugée importante dans la perception du timbre par certains auteurs (e.g. Krimphoff et al., 1994), ou au contraire moins significative par d'autres (Wedin & Goude, 1972 ; Kendall et al., 1999). Avant Krimphoff et al. (1994), les résultats d'Elliott (1975) (voir aussi Clark et al., 1963) montrent déjà une baisse significative des performances lorsque les attaques ainsi que la fin des sons sont supprimés. Wessel (1973) note quant à lui que couper les débuts et fins des sons réduit la dissemblance perçue entre des instruments à cordes et à vent, mais pas entre des cuivres.

Bien que Wedin & Goude (1972) relativisent le rôle de l'attaque dans la perception du timbre, leurs résultats montrent néanmoins que le degré d'identification correct des sons diminue drastiquement lorsque les transitoires d'attaque sont supprimés : quatre instruments seulement, sur les neuf testés, ne sont pas influencés en termes de pourcentage d'identification correct, lorsque l'attaque (premières 500 ms) et la décroissance (dernières 500 ms) sont supprimées, comparés aux sons complets. Le temps d'attaque n'a pas non plus été identifié comme un corrélat important dans l'étude de Kendall et al. (1999). Cependant, ces auteurs utilisent des sons continus comparés aux autres études avec des sons comprenant des impulsions ou des enveloppes temporelles plus marquées (Iverson & Krumhansl, 1993 ; McAdams et al., 1995).

Plus tard, Grey (1977) et Grey & Gordon (1978) ont mis en avant dans leur étude MDS deux corrélats spectro-temporels correspondant à la synchronisation des transitoires hautes-fréquences et à la présence de composantes hautes-fréquences à l'attaque. Ces corrélats ont toutefois été contestés dans l'étude d'Iverson & Krumhansl (1993) qui ont étudié spécifiquement la contribution de l'attaque dans la perception du timbre, et favorisant plutôt une dimension purement temporelle correspondant à l'enveloppe d'amplitude globale. Ainsi, les auteurs trouvent que l'un des deux axes perceptifs obtenus représente d'une part le caractère percussif (e.g. pour des cloches tubulaires) ou non (e.g. pour un saxophone) des sons. Aussi bien les jugements des sons avec que sans l'attaque ont contribué significativement au modèle perceptif. Afin d'identifier le corrélat acoustique pouvant correspondre à cette dimension, les auteurs ont calculé les enveloppes d'amplitude des sons. Cependant, les différences sont grandes entre les sons pour lesquels l'attaque a été supprimée et ceux pour lesquels elle a été conser-

vée. Les sons avec attaque sont alignés perceptivement en fonction de la rapidité de l'augmentation en amplitude, tandis que les sons sans attaque suivant la décroissance en amplitude. La contribution des différences générales des enveloppes d'amplitude sur les jugements de similarité est significative, et dans le cas des sons avec attaques, le temps d'attaque est identifié comme le facteur contribuant le plus à cette corrélation. Mais des sons avec des attaques similaires tendent à avoir des enveloppes d'amplitude similaires sur la globalité du son. Les auteurs en déduisent que les enveloppes d'amplitudes globales sont plus appropriées pour expliquer cette dimension perceptive.

Les résultats d'Iverson & Krumhansl (1993) confirment ceux de Miller & Carterette (1975), qui eux utilisent des enveloppes de synthèse en testant spécifiquement l'importance de l'attaque. En effet, Miller & Carterette (1975) présentent des sons avec ou sans l'attaque, ou encore présentent uniquement les attaques (premières 80 ms des sons). Les performances de reconnaissance des sons complets sont équivalentes à celles où seule l'attaque est présentée, montrant par-là l'importance perceptive de l'attaque pour les jugements de dissemblances par paires de sons. Toutefois, elles sont aussi équivalentes à celles où l'attaque est supprimée, montrant que l'attaque n'est pas seule en jeu dans la perception du timbre, mais que l'enveloppe temporelle y joue un rôle global.

Enfin, Suied et al. (2014) ont montré que des sons très courts peuvent être reconnus bien que l'attaque complète n'ait pas le temps d'apparaître. Les auteurs en ont déduit que l'attaque n'est pas toujours nécessaire pour la reconnaissance auditive, surtout dans le cas de sons stationnaires tels que des voyelles chantées. L'information spectro-temporelle présente initialement dans les sons originaux est réduite à de l'information spectrale dans les sons courts à cause de leur sélection. L'information spectrale seule semble donc pouvoir fournir les indices utilisés pour la reconnaissance du timbre dans le cas de sons courts (Suied et al., 2014).

Finalement, si la suppression de l'attaque des sons ne semble pas perturber la structure multidimensionnelle de l'espace de timbre (Miller & Carterette, 1975 ; Iverson & Krumhansl, 1993), mais que leur identification est compromise (Clark et al., 1963 ; Wedin & Goude, 1972 ; Elliott, 1975), il faut en conclure que l'attaque doit au minimum contenir des indices significatifs pour la reconnaissance des sons musicaux.

**Perception du CGS.** Le CGS a été identifié comme corrélat acoustique dans la très grande majorité des études de timbre, des plus anciennes (e.g. Wessel, 1973) jusqu'à très récemment (e.g. Elliott et al., 2013). Il s'agit d'un corrélat purement spectral bien défini mathématiquement, bien que certains auteurs aient testé des variantes comme l'étalement spectral (Marozeau et al., 2003). Selon Miller & Carterette (1975), la dimension spectrale est prédominante dans la perception du timbre, en particulier au travers du CGS. Même la récente étude d'Elliott et al. (2013), avec des techniques d'analyses basées sur des patterns de modulation spectro-temporelle, identifie le CGS comme corrélat important de la perception du timbre, aux côtés de quatre autres corrélats spectro-temporels. De plus, lorsqu'il est choisi comme corrélat acoustique, le CGS contribue toujours très fortement aux jugements perceptifs du timbre.

Le CGS est également bien identifié en tant que qualité sonore de la brillance. Iverson & Krumhansl (1993) différencient ainsi une brillance élevée (e.g. une trompette avec sourdine) ou non (e.g. un tuba), avec le CGS pour corrélat acoustique. Il est parfois plus difficile de qualifier ce qui est perçu dans le cas d'autres corrélats acoustiques plus complexes (e.g. Elliott et al., 2013).

**Perception du pattern spectro-temporel.** Les études ont plus de mal à faire consensus sur un corrélat spectro-temporel, bien que beaucoup d'entre elles en mentionnent au moins un. Ces corrélats peuvent mettre en jeu l'attaque du son, dans laquelle peuvent se concentrer des fluctuations spectrales importantes (Wessel, 1973 ; Ehresman & Wessel, 1978), voire remplacer le corrélat strictement temporel du temps d'attaque (Grey, 1977 ; Grey & Gordon, 1978). L'attaque n'est pas seule mise en cause puisque beaucoup d'études mentionnent aussi l'évolution temporelle des composantes spectrales comme corrélat acoustique (Krumhansl, 1989 ; McAdams et al., 1995 ; Kendall et al., 1999 ; Caclin et al., 2005). Cette fois, c'est donc avec un corrélat strictement spectral que pourrait se confondre le corrélat spectro-temporel initial.

Ce manque de consensus reflète une première difficulté qui est celle de la quantification du pattern spectro-temporel, et de la balance entre les deux dimensions temporelle et spectrale. Par exemple, deux études anciennes utilisant des stimuli naturels ne mentionnent aucun corrélat spectro-temporel (Wedin & Goude, 1972 ;

Wessel, 1973). Mais les auteurs se basent sur des observations qualitatives de spectrogrammes acoustiques, certainement insuffisantes pour expliquer la perception du timbre de sons naturels comportant des variations fines et complexes de modulations spectro-temporelles.

Dans leur étude, McAdams et al. (1995) ont réutilisé 18 des 21 sons de synthèse de Krumhansl (1989), soit 12 sons imitant des instruments et 6 hybrides. Les deux premières dimensions de leur espace de timbre 3D sont très corrélées avec celles nommées enveloppe temporelle et enveloppe spectrale, mais pas la troisième, nommée flux spectral, ce qui suggère des différences chez les participants ayant passé les expériences (9 participants dans l'étude de Krumhansl (1989) contre 88 dans l'étude de McAdams et al. (1995)). De plus, le modèle sélectionné initialement par les statistiques dans l'étude de McAdams et al. (1995) a six dimensions, et il s'avère que le flux spectral corrèle avec deux des six dimensions.

Caclin et al. (2005) remettent en cause, quant à eux, les contributions respectives du flux spectral et de l'atténuation d'harmoniques paires en fonction de leur saillance. Les auteurs notent que la perception du flux spectral a une influence limitée sur les jugements de dissemblance. Les auteurs soulèvent par la même occasion le problème des stimuli utilisés dans chaque étude, qui pourrait expliquer ces divergences entre les corrélats proposés. En l'occurrence, les stimuli instrumentaux classiquement utilisés peuvent présenter peu de variations spectro-temporelles sur des durées relativement courtes, et il est difficile de conclure quant à leur importance perceptive. Finalement, à la fois la nature des stimuli et la complexité de leur caractérisation, sans doute mieux appréhendée par des corrélats auditifs en particulier dans le cas de stimuli naturels (voir le paragraphe suivant), peuvent jouer sur l'interprétation des dimensions perceptives du timbre.

**Perception de spécificités.** Selon Krumhansl (1989), la définition classique du timbre conduit à exclure un grand nombre d'instruments et de modes de jeu possibles des tests perceptifs. En plus des propriétés émergentes, la représentation du timbre devrait tenir compte d'évènements sonores discrets et propres à certains instruments, appelés *spécificités*. Par exemple, Grey & Gordon (1978) notent que leur dimension perceptive liée à la présence d'inharmonicité haute-fréquence à l'attaque peut aussi s'expliquer par le caractère plus ou moins bruité de certaines

attaques. Selon les auteurs, certains instruments présentent ainsi une attaque légèrement grinçante, qui contraste avec l'attaque plus nette d'autres instruments, ou plus en basses-fréquences. La dureté de l'attaque peut aussi être caractérisée, dans une certaine mesure, en fonction de son pattern spectro-temporel. Ce type de saillances spectro-temporelles serait donc propre à certains instruments (e.g. l'attaque d'un clavecin). A nouveau, l'ensemble des stimuli utilisés pourrait influencer les interprétations acoustiques en fonction de leurs saillances timbrales. C'est ainsi qu'Iverson & Krumhansl (1993) expliquent leurs différences de résultats avec ceux de Grey (1975, 1977).

Les études de Krumhansl (1989) et de McAdams et al. (1995) ont utilisé un modèle plus performant pour ajuster les données, car pouvant tenir compte de la présence de spécificités pour générer leur espace multidimensionnel. Dans l'étude de Krumhansl (1989), les sons testés ont été créés par synthèse. Treize sons sont des reproductions de sons instrumentaux, tandis que les huit sons restants sont des hybrides, c'est-à-dire des combinaisons de deux sons produisant un son intermédiaire. Près des deux tiers des sons testés, comprenant des sons simples et aussi bien que des sons hybrides, ont une valeur de spécificité non-nulle, indiquant par-là la présence de caractéristiques uniques.

McAdams et al. (1995) ont d'abord évalué quel modèle permet de représenter au mieux les jugements de dissemblance obtenus avec des stimuli déjà utilisés dans l'étude de Krumhansl (1989). Ils ont choisi un modèle 3D avec spécificités, qui était proche du meilleur modèle 6D sans spécificités suggéré par l'algorithme, afin de pouvoir interpréter plus facilement les résultats. Dans leur analyse informelle des spécificités de valeurs importantes, les auteurs relèvent l'unicité des caractéristiques propres aux sons correspondants, et que la force perceptive de ces caractéristiques augmente avec la valeur de la spécificité. Les auteurs proposent deux sources à ces spécificités : des attributs continus (e.g. inharmonicité, grain) et des attributs discrets (e.g. présence d'un bruit sourd). Ces caractéristiques peuvent influencer le jugement de dissemblance par rapport à la globalité du son, faussant une représentation obtenue avec un modèle euclidien seul. En prenant en compte la présence de spécificités, le modèle s'accorde mieux aux données et il est ainsi rendu mieux interprétable en termes de corrélats acoustiques. En l'occurrence, on constate que les corrélats obtenus dans les études de Krumhansl (1989) et de

McAdams et al. (1995), tenant compte de la présence de spécificités, ont délaissé les deux dimensions spectro-temporelles de Grey (1977) et de Grey & Moorer (1977) concernant des fluctuations spectro-temporelles transitoires, comme à l'attaque, en proposant à la place le corrélat strictement temporel du temps d'attaque auquel s'ajoute des spécificités.

Caclin et al. (2005) notent quant à eux que l'influence (limitée) du flux spectral sur les jugements de dissemblance n'a pas de lien avec des spécificités dans les sons. En effet, dans la solution 2D avec spécificités, les sons avec de grandes valeurs de flux spectral n'ont pas de spécificités plus élevées. D'après les résultats précédents, cette remarque renforce l'intérêt de la caractérisation spectro-temporelle transitoire des spécificités, comparée au flux spectral qui concerne l'évolution globale des composantes spectrales.

**Corrélats auditifs.**

**Représentations temporelles et spectrales auditives.** La majorité des études sur le timbre utilisant la méthode MDS se sont attachées à trouver des corrélats strictement acoustiques aux dimensions perceptives qu'elles établissent d'après leurs résultats expérimentaux. Vue de cette façon, la méthode MDS permet de faire un lien direct entre les caractéristiques acoustiques des stimuli et leur impact perceptif. Cependant, comme on l'a présentée précédemment, la transformation qui a lieu depuis les stimuli physiques jusqu'à leur perception est complexe et non-linéaire. C'est au fond l'ensemble de ces transformations que l'on cherche à mettre à jour pour la compréhension du système auditif humain, et dont il faudrait rendre compte pour corréler plus fortement les variables acoustiques aux variables perceptives. Une partie de ces transformations ont été déjà bien explorées au moins jusqu'au nerf auditif. Quelques études ont intégré certaines modélisations auditives pour mieux expliciter les dimensions perceptives.

Déjà Plomp (1970), on l'a vu, proposait un calcul de distances entre des vecteurs d'énergie dans des bancs de filtre en tiers d'octave. Par la suite, Grey & Gordon (1978), en plus d'une recherche classique de corrélats acoustiques, ont proposé des modèles quantitatifs de la distribution de l'énergie spectrale de leur ensemble de stimuli. Les auteurs ont cherché à caractériser perceptivement et

de différentes manières le spectre physique des stimuli. D'abord, le spectre linéaire est dérivé de l'information spectro-temporelle utilisée pour la synthèse de leurs stimuli. Les niveaux des harmoniques sont caractérisés soit en prenant leur pic d'amplitude dans le temps, la moyenne temporelle de leur amplitude, ou la moyenne temporelle de leur énergie. Les auteurs décrivent ensuite les distributions d'énergie spectrale de quatre manières différentes : (1) en gardant l'unité du spectre linéaire, ou (2) en transformant l'échelle en décibels, ainsi qu'en introduisant des transformations perceptives avec le modèle de perception du niveau sonore de Zwicker & Scharf (1965), qui reprend des propriétés du système auditif telles que les bandes critiques et le masquage asymétrique des basses-fréquences vers les hautes-fréquences. A partir de ce modèle perceptif, ils dérivent des fonctions décrivant (3) le pattern d'excitation périphérique ainsi que (4) une fonction de niveau perçu. De cette façon, Grey & Gordon (1978) obtiennent des corrélations élevées en caractérisant numériquement ce pattern spectral par sa moyenne, et ce, non seulement avec leurs propres résultats, mais également avec ceux des études de Grey (1975, 1977), renforçant une interprétation auditive plutôt que strictement acoustique. Les corrélations sont plus élevées pour les caractérisations linéaires des spectres que celles en décibels ; elles sont plus élevées encore pour la fonction de niveau perçu que celles des caractérisations linéaires, ainsi que celles du pattern d'excitation, qui sont néanmoins déjà toutes élevées. Les meilleures corrélations sont obtenues lorsque le niveau des harmoniques a été calculé à partir de la moyenne temporelle de leur amplitude. Autrement dit, les transformations auditives apportées sur la dimension spectrale semblent garantir une interprétation de cette dimension plus représentative de la perception de l'auditeur.

Marozeau et al. (2003) ont aussi défini des mesures de CGS et d'étalement spectral calculées sur la base d'un spectrogramme auditif, obtenu en passant par un modèle de sonie partielle. Ils obtiennent de meilleures corrélations qu'avec la définition standard du CGS. En particulier, les corrélations entre les mesures de l'étalement spectral et les projections sur les dimensions perceptives concernées sont très fortes (autour de 0.9). De plus, en appliquant ce corrélat sur les données de McAdams et al. (1995), les auteurs trouvent une corrélation de 0.87, qui était bien supérieure à la valeur obtenue par les auteurs de la précédente étude (0.54) avec le flux spectral.

Concernant la dimension temporelle, Krimphoff et al. (1994) ont proposé d'utiliser le logarithme du temps d'attaque. Celui-ci commence avec le début de la perception du son et s'étend jusqu'au maximum d'amplitude de l'enveloppe temporelle. L'utilisation du logarithme permet de caractériser le fait que le début de perception du son ne correspond pas au seuil d'audition absolu mais dépend de l'amplitude maximale. L'utilisation du logarithme du temps d'attaque a été reprise notamment par McAdams et al. (1995), puis par Caclin et al. (2005) qui obtiennent également de meilleures corrélations en prenant le temps d'attaque logarithmique plutôt que le temps d'attaque linéaire. Plus que la perception du temps d'attaque, McAdams et al. (1995) ont cherché à caractériser celle du caractère impulsif des sons (voir aussi Iverson & Krumhansl, 1993). Avec la même idée, et plutôt que d'utiliser le logarithme du temps d'attaque, Marozeau et al. (2003) ont utilisé une mesure quantifiant l'impulsivité d'après la durée sur laquelle se concentre l'énergie. Les corrélations sont meilleures avec cette mesure qu'avec celle du logarithme du temps d'attaque, même si toutes étaient déjà supérieures à 0.9.

**Spectres de modulations spectro-temporelles.** Des études récentes sont venues changer la donne en proposant de nouvelles représentations auditives plus complexes des sons pour étudier la perception du timbre.

Patil et al. (2012) sont parvenus à classer automatiquement des sons d'instruments de musique, ainsi que reproduire des jugements de dissemblance humains, grâce à des enregistrements de neurones de A1 de mammifères dont les profils ont été analysés par un classificateur non-linéaire. Les réponses corticales auditives correspondent à des STRFs complexes, fournissant une représentation invariante des sons en fonction des hauteurs, des modes de jeu, des dynamiques, et des factures instrumentales. Les auteurs ont également modélisé la transformation du signal acoustique par les voies auditives jusqu'au traitement cortical, produisant un pattern d'activation spectro-temporel pour chaque instrument potentiellement utilisable pour le reconnaître. A nouveau, la classification automatique est très performante avec ce modèle. Enfin, ils comparent des espaces perceptifs multidimensionnels obtenus avec des données humaines et avec le modèle. L'espace obtenu avec les humains comporte deux dimensions, qui ont pour corrélats le temps d'at-

taque et le CGS. Le modèle cortical complet permet de décrire les performances humaines, mieux que le spectrogramme acoustique ou auditif (comprenant seulement les transformations de l'oreille interne).

L'étude d'Elliott et al. (2013) propose aussi une représentation auditive spectro-temporelle des sons, en se recentrant sur l'interprétation des dimensions perceptives du timbre. En effet, tandis que des détails de l'onde de pression ne sont pas perceptibles, les fluctuations d'intensités dans le spectre fréquentiel, représentées dans le spectre de modulation, contiennent de l'information sur les qualités du timbre. Les auteurs proposent quatre corrélats spectro-temporels sur la base d'une représentation auditive pertinente vis-à-vis du codage cérébral (Theunissen & Elie, 2014), leur permettant de réaffirmer le lien entre les dimensions temporelles et spectrales (par exemple, l'enveloppe des harmoniques peut jouer à la fois sur l'enveloppe strictement temporelle ou spectrale du fait du délai entre les harmoniques), même si certains corrélats ont des implications davantage temporelles ou spectrales.

### 1.3.5 Généralisation des indices du timbre

**Influence des stimuli sur l'espace perceptif multidimensionnel.** Les études employant la méthode MDS pour étudier le timbre sont contraintes de se limiter à des sons instrumentaux avec une hauteur définie. Pourtant, malgré un consensus assez marqué sur l'attaque et le CGS, des divergences subsistent concernant les corrélats acoustiques des dimensions perceptives. L'origine de ces divergences pourrait provenir du choix de l'ensemble des stimuli utilisés dans chaque étude, comme le font remarquer Iverson & Krumhansl (1993). Certaines saillances seraient favorisées par rapport à d'autres dans les attributs du timbre. Cette remarque est particulièrement valable dans le cas de stimuli de synthèse dans lesquels les composantes du timbre sont variées avec un nombre limité de valeurs (e.g. Miller & Carterette, 1975 ; Samson et al., 1997). Dans l'étude de Kendall et al. (1999), le troisième corrélat obtenu différencie justement les sons naturels des sons de synthèse, et dépend de la qualité de la synthèse.

Afin d'évaluer l'influence du choix des stimuli sur l'espace multidimensionnel, Caclin et al. (2005) ont proposé trois expériences, où pour chacune d'elles, les stimuli de synthèse sont contrôlés suivant des paramètres acoustiques qui varient

d'une expérience à l'autre. Dans la première expérience, les paramètres contrôlés sont le temps d'attaque, le CGS, et le flux spectral. Les analyses permettent de montrer l'importance du temps d'attaque et du CGS dans les jugements de dissemblance. En revanche, l'influence du flux spectral est limitée. Dans la deuxième expérience, les participants sont répartis en trois groupes correspondant à trois sous-ensembles de stimuli : (a) CGS constant, (b) temps d'attaque constant, (c) CGS et temps d'attaque constants. Pour le premier groupe, le temps d'attaque est identifié comme corrélat, ainsi que le flux spectral dans une moindre mesure. Pour le deuxième groupe, cette fois le CGS est identifié comme corrélat, et le flux spectral à nouveau dans une moindre mesure (avec un effet peut-être plus important lorsque le CGS est plus élevé en fréquence). Pour le troisième groupe, à nouveau les dissemblances ne sont pas bien prédites par le flux spectral, d'autant moins lorsqu'il prend des valeurs faibles dans les sons comparés. Les auteurs ont donc confirmé la significativité perceptive du temps d'attaque et du CGS avec des stimuli de synthèse contrôlés dans ces dimensions. Mais selon eux, le flux spectral peut aussi être pertinent dans certains contextes. Enfin, dans la troisième expérience, les auteurs ont testé s'il est possible de trouver une troisième dimension suffisamment saillante pour être utilisée quand le temps d'attaque et le CGS varient conjointement. Il s'agit cette fois de faire varier en plus l'irrégularité spectrale en atténuant les harmoniques paires relativement aux harmoniques impaires sur un intervalle de 0 à 8 dB. Comme dans les deux premières expériences, la significativité du temps d'attaque et du CGS en tant que corrélats est maintenue, avec en plus l'irrégularité spectrale comme paramètre saillant du timbre. En fin de compte, les timbres synthétiques utilisés par les auteurs leur ont permis de confirmer l'impact perceptif des dimensions acoustiques du temps d'attaque, du CGS, du flux spectral (plus limité), et de l'irrégularité spectrale.

L'étude de Caclin et al. (2005) a permis de montrer que la saillance perceptive du flux spectral décroît lorsque le nombre de paramètres concurrents augmente. Cette saillance perceptive dépend donc bien des stimuli utilisés, en fonction des indices disponibles dans les sons. De plus, des stimuli peuvent être jugés plus saillants que d'autres en fonction de la perception d'un indice prépondérant (e.g. l'irrégularité spectrale, cf. Chon & McAdams, 2012 ; Chon et al., 2013). Finalement, en fonction des paramètres concurrents dans un ensemble de stimuli donné,

certains de ces paramètres peuvent être mis en valeur ou au contraire négligés. Et ces paramètres concernaient jusqu'ici uniquement des sons instrumentaux, tandis que d'autres attributs du timbre pourraient être mis en évidence avec l'étude de sons non-instrumentaux. Selon Susini et al. (1999), la complexité de ce type de sons implique la nature multidimensionnelle de la perception qu'on en a, avec des corrélats acoustiques certainement différents de ceux obtenus pour des sons instrumentaux.

**Timbre de sons environnementaux.** Ballas (1993) a montré que des facteurs à la fois acoustiques, perceptifs, et cognitifs sont impliqués dans l'identification de sons environnementaux très divers. En particulier, leur temps d'identification résulterait à la fois d'une forme de choix entre des alternatives connues (dont le nombre et la familiarité déterminent l'incertitude sur la cause ayant produit le son), et d'une analyse étendue de l'information acoustique. Pour tous les sons testés, quelques secondes suffisent à les identifier correctement bien que des disparités apparaissent en fonction du type de sons. Ces disparités semblent davantage liées à des timbres stéréotypés qu'à la fréquence de l'exposition écologique. Par exemple, bien que des sons d'impacts soient plus fréquents que des signaux d'alertes (e.g. klaxons ou sonnettes), ils suscitent plus d'incertitude sur la cause de l'évènement sonore, et donc un temps d'identification plus long.

Gygi et al. (2007) sont arrivés à une conclusion similaire soulignant l'implication de deux types d'écoute, taxonomique [6] et qualitative. Les participants jugent la similarité entre des sons, mais également entre les labels des sons. En comparant les données respectives, les auteurs ont observé que pour effectuer leurs jugements de dissemblance, les participants examinent à la fois les propriétés acoustiques qui informent sur la source (qu'ils avaient mémorisées et stéréotypées dans le cas où seuls les labels étaient présentés ; voir aussi Rosch et al., 1976 ; Giordano et al., 2010) et les qualités sonores.

Si la reconnaissance de sons environnementaux reste (en partie) basée sur des indices acoustiques, on peut se demander dans quelle mesure les indices du timbre

6. La taxonomie est un système grâce auquel les catégories, qui regroupent des objets considérés comme équivalents, sont reliées les unes aux autres par l'inclusion de classes, en fonction des fréquences d'apparition et des dépendances entre les attributs perceptifs des objets réels (Rosch et al., 1976).

pourraient se généraliser à des sons non-instrumentaux. En comparaison à des sons d'instruments, les égalisations en hauteur et en niveau sonore de sons de l'environnement peuvent être plus difficiles à effectuer, avec des sons parfois assez longs (e.g. 5 s dans l'étude décrite par Susini et al. (1999)). Moyennant certains ajustements, des auteurs parlent néanmoins du timbre de sons environnementaux pour leur appliquer la méthode MDS et bénéficier des avantages de la comparaison de sons sans critères prédéfinis (e.g. véhicules, convecteurs d'air climatisé, klaxons ; Susini et al., 2005 ; Lemaitre et al., 2007 ; Minard et al., 2008 ; Misdariis et al., 2010). Les auteurs comparent généralement des classes restreintes de sources sonores. Susini et al. (1999) précisent bien l'importance de l'homogénéité des stimuli utilisés pour les jugements de dissemblance, afin d'éviter les jugements basés sur une connaissance a priori des objets sonores. Ils citent notamment un test réalisé avec des sons de l'environnement très hétérogènes et dont la structure des résultats prenait une forme fortement catégorielle, c'est-à-dire qu'ils étaient basés sur l'identification des sources sonores plutôt que sur des jugements de dissemblances perceptives.

De même, dans l'exemple de l'étude de Gygi et al. (2007), la solution MDS 3D obtenue répartit les sons de sources similaires en trois clusters : sons harmoniques, sons d'impacts discrets, et sons continus. Il est possible de repérer assez facilement ces cas de figure d'après la répartition hétérogène des sons dans l'espace multidimensionnel. A défaut de pouvoir complètement égaliser les sons en hauteur, en niveau sonore, ou en durée perçue, les auteurs peuvent être amenés à demander aux participants de ne pas baser leurs jugements sur ces trois dimensions (e.g. Lemaitre et al., 2003 ; Gygi et al., 2007).

Susini et al. (1997) (voir aussi Susini et al., 1998, 1999) ont réalisé une étude impliquant des jugements de dissemblance entre des sons de moteurs de véhicules. Les auteurs ont établi un espace perceptif à deux ou trois dimensions avec spécificités. Les corrélats acoustiques n'y sont pas explicitement précisés mais étant donné le caractère quasi-stationnaire des sons, ils sont essentiellement spectraux et entre autres inspirés de corrélats classiques (e.g. Krimphoff et al., 1994). Par la suite, Susini et al. (2001) ont appliqué ces méthodes à des sons générés par des systèmes d'air conditionné (voir aussi Susini et al., 2004). Cette fois, l'espace perceptif obtenu comporte trois dimensions avec spécificités, dont les corrélats

acoustiques sont respectivement : le rapport entre partie bruitée et partie harmonique[7], le CGS, et le niveau sonore perçu. De même, Lemaitre et al. (2003) (voir aussi Lemaitre et al., 2007) ont repris la méthode MDS pour l'appliquer à des sons de klaxons de voitures, et ont établi un espace perceptif à trois dimensions. En se basant sur les mêmes sources de la littérature que Susini et al. (1999) (qui n'indiquent pas leurs corrélats) et que Marozeau et al. (2003), les auteurs ont trouvé que le CGS, la rugosité, et la déviation spectrale (liée à la structure fine de l'enveloppe spectrale) corrélent à leurs dimensions perceptives. Comme pour Susini et al. (1997), la quasi-stationnarité des sons ne permet pas de retenir des dimensions temporelles et spectro-temporelles liées à l'attaque des sons (la rugosité est une propriété temporelle à court terme par ailleurs).

Pour évaluer la cohérence entre les études de timbre de sons environnementaux, Minard et al. (2008) (voir aussi Misdariis et al., 2010) ont comparé quatre études utilisant différents types de sons : klaxons (Lemaitre et al., 2007), intérieurs de voitures (McAdams et al., 1998), fermetures de portières de voitures (Parizet et al., 2008), convecteurs d'air conditionné (Susini et al., 2004). L'ensemble des sons sont regroupés en trois classes : sons d'impacts (portières de voitures), sons de moteurs (intérieurs de voitures et air conditionné), sons similaires à des instruments (klaxons). Les espaces de timbre générés pour chacune de ces classes sont différents, avec deux ou trois dimensions, avec ou sans spécificités, avec ou sans classes latentes. Plusieurs corrélats acoustiques candidats sont testés à l'aide notamment des descripteurs de Peeters (2004) : niveau RMS, niveau sonore perçu, HNR, CGS, étalement spectral, brillance complexe (CGS tenant compte d'une partie bruitée dans le signal), netteté, rugosité. Le CGS (complexe ou non) est

7. L'harmonicité est un paramètre déterminant pour la perception de la hauteur (Meddis & Hewitt, 1991a), et plus généralement du caractère bruité du son. Elle peut être quantifiée à l'aide du rapport entre composantes harmoniques et composantes bruitées dans le signal (HNR, pour *Harmonic-to-Noise Ratio*, cf. Boersma (1993) ; soit l'inverse du rapport calculé dans l'étude de Susini et al. (2001)). Le HNR est utile, par exemple, pour caractériser le niveau de bruit dans la voix (e.g. Qi & Hillman, 1997 ; Lewis et al., 2009) ou dans des vocalisations animales (e.g. aboiements de chiens ; Riede et al., 2001), pour lesquelles il prend des valeurs plus importantes que pour d'autres catégories de sons de l'environnement. En particulier, la Figure 6 de l'article de Lewis et al. (2009) donne des intervalles de valeurs que peut prendre le HNR pour différentes sous-catégories de vocalisations. Il a également été beaucoup utilisé dans des études d'imagerie cérébrale comme contrôle acoustique de stimuli naturels très divers (voir paragraphe I.1.4.3 ; e.g. Lewis et al., 2005 ; Murray et al., 2006 ; Lewis et al., 2009 ; Staeren et al., 2009 ; Leaver & Rauschecker, 2010 ; Giordano et al., 2013 ; Latinus et al., 2013).

trouvé de façon commune aux trois classes de sons, tandis que d'autres corrélats sont spécifiques à chaque classe : valeur RMS et netteté pour les sons d'impact, indiquant leur caractère impulsif, le HNR et l'étalement spectral ou le niveau perçu pour les sons de moteurs, la rugosité et l'étalement spectral pour les sons similaires à des instruments. Hormis pour les sons d'impact, le timbre des autres classes de sons n'est pas décrit par des caractéristiques temporelles, indiquant que l'impulsivité semble être un paramètre important pour distinguer ces deux grandes classes de sons.

Ces différentes études ont montré l'applicabilité relative de la méthode MDS à des sons de l'environnement, dans la mesure où les sons utilisés sont suffisamment homogènes. Cependant, les corrélats trouvés dépendent justement des sons utilisés dans chaque étude, et surtout de leur impulsivité. Dans le cas de sons stationnaires, les corrélats sont avant tout spectraux (CGS, étalement spectral, HNR, structure fine du spectre) mais aussi temporels (rugosité). L'égalisation en hauteur et en niveau sonore requise par la définition du timbre reste le problème majeur avec des sons de l'environnement, des études trouvant par exemple le niveau sonore comme corrélat d'une dimension perceptive (Susini et al., 2001 ; Minard et al., 2008). D'autres facteurs cognitifs peuvent aussi entrer en jeu, comme la familiarité avec les sources sonores (Ballas, 1993 ; Murray et al., 2006).

### 1.3.6 Bilan sur le timbre

**Identification de variations perceptives induites par des variations acoustiques contrôlées.** Les études sur le timbre ont apporté une grande contribution à la compréhension de la reconnaissance auditive, mais ont aussi montré leurs limites. Depuis les années 70, la méthode MDS a été calibrée et reprise dans la majorité des études pour tester des stimuli instrumentaux facilement égalisables, d'autant plus lorsqu'ils sont créés par synthèse sonore. Ces études ont ainsi montré qu'il est possible de faire des liens forts entre les dimensions perceptives qui permettent de distinguer des sons entre eux, et leurs caractéristiques acoustiques. En majorité, les corrélats acoustiques prennent part aux trois dimensions : temporelle, spectrale, et spectro-temporelle, avec respectivement le temps d'attaque, le CGS, et le pattern spectro-temporel des harmoniques, notamment lors des transitoires. Mais il faut rappeler que la définition du timbre se fait par la négative,

et que par conséquent, ses dimensions sont établies à partir de comparaisons de sons. Ces comparaisons contraignent les conditions expérimentales à se restreindre à certains groupes de sons à comparer entre eux. Par suite, elles ne permettent pas de généraliser directement les indices à des sons de natures différentes (e.g. d'un son d'instrument à un son de véhicule).

Le degré de généralisation des indices du timbre à de nouveaux stimuli a néanmoins été testé dans plusieurs études du timbre de sons environnementaux, en appliquant avec succès la méthode MDS, du moins lorsque leur homogénéité est contrôlée. Si ce n'est pas le cas, le risque est d'induire des critères catégoriels abstraits plutôt que strictement perceptifs dans la tâche de jugement de dissemblance. Les espaces perceptifs ont permis de mettre à jour des corrélats acoustiques proches de ceux obtenus pour la description du timbre de sons instrumentaux, tels que le CGS ou la déviation spectrale (Susini et al., 2005). Mais d'autres corrélats supplémentaires sont nécessaires pour expliquer complètement ces espaces, comme le HNR ou la rugosité. A l'inverse, le temps d'attaque n'a pas été mis en évidence dans les études avec des sons environnementaux de nature quasi-stationnaire, mais reste un paramètre important si les sons peuvent se distinguer par leur impulsivité (Minard et al., 2008). La manipulation de ces paramètres acoustiques, par analogie aux essais d'intervalles de timbres entre sons instrumentaux, est l'une des voies envisagées pour travailler sur l'identité de la source sonore (Susini et al., 2005). Bien que la saillance des dimensions perceptives reste assez spécifique aux stimuli utilisés, avec le risque de passer à côté de dimensions importantes du timbre, l'ensemble de ces études montre que certains paramètres acoustiques sont prépondérants pour la reconnaissance des sons, incluant le temps d'attaque, le CGS, le HNR, le flux spectral.

La prise en compte de spécificités a permis d'affiner l'ajustement algorithmique de l'espace multidimensionnel avec les données perceptives. On verra que ces aspects discrets du timbre méritent d'être développés dans un cadre théorique de reconnaissance auditive parcimonieuse (voir plus bas, paragraphe I.2.2.1 et la Figure 13). Entre temps, l'observation des évolutions temporelles et spectrales sur le spectrogramme des sons a aussi été améliorée avec l'utilisation de corrélats auditifs (caractérisations auditives des enveloppes temporelle et spectrale), permettant de quantifier plus fidèlement les qualités du timbre utilisées pour la

reconnaissance des sons. Sur ce point, l'étude d'Elliott et al. (2013) va plus loin en proposant de se baser sur l'observation des modulations spectro-temporelles pour se rapprocher encore plus près du traitement effectué par le système auditif. Les corrélats sont difficilement interprétables mais permettent de décrire avec plus de précision l'espace du timbre et peuvent avoir des applications en apprentissage automatique (voir aussi Patil et al., 2012). De toute évidence, la représentation spectro-temporelle peut fournir les indices pertinents pour la reconnaissance du timbre.

**Timbre et reconnaissance auditive.** Les études sur le timbre permettent de caractériser certains indices susceptibles de conduire à la reconnaissance de sons naturels. Leur complexité est réduite à l'aide des corrélats des dimensions perceptives, et d'autant plus lorsqu'y sont intégrés des modèles auditifs. On parvient ainsi à mieux saisir les sources potentielles d'indices auditifs pertinents pour la reconnaissance auditive.

Pourtant, cette solution méthodologique soulève deux problèmes. Premièrement, même si l'on utilise des sons naturels, en cherchant à isoler des dimensions perceptives sur lesquelles pourraient se répartir des indices de reconnaissance, on contredit le principe selon lequel ces indices seraient présents dans les patterns complexes et complets des sons naturels, toutes dimensions confondues. Des indices qu'on élimine suivant l'égalisation du timbre pourraient potentiellement servir dans d'autres contextes (e.g. une hauteur non-stationnaire pour reconnaître un chant d'oiseau), tandis qu'on pourrait être tenté d'égaliser les stimuli dans plus de dimensions encore. Deuxièmement, le champ des qualités perceptibles reste de toute façon très vaste, sa caractérisation est très variable suivant les études, et des indices taxonomiques issus d'une connaissance sémantique sur les sources sonores, qu'on ne considérait pas jusque-là, doivent s'y adjoindre dans le contexte d'une écoute naturelle (Ballas, 1993 ; Gygi et al., 2007).

Un certain nombre d'auteurs ont retourné ces problèmes en prenant le parti d'évaluer l'appartenance de sons naturels à des catégories sonores d'après la sélectivité des aires cérébrales à celles-ci. De fait, ces études ont la possibilité d'aborder le traitement auditif de n'importe quelle catégorie sonore pour ensuite seulement comparer des paramètres plus ou moins bas-niveaux (dont ceux du timbre) afin

de tenter d'expliquer ces sélectivités cérébrales (Staeren et al., 2009 ; Giordano et al., 2013).

## 1.4 Corrélats cérébraux de la reconnaissance auditive

### 1.4.1 Hiérarchisation et abstraction de l'information auditive

A quel moment et sous quelle forme l'abstraction des propriétés du signal acoustique permettant la reconnaissance auditive a-t-elle lieu au cours du cheminement des caractéristiques temporelles et fréquentielles et de leur analyse le long des voies auditives ? Au cours de leur traitement, les représentations auditives véhiculent une quantité d'informations qui ne prennent pas toutes le même statut en fonction de la tâche auditive. En premier lieu, pour expliquer l'efficacité du codage des caractéristiques pertinentes parmi la très vaste quantité d'informations sensorielles à traiter, celles non-pertinentes doivent être filtrées (e.g. du bruit de fond ; Rabinowitz et al., 2013). Puis, en fonction de la tâche auditive, certains paramètres acoustiques semblent suivre des étapes hiérarchiques pour ignorer l'implication de paramètres utiles à d'autres tâches auditives. Cette hypothèse implique donc des représentations sensorielles multiples donnant accès aux différents aspects à traiter d'un même objet sensoriel (Kaas & Hackett, 1999).

De telles représentations invariantes à d'autres paramètres acoustiques comme ceux de localisation ont déjà été trouvées dans des aires auditives secondaires chez le furet (Walker et al., 2011). Des études antérieures chez le primate non-humain montrent similairement l'implication d'aires corticales distinctes dans le traitement de l'identité et de la localisation d'un objet auditif (Rauschecker, 1998 ; Romanski et al., 1999). Ces deux types de traitements nécessitent de l'information spectrale, qui suivrait deux flux indépendants dans un processus d'abstraction à haut-niveau : un flux ventral pour le traitement de reconnaissance de patterns auditifs (e.g. vocalisations ; le "quoi"), et un flux dorsal pour le traitement spatial (le "où"), originaires tous deux de la région centrale du cortex auditif (Rauschecker, 1998).

Les résultats de Romanski et al. (1999) appuient cette hypothèse d'un double flux [8] . Les auteurs ont notamment observé anatomiquement que deux aires au-

8. Sur ce sujet, voir aussi les études de Rauschecker & Tian (2000), Tian et al. (2001), ainsi

ditives du lobe temporal projettent dans des régions différentes du lobe frontal : l'une impliquée dans des fonctions spatiales et l'autre non. La convergence des deux flux dans le lobe frontal permettrait d'intégrer les deux types d'informations avec d'autres modalités (Romanski et al., 1999). A noter que si les deux principaux flux du "quoi" et du "où" semblent bien identifiés dans la littérature, des auteurs évoquent la possibilité d'un nombre plus important de flux, notamment d'après les multiples facettes que peut prendre un objet auditif (Lomber & Malhotra, 2008 ; Rauschecker & Scott, 2009).

Pourtant, ce sont aussi ces multiples facettes que peuvent présenter les objets auditifs qui les rendent difficiles à cerner (Griffiths & Warren, 2004). Par exemple, des sons de l'environnement reconnaissables peuvent être déterminés par la source sonore et ses propriétés mécaniques connues d'avance, à l'inverse de sons non-reconnaissables. De plus, les sources sonores peuvent combiner plusieurs caractéristiques acoustiques (e.g. les informations d'une hauteur et d'une voyelle peuvent être extraites séparément d'une même voyelle chantée), dont certaines peuvent être invariantes et ainsi permettre une reconnaissance indépendamment d'autres caractéristiques (e.g. la reconnaissance d'une voix indépendamment de son intensité), ou encore se définir en fonction des frontières des patterns spectro-temporels (deCharms et al., 1998 ; Shamma, 2001).

Dans une étude comportementale, Giordano et al. (2010) montrent que des objets sonores peuvent générer des représentations mentales abstraites et symboliques indépendantes du signal acoustique. En effet, désigner la source à l'origine d'un son implique des connaissances sémantiques pour symboliser l'entrée sensorielle. Or, les auteurs montrent que des sons d'objets vivants comparés à des sons d'objets non-vivants sont identifiés plus rapidement, plus précisément, et avec un vocabulaire moins hétérogène, indiquant qu'ils sont davantage évalués avec de l'information symbolique qu'acoustique. On verra dans le paragraphe suivant que ces deux grandes classes d'objets sonores, vivants (en particulier les vocalisations, cf. Theunissen & Elie, 2014) et non-vivants, ont justement fait l'objet d'études pour

que plus spécifiquement chez l'humain les études d'Alain et al. (2001), Maeder et al. (2001), Warren & Griffiths (2003). On pourra notamment se reporter au tableau proposé par Warren & Griffiths (2003), en annexe de leur article, qui récapitule les résultats d'un grand nombre d'études d'imagerie fonctionelle chez l'humain quant aux régions corticales auditives distinctes activées pour le traitement de patterns spectro-temporels ou bien de localisation.

observer l'existence d'activations cérébrales spécifiques, d'après leur familiarité écologique ainsi que leur caractérisation acoustique ou sémantique.

### 1.4.2 Catégories auditives

Le flux auditif du "quoi" passe par l'extraction successive de caractéristiques auditives utiles à différentes étapes du traitement jusqu'à en extraire l'objet auditif. La sélectivité cérébrale à différentes catégories auditives chez l'humain semble refléter ce traitement hiérarchisé (Murray et al., 2006 ; Lomber & Malhotra, 2008 ; Leaver & Rauschecker, 2010), et en particulier pour la voix (Belin et al., 2000, 2002 ; Fecteau et al., 2004 ; Warren et al., 2006 ; Charest et al., 2009 ; Lewis et al., 2009). En effet, les techniques récentes d'imagerie cérébrale ont permis d'observer des répartitions différenciées des régions corticales activées en fonction des catégories sonores utilisées en stimulation et de leurs propriétés. Pour des catégories sonores hétérogènes et complexes qu'il serait difficile de caractériser par des propriétés basiques, les études d'imagerie présentent au moins l'avantage de montrer que leur traitement auditif est spécifique en fonction de catégories complexes distinctes (e.g. outils, animaux, voix, instruments), comme cela a aussi pu être mis en évidence en vision (e.g. Kanwisher et al., 1997 ; Haxby et al., 2001). C'est le cas des études sur la voix, dont le traitement cortical spécifique est particulièrement bien documenté depuis les années 2000.

**La voix.** Un grand nombre d'études en éthologie ont révélé des circuits cérébraux sélectifs aux vocalisations conspécifiques (cf. Theunissen & Elie, 2014). La voix semble également bénéficier d'un statut particulier dans le traitement auditif chez l'humain. Belin et al. (2000) ont comparé l'activité cérébrale en réponse à des sons de voix et à des sons environnementaux. Le contraste des activités cérébrales respectives montre plus d'activité en réponse à des sons de voix le long du sillon temporal supérieur plutôt qu'à des sons de l'environnement non-vocaux (sons naturels, d'animaux, mécaniques), et avec une sélectivité plus forte dans l'hémisphère droit. Belin (2006) note cependant que la latéralisation du traitement de la voix est, dans la majorité des études (incluant des études comportementales et anatomiques chez le primate non-humain ainsi que des études de lésions cérébrales chez l'humain), plutôt marquée dans l'hémisphère gauche, tandis que l'hémisphère

droit serait davantage dédié au traitement de la prosodie ou à l'identification du locuteur.

Cette spécialisation corticale pour la voix a été confirmée par des études d'imagerie fonctionnelle chez l'humain (e.g. Belin et al., 2002 ; Fecteau et al., 2004 ; Kriegstein & Giraud, 2004 ; Warren et al., 2006 ; Pernet et al., 2015) [9] ou d'enregistrements électrophysiologiques (e.g. Levy et al., 2001 ; Charest et al., 2009), en comparaison à des sons de différentes complexités pour tester la robustesse de cette spécialisation en fonction des caractéristiques acoustiques (e.g. sons humains non-vocaux, instrumentaux, environnementaux, hybrides, etc.). L'activité cérébrale en réponse à des sons de voix ne semble pas être seulement générée par des caractéristiques bas-niveaux de la voix, mais reflète plutôt un traitement sélectif en tant que catégorie sonore complexe (voir aussi Norman-Haignere et al., 2015).

Des études comportementales chez l'humain indiquent que ce traitement spécialisé de la voix peut se traduire par un avantage pour cette catégorie auditive particulière dans certaines tâches auditives (e.g. temps de réaction plus rapides, reconnaissance de sons courts ; Agus et al., 2012 ; Suied et al., 2014). Il s'expliquerait par l'importance sociale des communications humaines dès le plus jeune âge et dès les plus anciens temps, dont la voix est le support principal. En effet, les capacités d'identification de ces paramètres semblent très précoces et n'ont pas été acquises avec des codes de la parole (par exemple, autant les primates non-humains que les nouveau-nés humains préfèrent les vocalisations de leur mère, c'est-à-dire les reconnaissent ; voir aussi Belin & Grosbras, 2010).

Le mécanisme de production de la voix est initié par la vibration des cordes vocales qui crée un son harmonique avec une hauteur tonale. Puis les différentes cavités par lesquelles l'onde sonore passe avant d'être émise à l'extérieur la filtrent en marquant certaines résonances caractéristiques appelées formants. Ce mécanisme est partagé par différentes espèces animales qui produisent des vocalisations (Belin, 2006). Dans la parole humaine, c'est le codage des patterns formantiques qui permet de discriminer les voyelles de façon immédiate et indépendamment de la

9. On pourra se reporter au tableau proposé dans l'article de Warren et al. (2006) et récapitulant les résultats d'un grand nombre d'études d'imagerie fonctionnelle chez l'humain quant aux régions corticales auditives distinctes activées pour le traitement de la voix.

hauteur, du niveau sonore, ou du timbre (Figure 9 ; Liberman & Mattingly, 1989).

A la fois la structure harmonique et les patterns formantiques semblent jouer un rôle dans le codage de la voix. D'après Lewis et al. (2009), la structure harmonique, quantifiée par le HNR, contribue à l'activation de gabarits spectraux sensibles à la voix au niveau du cortex auditif humain (voir aussi Latinus et al., 2013). La sensibilité au filtrage formantique (Figure 9) a aussi été étudiée en combinant des données d'IRMf et des analyses acoustiques. Formisano et al. (2008) ont estimé l'invariance des représentations cérébrales de sons de voix en fonction des variations acoustiques induites par différents locuteurs et différentes prononciations de certaines voyelles. Cette invariance est suffisante pour identifier automatiquement les voyelles prononcées à partir des patterns d'activations cérébrales. Les auteurs donnent des conclusions similaires à celles de Lewis et al. (2009), indiquant que la représentation auditive abstraite des informations contenues dans la voix émerge de l'encodage de l'information provenant de régions haut-niveaux, mais aussi des régions auditives précoces qui traitent des caractéristiques acoustiques basiques comme ici des fréquences formantiques.

En effet, la voix permet d'extraire une grande quantité d'informations sur le locuteur, comme le genre, l'identité, ou l'état émotionnel, de façon similaire au traitement des visages en vision (Belin, 2006). Cette analogie a conduit des auteurs à décrire la voix comme un "visage auditif" (Belin et al., 2002, 2004). En outre, le partage d'une grande quantité d'informations entre la voix et le visage correspondant faciliterait leur intégration sensorielle (Griffiths & Warren, 2004 ; Ghazanfar et al., 2005 ; Charest et al., 2009).

Enfin, la multiplicité des informations transmises au sein même de la voix sont aussi certainement traitées dans un certain ordre chronologique (e.g. l'identité du locuteur avant les caractéristiques phonétiques). C'est pourquoi des auteurs ont cherché à caractériser le décours temporel des caractéristiques perceptives des voix. Levy et al. (2001) ont observé des dynamiques cérébrales, à l'aide d'enregistrements électrophysiologiques, révélant un pic temporel à 320 ms après le début du stimulus sur les potentiels évoqués (PEs) dans le cas de voix comparées à des sons d'instruments jouant aux mêmes hauteurs tonales. Ces résultats confirment la spécificité du traitement des voix, mais ici peut-être pour un traitement préphonologique déjà à haut-niveau, en tout cas trop tardif pour être associé à un

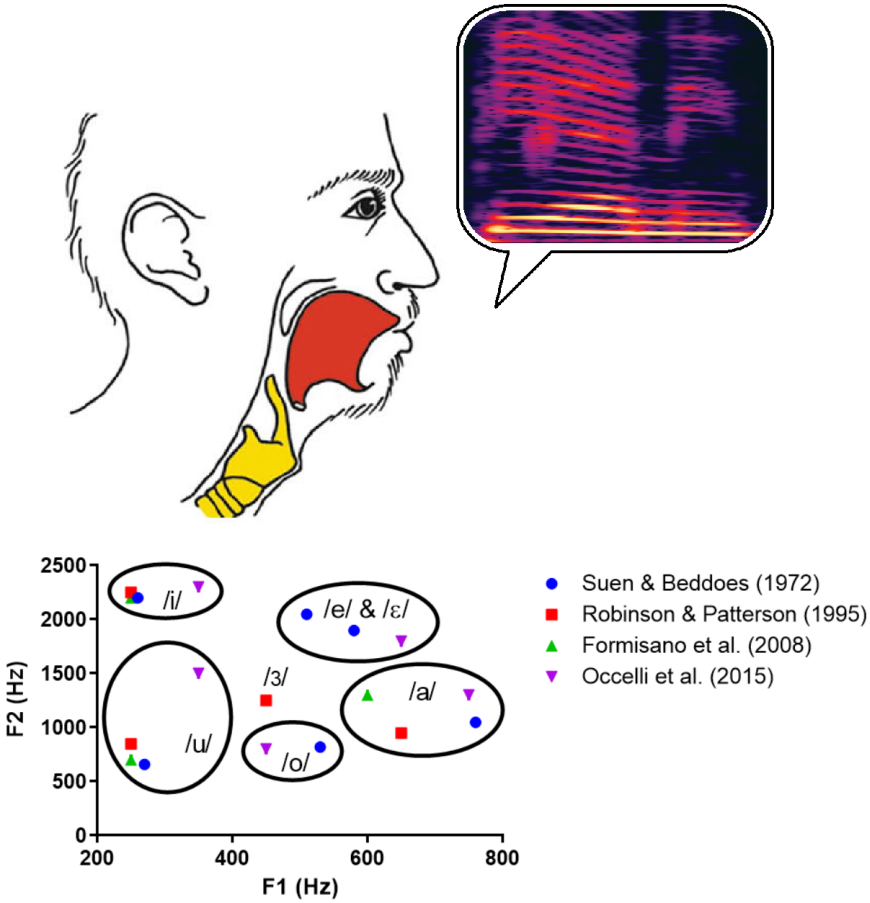


FIGURE 9 – Mécanisme de production de la voix illustré par le codage des fréquences formantiques en voyelles. En haut : vue sagittale de l'anatomie du tractus vocal chez l'humain. En rouge, la langue ; en jaune, le larynx. De cette anatomie découle la théorie source-filtre, selon laquelle les vocalisations résultent de la vibration d'une source (les cordes vocales dans le larynx) combinée au filtrage du tractus vocal, constitué de plusieurs formants, et qu'on peut observer sur le spectrogramme (0-5500 Hz) de la parole humaine (figure adaptée de Belin, 2006). En bas : représentation de la répartition des voyelles en fonction des deux premières fréquences formantiques indiquées dans quatre études différentes.

mécanisme de A1.

Charest et al. (2009) se sont interrogés sur la latence importante du pic du PE spécifique à la voix trouvé par Levy et al. (2001), étant donné que l'étude de Murray et al. (2006) (voir le paragraphe suivant) montre que 70 ms suffisent à discriminer des sons vivants comparés à des sons non-vivants. Les auteurs ont donc réalisé une nouvelle étude avec des stimuli de voix, de chants d'oiseaux, et de l'environnement, contrôlés dans les domaines temporel, fréquentiel, et temps-fréquence. Cette fois, des différences d'amplitudes des PEs en réponse aux sons de voix comparés aux sons d'oiseaux et environnementaux sont significatives dès 164 ms après le début du stimulus et surtout autour de 200 ms, au niveau des électrodes fronto-temporale et occipitale, soit presque deux fois plus tôt que celles observées par Levy et al. (2001). Les auteurs expliquent cela par des différences de matériel employé et/ou de protocole expérimental. De plus, des différences significatives entre les réponses aux sons d'oiseaux et aux autres catégories sont observées dès 80 ms, ce qui correspond davantage à l'ordre de grandeur rapporté par Murray et al. (2006), mais elles semblent s'expliquer par des différences acoustiques plus marquées pour les sons d'oiseaux. Dans tous les cas, ces études électrophysiologiques apportent des gages supplémentaires quant au traitement cérébral spécifique de la voix, ici d'après son traitement temporel.

### Des modules corticaux par catégorie sonore ?

**Gabarits de composantes acoustiques.** La notion de module cortical (e.g. Kanwisher et al., 1997) peut porter à confusion, dans la mesure où cela évoque des modules séparés et indépendants. En effet, si la voix semble bien solliciter une région corticale en particulier, il est difficilement envisageable de proposer des modules distincts pour chaque catégorie sonore. En revanche, l'activation de gabarits de composantes basiques combinées pour chaque catégorie sonore semblerait plus plausible (Lewis et al., 2009). En-dehors des sons de voix, des auteurs ont donc testé cette hypothèse avec d'autres types de catégories sonores.

En partant du constat de la spécialisation de certaines régions corticales pour le traitement de la voix, Leaver & Rauschecker (2010) ont testé si ce type de spécialisation corticale pouvait s'étendre à d'autres catégories sonores telles que

des chants d'oiseaux, d'autres animaux, de la parole humaine, et des instruments de musique. Les auteurs contrôlent les valeurs de six caractéristiques bas-niveaux du contenu spectral et de la variabilité temporelle, leur permettant de vérifier par la même occasion si les études antérieures sur le traitement spécialisé de la voix n'étaient pas biaisées d'après des profils acoustiques communs aux voix. Les auteurs n'ont pas identifié de régions sélectives aux chants d'oiseaux ni aux autres animaux. Cependant, ils identifient des réponses sélectives aux instruments de musique et à la parole dans des régions temporales supérieures antérieures, tandis que des régions plus proches de A1 ne sont pas sélectives aux catégories mais aux caractéristiques acoustiques basiques. Ces résultats semblent confirmer l'hypothèse d'une augmentation de la complexité des profils acoustiques représentés par les neurones ou les réseaux de neurones le long des voies auditives dans les régions temporales, jusqu'à encoder spécifiquement des instruments de musique et de la voix en tant qu'objets auditifs.

Toujours dans le but d'évaluer le degré d'abstraction de la structure bas-niveau des stimuli par des représentations corticales optimisées pour l'analyse d'objets auditifs, Giordano et al. (2013) ont utilisé la méthode multivariée d'analyse de similarité représentationnelle (cf. paragraphe II.1.4.1 ; Kriegeskorte et al., 2008), qui permet de mesurer la dissemblance entre des patterns d'activations cérébrales mesurés en IRMf pour un grand nombre de stimuli naturels. Cette méthode permet d'éviter de contraindre la sélection des stimuli, comme c'était le cas avec les stimuli très harmoniques de Leaver & Rauschecker (2010). Giordano et al. (2013) ont utilisé des sons très hétérogènes issus d'actions humaines (incluant des sons vocaux) et des sons non-humains. Les auteurs comparent la contribution de différentes caractéristiques acoustiques calculées sur les sons et par catégorie sonore. Des structures bas-niveaux semblent médier la sensibilité à des catégories, mais les auteurs notent aussi la présence de modules corticaux encodant des catégories auditives abstraites. En particulier, les auteurs observent un encodage auditif d'actions humaines dans des régions autres que le lobe temporal supérieur (cortex temporal moyen à postérieur), indépendamment de leur structure acoustique basique. Cependant, des caractéristiques telles que la hauteur, le niveau sonore, le CGS, le HNR sont encodées sélectivement par ailleurs, limitant à nouveau les conclusions sur le degré de la sélectivité cérébrale à des catégories auditives.

En partant d'une définition plus large de modules corticaux potentiellement spécialisés pour la reconnaissance de catégories auditives, c'est-à-dire en termes de patterns d'activations cérébrales, Staeren et al. (2009) ont testé leur différenciation en réponse à des sources sonores aussi variées que des chats, des chanteuses, des guitares acoustiques, et des tons purs. Les caractéristiques acoustiques des sons sont contrôlées par ailleurs et égalisées dans plusieurs dimensions acoustiques (durée, niveau RMS, enveloppe d'amplitude, HNR, profil temporel du spectre). Les auteurs utilisent une technique d'apprentissage machine similaire à celle utilisée par Formisano et al. (2008) pour la voix et sont de la même façon capables de classer automatiquement et au-dessus de la chance les patterns d'activations cérébrales après un entraînement sur des patterns labélisés par les catégories sonores testées. Selon les auteurs, ces résultats suggèrent que les patterns distribués spatialement dans les régions temporales supérieures d'après des caractéristiques basiques encodent globalement l'information abstraite de catégorie auditive, une conclusion similaire à celle proposée pour la voix (e.g. Formisano et al., 2008 ; Lewis et al., 2009).

Comme pour la voix (Levy et al., 2001 ; Charest et al., 2009), des auteurs se sont intéressés aux dynamiques temporelles en réponse à des sons plus divers. Murray et al. (2006) ont réalisé une tâche de détection de cible tout en mesurant les PEs chez l'humain pour estimer la rapidité de discrimination cérébrale d'objets vivants et d'objets fabriqués. Les auteurs ont observé un traitement différencié localisé dans l'hémisphère droit (gyri temporaux supérieur et moyen), dans des régions proches des aires auditives identifiées pour représenter des intermédiaires hiérarchiques du traitement auditif, et suivi rapidement par un traitement dans l'hémisphère gauche et par de l'activité bilatérale. Les auteurs citent d'autres études (e.g. Maeder et al., 2001 ; Lewis et al., 2005) qui suggèrent un traitement davantage bilatéral de l'objet auditif. Cependant, cette différence est selon eux imputable à la faible résolution temporelle des techniques utilisées, ne permettant pas d'éliminer la possibilité d'une activation précoce dans l'hémisphère droit. En effet, les auteurs estiment que les premières différences acoustiques (quantifiées à l'aide des spectrogrammes et des HNRs) surviennent au plus tôt à 140-145 ms après le début du stimulus, et que les catégories sonores sont explicitement reconnues au bout d'une seconde environ, d'après les résultats comportementaux de

temps de réaction. Or, l'amplitude des activations est plus importante pour les objets fabriqués dès la période analysée entre 70 et 119 ms, et une latence de 12 ms entre les formes d'onde des champs électriques des deux catégories sonores est également constatée dans la période allant de 155 à 257 ms. Les auteurs en concluent que ce traitement différencié entre catégories sonores n'est pas attribuable aux différences acoustiques bas-niveaux. Selon eux, une centaine de millisecondes correspond à la limite supérieure d'initiation de la discrimination corticale d'objets auditifs.

**Implication de facteurs extra-auditifs dans des représentations haut-niveaux.** Plusieurs auteurs se sont intéressés particulièrement aux activations cérébrales haut-niveaux de sons d'outils et d'actions, potentiellement destinées à un partage multimodal de l'information auditive. C'est par exemple le cas de Lewis et al. (2005), qui comparent les activations cérébrales en réponse à des sons d'animaux et à des sons d'outils. Ecouter et catégoriser correctement ou non des vocalisations animales génère préférentiellement une activité des portions moyennes des gyri temporaux supérieurs gauche et droit, pouvant refléter le traitement de certaines composantes acoustiques comme les contenus harmoniques et de phases qui différencient les deux catégories sonores.

En revanche, les sons d'outils correctement catégorisés, ou même des sons d'animaux catégorisés comme outils, activent préférentiellement un large réseau neuronal recouvrant des portions des cortex moteurs. Ce réseau de neurones dits "miroirs" est aussi activé indépendamment lorsque les participants miment la manipulation des outils. Son implication pour la reconnaissance des sons d'outils signifierait qu'à ce niveau du traitement cérébral, des mécanismes de raisonnement causal entreraient en jeu. Autrement dit, l'association à haut-niveau de caractéristiques acoustiques distinctives et de connaissances sur la production des sons d'outils serait à l'origine de représentations mentales riches et abstraites sur la forme et la spatialité des outils, potentiellement partagées avec d'autres modalités sensorielles pour aider à la discrimination.

Kaplan & Iacoboni (2005, 2007) font encore davantage pencher cette hypothèse du côté d'un traitement abstrait pour les sons d'actions en général, qui serait issu du développement de représentations symboliques indépendantes du signal

acoustique par association entre des sons et des actions effectuées ou observées au même moment. Dans le cas de sons d'utilisations d'outils, leur reconnaissance serait induite par la compréhension de l'action motrice associée plutôt que par les propriétés acoustiques du son (Lewis, 2006). Une étape supplémentaire serait impliquée dans le processus de reconnaissance auditive : il ne s'agirait pas simplement d'activer un gabarit codant pour une catégorie avec une combinaison de caractéristiques, mais d'inclure ensuite une étape de raisonnement sur ce gabarit. Toutefois, il est difficile de savoir avec ces données si cette étape supplémentaire correspond vraiment à celle de la reconnaissance du son ou d'une étape ultérieure visant par exemple l'intégration multimodale, étant donné que les caractéristiques acoustiques sont présentes initialement et doivent nécessairement conduire à cette reconnaissance.

A nouveau, le décours temporel du traitement de reconnaissance auditive peut apporter des indices relatifs à cette question. Comme pour la voix (Levy et al., 2001 ; Charest et al., 2009) et des sons plus divers (Murray et al., 2006), des auteurs se sont intéressés aux dynamiques temporelles en réponse à des sons d'actions. Pizzamiglio et al. (2005) ont mesuré les PEs en réponse à des sons produits par une action de la main ou de la bouche (pouvant théoriquement être effectuée par le participant) comparés à des sons de non-actions. Les résultats ont permis de mettre en évidence la séparation de deux systèmes : les sons d'actions modulent l'activité de l'aire temporale supérieure postérieure gauche et le cortex prémoteur gauche, tandis que les sons de non-actions modulent bilatéralement l'activité du pôle temporal. Dans le cas des sons d'actions en particulier, le cortex prémoteur gauche s'active environ 10 ms après le sillon temporal supérieur gauche (dont le pic d'activité s'établit à environ 290 ms), en impliquant un système miroir relatif à un programme moteur. Cette hiérarchisation de l'information reflète probablement une première description du son d'action dans le sillon temporal supérieur, avant de former un percept plus précis du son d'action dans le cortex prémoteur.

### 1.4.3 Bilan sur les représentations cérébrales de catégories auditives

Un grand nombre d'études sur les représentations cérébrales de catégories auditives mentionnent des traitements différenciés à haut-niveaux en fonction des catégories sonores testées (e.g. Belin et al., 2000 ; Fecteau et al., 2004 ; Kaplan &

Iacoboni, 2005, 2007), mais tiennent aussi souvent compte d'un encodage de caractéristiques acoustiques bas-niveaux (e.g. Formisano et al., 2008 ; Lewis et al., 2009 ; Staeren et al., 2009 ; Giordano et al., 2013). L'ensemble de ces études semble globalement trouver un consensus sur le processus d'abstraction des caractéristiques acoustiques le long de voies auditives hiérarchisées, permettant de former l'objet auditif à reconnaître. Cependant, le lieu exact de la formation de l'objet auditif pourrait se situer à plus bas-niveau avec des combinaisons de caractéristiques basiques pour représenter des objets auditifs à part entière dans A1 (Nelken, 2004 ; Santoro et al., 2014 ; Norman-Haignere et al., 2015). Cette étape, susceptible de conduire à la reconnaissance auditive, devrait se traduire en termes de combinaisons d'indices auditifs analysés (e.g. des gabarits spectro-temporels), et serait donc informative à la fois sur les composantes acoustiques nécessaires à la reconnaissance et sur le temps pris pour y parvenir. Il est cependant encore trop tôt pour conclure sur la base de ces seuls résultats.

Dans le cas des régions cérébrales sélectives à différentes catégories auditives pouvant inclure des sons d'une grande variabilité, on peut penser que l'information auditive qui y parvient a effectivement été reconnue. Il peut notamment s'agir de régions multimodales, puisque le partage de l'information entre différentes modalités sensorielles nécessite certainement de reposer sur des représentations plus abstraites de l'objet à reconnaître, et qu'on pourrait associer au processus de reconnaissance auditive en particulier (Belin et al., 2004 ; Kaplan & Iacoboni, 2005). Pourtant, si ces raisons justifient que la reconnaissance auditive ait bien déjà eu lieu une fois atteintes ces aires cérébrales, elle pourrait avoir eu lieu plus tôt.

C'est pour tenter de lever l'ambiguïté sur l'instant de la reconnaissance auditive que des auteurs ont utilisé des techniques d'enregistrements de l'activité cérébrale avec plus de précision temporelle (e.g. Levy et al., 2001 ; Pizzamiglio et al., 2005 ; Murray et al., 2006 ; Charest et al., 2009). En réalité, dans ces études, le consensus est encore moindre. Certaines d'entre elles indiquent des pics de PEs assez tardifs (Levy et al., 2001 ; Pizzamiglio et al., 2005), relevant certainement d'un traitement à haut-niveau mais peut-être aussi dépassant à nouveau le cadre strict de la reconnaissance auditive (e.g. traitement multisensoriel). D'autres études mentionnent de l'activité à des durées inférieures à 100 ms, que les auteurs préfèrent imputer à un traitement de différences acoustiques bas-niveaux (Cha-

rest et al., 2009), ou même à un traitement de reconnaissance antérieur à celui de caractéristiques bas-niveaux (Murray et al., 2006), ce qui semble pourtant aller à l'inverse du processus de reconnaissance auditive (un autre traitement acoustique bas-niveau non-identifié en était certainement la cause). D'autres voies théoriques et expérimentales sont à envisager pour tirer au clair l'influence des caractéristiques acoustiques spectro-temporelles sur le processus de reconnaissance auditive.

# 2 La parcimonie auditive

La complexité des sons naturels se traduit par une grande diversité de combinaisons de composantes acoustiques spectro-temporelles. Certaines de ces composantes acoustiques ne sont pas forcément utiles pour une tâche de reconnaissance auditive. La parcimonie du traitement auditif traduit la façon avec laquelle l'information strictement nécessaire pour la reconnaissance auditive est mise en évidence.

Dans cette section, nous verrons que le traitement auditif est économe et susceptible de sélectionner l'information auditive à transmettre à plus haut-niveau pour optimiser son traitement (paragraphe I.2.1). Cela nous amènera à proposer et discuter le concept d'*esquisse auditive*, i.e. des sons contenant des indices auditifs parcimonieux (paragraphe I.2.2). Enfin, nous listerons différentes méthodes existantes permettant d'extraire des indices parcimonieux dans les sons (paragraphe I.2.3).

## 2.1 Traitement auditif parcimonieux

### 2.1.1 Un encodage parcimonieux de l'information perceptive

**Réduction efficace de la complexité naturelle.** L'une des premières tâches du traitement perceptif est de réduire l'ensemble des données naturelles complexes et variant en permanence dans le temps. Toutefois, dans les années 70 en audition (Voss & Clarke, 1975, 1978), puis en vision (Field, 1987), des auteurs ont observé que la distribution de l'énergie de stimuli naturels n'est pas purement aléatoire mais suit une pente de $1/f$ en fonction de la fréquence $f$ ($1/f^2$ en vision en fonction de la fréquence spatiale $f$). Ces distributions compactes et localisées de l'énergie suggèrent l'existence de régularités entre les stimuli naturels qui pourrait expliquer aussi bien leur appartenance à une catégorie perceptive que leur codage cérébral (Field, 1987). Ainsi en audition, ces lois de puissance illustreraient des dépendances temporelles spécifiques aux sons naturels (e.g. vocalisations), qui sont globalement dessinés par une enveloppe temporelle variant lentement, tandis que celle-ci est combinée à une structure harmonique fine propre à chaque son (Theunissen & Elie, 2014).

La représentation cérébrale de la structure des sons naturels semble justement bien optimisée dès les filtres du nerf auditif suivant différentes échelles temps-fréquence : étroits en basses-fréquences pour les sons harmoniques des communications animales, et larges en hautes-fréquences pour les transitoires des sons de l'environnement (Lewicki, 2002 ; Olshausen & O'Connor, 2002). Ce compromis de décomposition temps-fréquence des filtres auditifs serait adapté aux sons naturels pour encoder le maximum d'information en utilisant leurs caractéristiques indépendantes. Le système auditif viserait ainsi à maximiser l'information transmise pour résoudre une tâche donnée, tout en minimisant les ressources cérébrales employées.

Ces hypothèses avaient été formalisées dès le milieu du XX$^{e}$ siècle dans la théorie du codage efficace par Attneave (1954) et Barlow (1961). Selon ces auteurs, les relais sensoriels effectueraient un recodage des messages sensoriels de manière à réduire leurs redondances en formant des représentations plus spécifiques. Ces mécanismes semblent effectivement refléter les stratégies du codage auditif (Schwartz & Simoncelli, 2001 ; Lewicki, 2002 ; Woolley et al., 2005). Plus particulièrement, le codage de l'information sensorielle est dit "parcimonieux" dans la mesure où il vise à réduire le nombre de caractéristiques à transmettre aux étapes ultérieures du traitement perceptif, de façon à utiliser un petit nombre de neurones actifs à chaque instant (Olshausen & Field, 2004 ; Felsen & Dan, 2005). Son optimisation consisterait à maximiser l'indépendance des composantes sensorielles impliquées (le codage parcimonieux n'est applicable que si les stimuli ont une structure qui s'y prête ; Olshausen & Field, 1997). Olshausen & Field (2004) (voir aussi les deux textes annexes à l'étude de Hromadka et al., 2008 ; Hromadka & Zador, 2009) y voient quatre avantages pour le traitement perceptif :

1. augmenter les capacités de stockage de patterns ;
2. obtenir une structure explicite des signaux naturels ;
3. obtenir des représentations plus lisibles de données complexes à différents niveaux de traitement ;
4. économiser l'activité cérébrale pour d'autres tâches.

Ces approches théoriques proposent des descriptions dites "surcomplètes" des stimuli obtenues à partir des représentations neuronales parcimonieuses et permet-

tant de suivre les transformations continues des stimuli avec les contraintes du codage parcimonieux (Figure 10 ; Olshausen & Field, 1997). Intuitivement, on peut comprendre qu'il soit plus facile de reconnaître un pattern parcimonieux dans un espace de haute dimensionnalité, plutôt qu'un pattern dense dans un espace de faible dimensionnalité (Hromadka et al., 2008). De plus, cette sur-représentation permet de produire un degré de parcimonie plus élevé en rendant les neurones plus sélectifs à des patterns à chaque étape de traitement. En vision, la complexité de l'environnement visuel naturel serait réduite grâce au traitement perceptif de ces régularités dans le cortex visuel primaire (e.g. des corrélations spatiales et temporelles ; cf. Felsen & Dan, 2005). De même en audition, les propriétés de traitement temps-fréquence du système auditif permettraient d'encoder l'information sans redondance (Smith & Lewicki, 2006).

Dans un tout autre champ de recherche, c'est également la conclusion à laquelle sont arrivés des auteurs en traitement du signal cherchant à coder des signaux audios à l'aide de représentations parcimonieuses (e.g. Plumbley et al., 2010 ; Sivaram et al., 2010 ; Elad, 2012). Dans leur cas, les signaux sont décomposés sur une base dont il faut trouver les coefficients non-nuls (e.g. un spectrogramme), en nombre réduit dans le cas des représentations parcimonieuses. Les avantages des représentations parcimonieuses en traitement du signal font écho à ceux supposés du codage cérébral : d'abord, seules quelques valeurs non-nulles doivent être codées et transmises (gain du codage) ; ensuite, signal et bruit ne sont pas représentés par les mêmes coefficients ; enfin, les ensembles de coefficients utilisés pour différents signaux ont peu de chances de se recouvrir (Plumbley et al., 2010). L'analogie entre le codage informatique et le codage cérébral présente toutefois des limites : le dictionnaire de codage et la capacité de transmission des sous-ensembles de paramètres significatifs dépendent des ressources computationnelles allouées.

**Données expérimentales.** Comme dans d'autres domaines, la théorie du codage parcimonieux a précédé de plusieurs décennies les investigations empiriques avec des stimuli écologiquement valides et dans différentes modalités sensorielles. Plusieurs études en vision montrent que le codage neuronal effectué dans les aires visuelles à haut-niveau pour représenter des objets visuels est sans doute de type parcimonieux, d'après des données physiologiques (e.g. Young & Yamane, 1992 ;

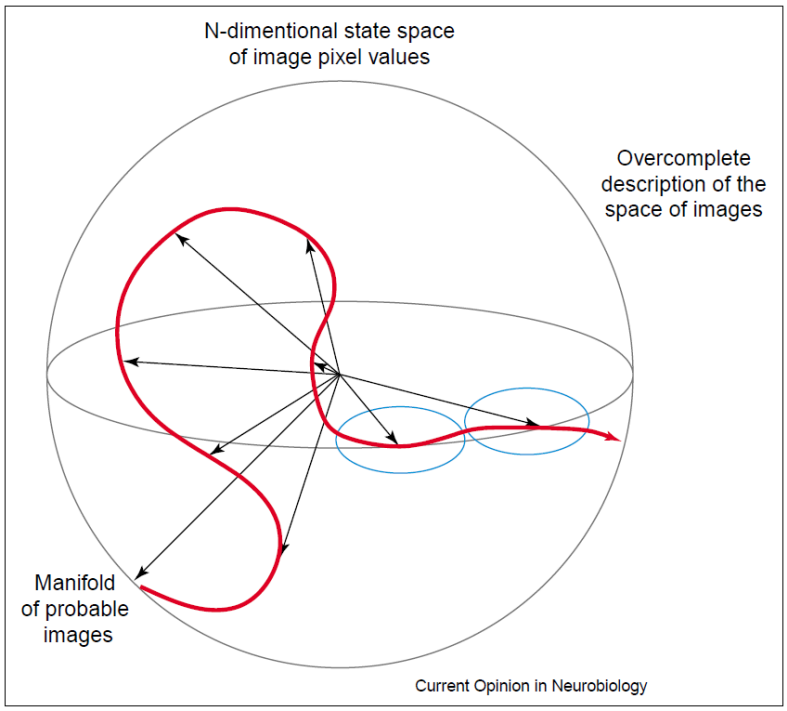


FIGURE 10 – Espace d'état de scènes naturelles et codes surcomplets. La sphère représente l'espace d'état N-dimensionnel de scènes naturelles, i.e. l'espace de tous les stimuli possibles composés de N échantillons (e.g. pixels). Les stimuli naturels sont supposés reposer le long de patterns de faible dimensionnalité inclus dans cet espace. La courbe rouge indique la trajectoire hypothétique d'une caractéristique de stimulus (e.g. le bord d'un objet visuel). Chaque flèche noire correspond à la caractéristique préférée d'un neurone. Les ellipses bleues représentent la zone de réponse du neurone. Cette représentation est surcomplète lorsqu'il y a plus de vecteurs de patterns que de dimensions d'entrée (e.g. pixels d'une image), permettant de simplifier les représentations pour des étapes plus élevées de l'analyse. Source : Olshausen & Field (2004).

Rolls & Tovee, 1995 ; Vinje & Gallant, 2000) ou des analyses computationnelles (e.g. Olshausen & Field, 1996, 1997). Par exemple, Young & Yamane (1992) analysent les réponses à des visages de deux populations de neurones. Leurs résultats montrait que la première, dans le cortex inférotemporal, véhicule de l'information correspondant aux propriétés physiques (des distances calculées entre les éléments du visage), tandis que la seconde, dans l'aire polysensorielle temporale supérieure, véhicule de l'information correspondant à des aspects sociaux plus complexes liés à la familiarité des visages. En particulier dans le cas du codage des caractéristiques physiques, les réponses de la première population de neurones sont plus similaires lorsque les visages sont plus similaires, d'après les redondances dues aux covariations de certaines caractéristiques comme des distances entre des éléments symétriques du visage. De plus, leur analyse montre que chaque neurone participe à la représentation de beaucoup de visages, et que pour obtenir l'acuité nécessaire pour identifier des visages, quelques dizaines de neurones seulement suffisent. Les auteurs estiment qu'une population de neurones complète utilisant un codage parcimonieux, comme celui qui semblait être à l'œuvre dans leur expérimentation, serait suffisante pour expliquer le comportement visuel réel.

En audition, certains auteurs ont d'abord cherché à déterminer les caractéristiques codées par des neurones seuls (e.g. deCharms et al., 1998 ; O'Connor et al., 2005). deCharms et al. (1998) ont utilisé une technique de corrélation inverse pour montrer que les STRFs de neurones de A1 chez le primate éveillé décomposent les sons en caractéristiques auditives locales (bords fréquentiels et temporels, transitions en fréquence ou intensité), par analogie à des caractéristiques visuelles locales (orientation des bords visuels, etc. ; voir aussi Shamma, 2001). Leurs résultats leur permettent d'estimer les stimuli optimaux qui peuvent générer une forte réponse des neurones, et donc de tester leur sélectivité et la décomposition en caractéristiques temporelles et fréquentielles. O'Connor et al. (2005) ont utilisé une autre technique d'optimisation pour trouver les caractéristiques spectrales préférées de neurones de A1 de macaques, consistant à modifier la composition du spectre du stimulus en fonction de la réponse du neurone. Les spectres préférés obtenus semblent prendre la forme que leur aurait donné le filtrage par des filtres auditifs optimisés pour des sons naturels.

Cependant, comme l'avaient déjà noté des auteurs en vision (Young & Ya-

mane, 1992 ; Olshausen & Field, 2004), le codage effectué par un neurone seul n'équivaut pas à la parcimonie du codage d'une population de neurones accordés ensemble. C'est pour cette raison que Hromadka et al. (2008) ont critiqué l'approche consistant à élaborer des stimuli optimisés pour générer des taux de décharges importants dans des neurones isolés. Ces auteurs se sont intéressés au contraire aux représentations de stimuli de synthèse et naturels à travers des populations de neurones dans le cortex auditif de rats non-anesthésiés. Les auteurs ont montré que ces représentations sont parcimonieuses car les taux de décharges sont élevés sur des périodes courtes dans moins de 5% des neurones de la population à chaque instant, tandis que l'activité globale augmente de 50%.

Toutefois, l'opposition entre les études portant sur des neurones seuls ou au contraire des populations de neurones n'est pas complètement tranchée. Dans l'étude de Hromadka et al. (2008), peu de neurones sont sollicités pour représenter un stimulus, et justement ces quelques neurones pourraient avoir fait partie de ceux testés dans les études avec des neurones seuls répondant préférentiellement à des stimuli optimisés. D'après Hromadka & Zador (2009), les stimuli optimaux seraient très divers et des neurones voisins ne répondraient pas préférentiellement aux mêmes stimuli. Pourtant, les stimuli naturels se distinguent aussi par leur diversité et pourraient avoir de ce fait excité des neurones proches les uns des autres. En outre, on constate que les taux de décharge sont comparables dans le cas de neurones isolés et dans le cas de populations de neurones (supérieurs à 20 décharges par seconde et par neurone). Ce qui laisse penser que des combinaisons de caractéristiques isolées des stimuli optimaux, codés par des neurones seuls, pourraient avoir été présentes dans les stimuli naturels, codés par des populations de neurones. Le passage de l'un à l'autre semble donc possible, que ce soit du point de vue des stimuli (naturels ou optimaux) ou du point de vue du codage neuronal.

L'étude de Smith & Lewicki (2006) propose une simulation de codage par potentiels d'action très efficace des caractéristiques spectro-temporelles de sons naturels (Figure 11). Cette abstraction théorique suit une approche représentationnelle du signal, c'est-à-dire que celui-ci est représenté par un ensemble d'impulsions neuronales lisibles sur une représentation temps-fréquence (sans tenir compte de la fréquence des décharges neuronales par exemple). Plus précisément,

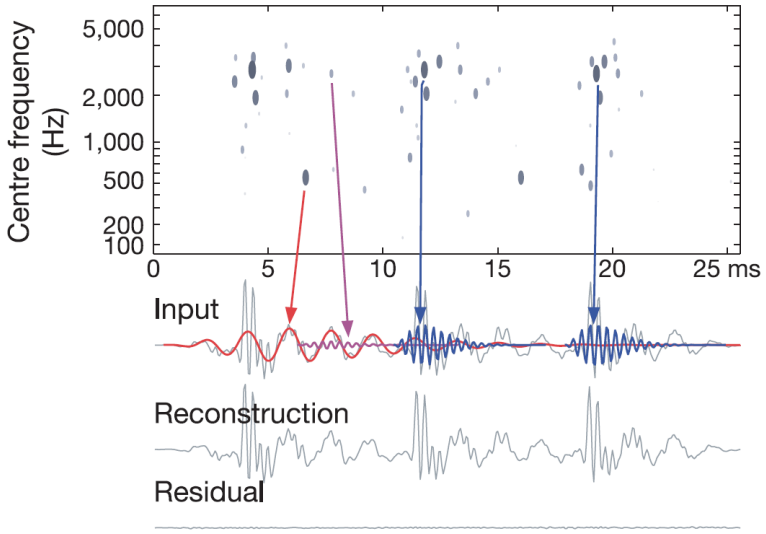


FIGURE 11 – Représentation de sons naturels avec des potentiels d'action. Un bref segment du mot "canteen" est représenté par un code de potentiels d'action (en haut). Chaque potentiel d'action (ovale) représente la position temporelle et fréquentielle d'une fonction noyau sous-jacente. Les flèches colorées illustrent la correspondance entre les potentiels d'action et la structure acoustique sous-jacente représentée par les fonctions noyaux. Source : Smith & Lewicki (2006).

elle consiste à décomposer le signal en éléments discrets pour ensuite manipuler des représentations de populations locales de potentiels d'action du nerf auditif. Un signal $x(t)$ est encodé à l'aide d'un ensemble de fonctions noyaux $\phi_1$, ..., $\phi_m$ positionnées arbitrairement et indépendamment dans le temps, et dont la forme et la longueur sont adaptées pour optimiser l'efficacité du codage. Leur formulation mathématique présente en outre l'avantage d'être flexible pour encoder des signaux acoustiques arbitraires.

Pour trouver la représentation parcimonieuse optimale, les auteurs utilisent une technique itérative en approximant le signal original. Les fonctions noyaux sont également optimisées pour adapter le code aux statistiques de l'environnement sensoriel (vocalisations et sons de l'environnement). Les auteurs observent qu'un petit ensemble de potentiels d'action est suffisant pour obtenir une reconstruction très précise du son. De plus, les fonctions noyaux obtenues présentent de fortes similarités avec des estimations de filtres obtenues elles à partir de données physiologiques du nerf auditif de chats, bien qu'elles aient été dérivées de façon

indépendante (voir aussi Lewicki, 2002). Selon les auteurs, il ne s'agit pas d'un hasard. Au contraire, les filtres auditifs semblent s'être adaptés de façon idéale à la structure statistique des sons naturels pour résoudre des tâches auditives.

A noter que Kording et al. (2002) avaient obtenu des résultats similaires en montrant que des représentations de neurones simulés peuvent générer une activité parcimonieuse en réponse à des données spectro-temporelles de parole. Pour ce faire, dans une représentation sont attribuées des propriétés différentes aux neurones d'une même population, et ceux-ci ont une activité souvent nulle et parfois très élevée. Les auteurs caractérisent les STRFs des neurones en leur attribuant un poids par une analyse en composantes principales sur des spectrogrammes auditifs. Les STRFs simulés sont ensuite comparés à des données physiologiques et montrent le partage de leurs propriétés.

### 2.1.2 L'exemple des textures sonores

En audition, les textures sonores correspondent au résultat global d'un nombre suffisant d'évènements acoustiques similaires (e.g. le flot d'un ruisseau, un essaim d'insectes, ou le crépitement d'un feu), par analogie à des textures visuelles (Figure 12). Elles donnent un exemple flagrant du cas où la quantité d'informations excède les capacités du système perceptif et où celui-ci doit adopter des stratégies pour ne pas se laisser submerger d'analyses et ainsi économiser des ressources cérébrales. En effet, même si l'information perceptive d'une texture est non-prédictible, le traitement du système perceptif vise à en extraire des homogénéités plutôt que d'analyser chaque détail (e.g. formes, directions ; Attneave, 1954).

On pourrait assimiler les textures sonores à du bruit à cause de la présence importante de composantes inharmoniques. Pourtant, le système auditif permet de les reconnaître efficacement en dépassant l'analyse de caractéristiques basiques comme la fréquence ou l'amplitude de composantes isolées. En effet, McDermott et al. (2013) ont montré que la reconnaissance de textures sonores est liée à leur redondance, du point de vue de leur contenu statistique. Pour ce faire, les auteurs ont développé une analyse similaire à celle de Field (1987), qui montre que certains neurones du cortex visuel répondent préférentiellement à des régularités statistiques dans des images naturelles. Field (1987) parvient ensuite à modéliser ses résultats physiologiques en montrant que des textures visuelles, superpositions

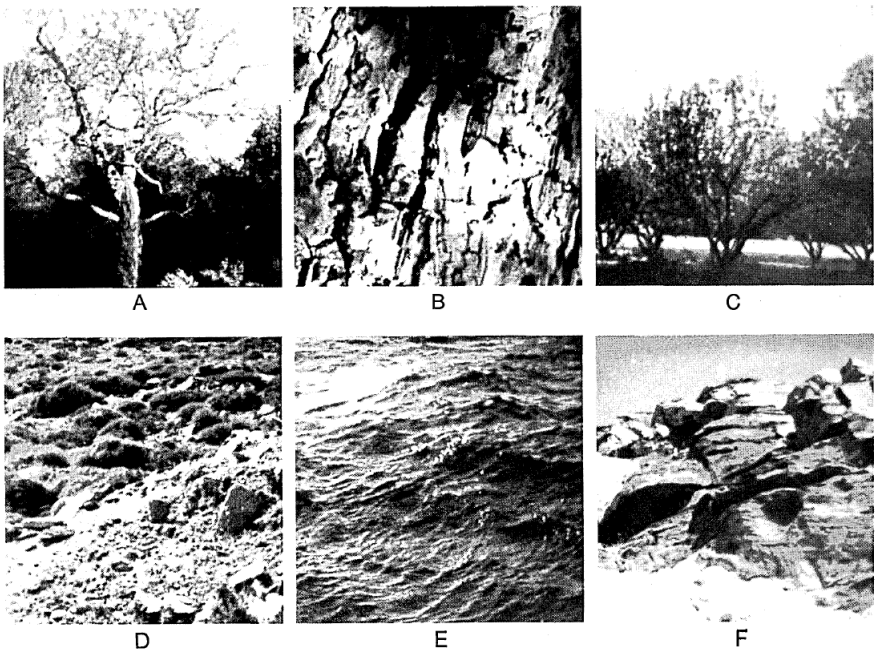


FIGURE 12 – Exemples de textures visuelles. Source : Field (1987).

de motifs se répétant dans l'image de façon plus ou moins complexe, peuvent se distinguer les unes des autres suivant la distribution de l'intensité des pixels (moyenne et variance), et le gradient des intensités.

Les textures sonores sont aussi des superpositions de motifs sonores simples qui se répètent. En tenant compte de ces répétitions, McDermott et al. (2013) sont parvenus à créer un résumé statistique qui consiste en un petit nombre de paramètres statistiques suffisant pour coder les sons de façon efficace. Pour cela, ils utilisent un modèle du système auditif périphérique et central pour découper le signal sonore en 600 bandes de modulation, pour n'en retenir que celles qui présentent une variance maximale entre elles. Ensuite, deux descripteurs statistiques (variations et covariations des enveloppes d'amplitudes au cours du temps) permettent de mettre en évidence des différences d'une texture sonore à une autre. Le résumé statistique obtenu permet de créer des textures de synthèse, à l'aide de bruits modulés par les descripteurs statistiques, fidèles aux sons originaux.

Au moins dans le cas des textures sonores, les résultats de McDermott et al. (2013) suggèrent que la reconnaissance et la mémorisation humaine s'appuieraient sur ces propriétés statistiques capturées par leurs résumés statistiques. Dans

une première expérience perceptive, les participants entendent une séquence de 3 échantillons de textures, dont 2 proviennent de la même texture, et doivent indiquer celui produit par une texture différente. La reconnaissance augmente quand la durée de l'échantillon augmente, ce qui semble aller dans le sens d'un moyennage temporel des caractéristiques statistiques. Dans une seconde expérience perceptive, les participants entendent une séquence de 3 échantillons de texture, dont 2 identiques, et doivent indiquer celui qui est différent des deux autres. Dans ce cas, leur score de reconnaissance augmente quand la durée de l'échantillon diminue. Ce résultat valide l'hypothèse d'un moyennage temporel, et l'approche statistique semble bien être mise en œuvre dans la perception des textures sonores. En définitive, résumer une texture sonore à un contenu statistique plus simple permettrait de réduire leurs coûts de stockage dans la mémoire à court terme et de supprimer des détails ne présentant pas d'intérêt écologique (voir aussi Nelken & de Cheveigne, 2013). De plus, ce processus d'abstraction induit par des mécanismes statistiques pourrait favoriser l'intégration multisensorielle en compensant les disparités entre les signaux physiques traités dans les différents canaux sensoriels.

Des mécanismes corticaux séparés ont été identifiés pour la ségrégation (i.e. la détection de frontières) et la représentation (i.e. l'abstraction et le maintien de caractéristiques) d'objets auditifs dans le cadre des analyses perceptives des propriétés statistiques de textures sonores (Overath et al., 2010). Un changement local des statistiques déterminerait la détection des frontières de l'objet auditif dans le cortex auditif incluant A1, tandis que la cohérence de segments plus étendus serait analysée dans des aires plus lointaines de A1. Selon Overath et al. (2010), ces analyses perceptives dépassent le cadre des textures auditives et feraient partie d'un mécanisme hiérarchique plus fondamental d'analyse d'objets auditifs. Le fait qu'un son soit reconnu dans différentes conditions acoustiques (e.g. réverbération, distorsions), voire malgré la présence éventuelle d'autres sources sonores concurrentes, semble indiquer que l'analyse de certaines caractéristiques particulières est maintenue tandis que d'autres sont négligées pour maintenir l'efficacité de la reconnaissance auditive.

## 2.2 L'esquisse auditive

### 2.2.1 Traits auditifs

Plusieurs paramètres acoustiques ont été identifiés pour expliquer la reconnaissance auditive, en particulier grâce aux études sur le timbre au moins dans le cas des sons musicaux. Malgré l'homogénéité des sons testés, ceux-ci peuvent parfois présenter des spécificités, ces qualités distinctives propres à certains sons et permettant de les reconnaître (Krumhansl, 1989 ; McAdams et al., 1995). Ces sons se démarquent donc des autres par un détail spectro-temporel qui leur est propre et qui n'est pas ignoré au profit des dimensions continues car ce type d'indices influence les jugements de dissemblance. On a vu aussi que la méthode MDS présente des limites dans le cas où les sons ne sont pas homogénéisés entre eux, car dans les comparaisons par paires, ils risquent d'être différenciés à cause de facteurs catégoriels plutôt que de ceux impliqués dans la perception du timbre. Néanmoins, cela n'empêche pas de penser que des sons de catégories différentes soient différenciés les uns des autres sur la base de traits distinctifs avant d'être reconnus de façon catégorielle. Par exemple, la distinction perçue entre un son d'instrument et un son de moteur, comme entre deux sons d'instruments, pourrait reposer sur la perception de traits auditifs, distincts du timbre cependant, dans la mesure où les égalisations requises en hauteur, niveau sonore, et durée sont impossibles.

Après tout, la reconnaissance auditive pourrait tenir compte uniquement des traits spécifiques des sons, discrets et parcimonieux, sans se baser sur des dimensions perceptives continues (McAdams, 1994 ; Pressnitzer et al., 2013). Dans cette approche, la perception d'un son ne correspondrait plus à un point placé dans un espace continu de faible dimensionnalité pour ainsi être réduit à quelques valeurs sur les dimensions perceptives, mais correspondrait au contraire à un ensemble fini de traits complexes discrets (Figure 13).

Cette nouvelle approche de la reconnaissance auditive sous-tend la problématique de ce qu'on considère comme un trait isolé, sa complexité, et les enjeux computationnels impliqués dans l'apprentissage, la mémorisation, et le traitement de l'ensemble de l'information de reconnaissance auditive prise sous forme de traits. Pressnitzer et al. (2013) soulignent néanmoins que plusieurs résultats de la littéra-

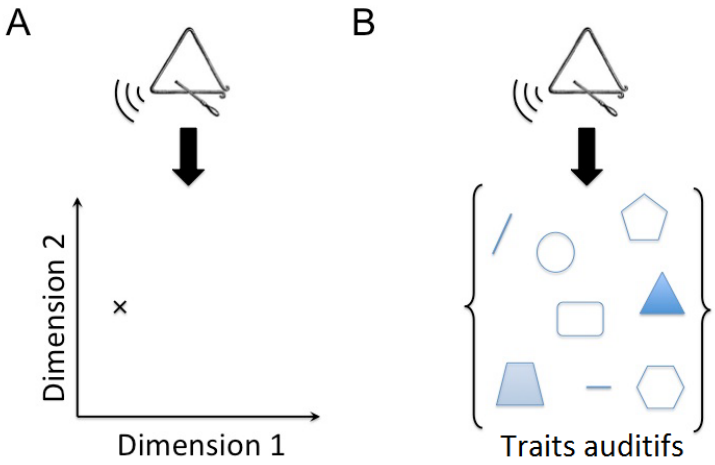


FIGURE 13 – Représentation schématique des approches du timbre : par espace multidimensionnel continu et par traits auditifs discrets. A : approche multidimensionnelle : les timbres sont projetés dans un espace de faible dimensionnalité et de dimensions continues ; B : approche par traits auditifs : chaque timbre est défini par un ensemble de caractéristiques distinctives parmi un ensemble très large et non-ordonné de caractéristiques possibles. Source : Pressnitzer et al. (2013).

ture semblent suggérer la présence, dans les sons, de traits mémorisés par les auditeurs, à l'échelle physiologique (e.g. des enregistrements neuronaux montrant une sensibilité accrue pour certaines propriétés spectro-temporelles ; Shamma, 2001) aussi bien qu'à l'échelle comportementale (e.g. la familiarité de certaines catégories sonores caractérisée par des temps de réaction plus courts comme pour la voix ; Agus et al., 2012).

En définitive, au vu des difficultés à saisir précisément le timbre, ses dimensions et ses spécificités, on peut penser qu'une combinaison des deux approches, dimensionnelle et par traits auditifs, ajustée par un mécanisme adaptatif en fonction du contexte acoustique, permettrait de résoudre ces difficultés, voire d'optimiser la reconnaissance auditive. En effet, un nombre limité de traits appris sur des sources familières pourrait constituer un ensemble de points de repère rigides dans un espace de timbre multidimensionnel. Ces points de repère permettraient aussi de faire face à la variabilité des configurations acoustiques dans lesquelles peut être présentée une source sonore. La perception d'un son connu mais distordu, réverbéré, ou bruité, ou bien d'un nouveau son, pourrait être évaluée autour de

ces repères utiles à la reconnaissance auditive.

### 2.2.2 Esquisses auditives : analogies avec la vision

**Essence et saillance.** Selon Harding et al. (2007), l'essence d'un stimulus correspond à l'ensemble des représentations construites pendant la perception permettant de déterminer son contenu pour le reconnaître. Le traitement de l'essence serait initial et rapide, permettant de déployer a posteriori l'attention pour une analyse détaillée des parties signifiantes avec cette connaissance a priori. En vision comme en audition, le système perceptif capterait une essence globale plutôt que de traiter séparément des caractéristiques physiques bas-niveaux (voir aussi Navon, 1977).

Plusieurs résultats de la littérature semblent appuyer cette hypothèse dans les deux modalités visuelle et auditive. Par exemple, avec des paradigmes de présentation sérielle et rapide de stimuli, des auteurs montrent que des catégories basiques sont reconnues très rapidement en vision (e.g. Buffat et al., 2013) ou en audition (e.g. Suied et al., 2013a) [10]. L'extraction de l'essence se ferait donc de façon très rapide et pré-attentive, dans les premières 100 ms du stimulus (cf. Harding et al., 2007). Des phénomènes de surdité au changement, analogue de la cécité au changement en vision, ont été mis en évidence par Vitevitch (2003), qui monte que des participants qui doivent répéter des mots ne remarquent pas que le locuteur change au milieu de la tâche. Des études sur l'effet *pop-out* (la reconnaissance d'une cible parmi des distracteurs indépendamment du nombre de distracteurs), en vision comme en audition, ont aussi permis de montrer un traitement efficace de l'information avec la détection et la focalisation sur des caractéristiques saillantes, tandis que le reste de l'information n'est pas traité dans le détail (e.g. Klein & Stolz, 2015). Le cas des textures, décrit précédemment, est un autre exemple similaire de traitement de l'invariance globale ignorant les détails visuels ou auditifs (voir aussi King & Nelken, 2009).

L'essence transmet les contours généraux du son pour le reconnaître rapidement. On peut donc l'interpréter comme une forme réduite de la saillance auditive,

10. Ces questions sur les capacités de reconnaissance rapide avec peu d'information auditive ont été traitées expérimentalement à l'aide d'un paradigme de présentation auditive séquentielle et rapide (cf. section II.2).

qui doit déterminer les évènements importants d'une scène sonore (Kayser et al., 2005 ; Tsuchida & Cottrell, 2012). Kayser et al. (2005) ont relevé plusieurs caractéristiques bas-niveaux (e.g. intensité, contrastes fréquentiel et temporel) qui seraient susceptibles de biaiser l'attention à un stade précoce du traitement vers ces traits auditifs perçus comme saillants, de la même façon qu'en vision (e.g. orientation, intensité, couleur ; Livingstone & Hubel, 1988 ; Itti et al., 1998). A partir de ces caractéristiques, les auteurs ont créé des cartes temps-fréquence basées sur des modèles auditifs pour mettre en évidence la saillance auditive et prédire la détectabilité d'évènements dans du bruit ou décrire l'importance potentielle d'un stimulus sur notre perception. Dans une tâche comportementale, des participants comparent la saillance de scènes auditives complexes. Le modèle prédit bien les notations humaines de la saillance, qui ne s'explique pas uniquement avec une variation d'intensité sonore. En comparant leurs résultats à ceux en vision, les auteurs en déduisent que les différents systèmes sensoriels doivent être basés sur des principes communs de détection d'évènements. Ce guidage attentionnel permettrait un traitement spécifique des évènements saillants. Inversement, le traitement du détail, non-perçu initialement, nécessiterait de se focaliser dessus par l'attention, voire grâce à un entraînement (Lively et al., 1994). Toutefois, ces résultats en audition concernent des stimuli de longues durées : la saillance conduit un mécanisme attentionnel à détecter un évènement dans une scène auditive. On s'interroge ici sur les composantes primordiales qui ont conduit à la formation d'un même objet auditif et donc à sa reconnaissance.

**Saisir l'essence dans une esquisse.** De tout temps, les hommes ont capté les traits essentiels de leur environnement pour les reproduire dans des représentations visuelles (Cavanagh, 2005). Malgré la diversité des œuvres d'art à travers les âges, Cavanagh (2005) note que même des peintures abstraites récentes peuvent transmettre un sens de l'espace commun avec celui d'œuvres plus anciennes. En effet, un peintre ne se base pas sur les lois de la physique pour reproduire une scène visuelle mais sur sa compréhension perceptive de cette scène. Cette "physique alternative", avec des ombres, des couleurs, ou des contours irréels voire impossibles, est plus simple pour comprendre le monde. L'artiste reproduit sa perception du monde. Surtout, cette physique alternative n'interfère pas avec la compréhension

FIGURE 14 – Les lignes permettent de représenter les contours de façon similaire dans ces dessins. A gauche : cheval chinois, vers 15 000 av. JC, grotte de Lascaux, France ; à droite : Le cheval maigre et le cheval gras, Jen Jen-fa, 1300 ap. JC, Musée National, Chine. Source : Cavanagh (2005).

de la scène chez les spectateurs. Il ne s'agit donc pas d'un codage propre à l'artiste ou partagé par un cercle restreint de connaisseurs [11].

Selon Cavanagh (2005), ces raccourcis perceptifs sont des sortes de découvertes de la perception rapide et efficace du système visuel, qui ont été confirmées par des expériences comportementales et physiologiques. Ainsi, des dessins de lignes caractérisent la forme d'objets complexes. Ce type de dessins existe depuis les arts primitifs et sont reconnus par des êtres humains quels que soit leur âge et leur génération (Figure 14). Des peintures présentent quelques fragments, un squelette minimal pour évoquer des images plus complexes, c'est-à-dire suggérer davantage que détailler [12].

Les représentations visuelles artistiques semblent bien refléter le traitement parcimonieux du système visuel, puisque le codage de stimuli naturels passerait par un traitement statistique dépendant de la forme générale (où se concentre l'information selon Attneave (1954)) plutôt que du fond (Redies, 2007). Surtout, la distinction entre une reproduction parfaite (qui reste à définir) et une reproduc-

11. On pourrait développer une analogie avec les nouveaux langages musicaux : "Si la musique n'était perceptible que par les gens qui la connaissent, on n'aurait jamais qu'un auditoire très limité." (Boulez & Archimbaud, 2016).

12. Des méthodes automatiques, utilisant de la synthèse sur des bases de données, permettent de générer des images complexes à partir de propositions d'esquisses (Turmukhambetov et al., 2015). En audition, des auteurs s'intéressent également à la façon de suggérer des sons complexes en utilisant la voix combinée à des gestes (Scurto et al., 2015 ; Rocchesso et al., 2016).

tion dégradée (artistique), mais pertinente pour la perception, permet justement de comprendre qu'il n'est pas nécessaire de tout reproduire, ni de reproduire toujours la même chose, en fonction de la tâche perceptive donnée. Ainsi, comme l'ont noté Smith & Lewicki (2006), toute l'information acoustique n'est pas pertinente pour une tâche auditive donnée. Et une représentation moins exacte de l'onde acoustique pourrait s'avérer plus pertinente biologiquement. A noter que des outils de compression très performants existent pour reproduire un objet perceptivement identique à son original, mais il s'agit pour ces outils d'omettre les détails imperceptibles. Dans notre cas, des éléments perceptibles pourront aussi être ignorés parce qu'ils n'entrent pas en compte dans la tâche de reconnaissance de l'objet initial (voir aussi Simoncelli & Olshausen, 2001).

**Le concept d'esquisse auditive.** A l'instar des œuvres d'art que l'on vient d'évoquer, des stimuli simplifiés pourraient être finalement plus proches de la perception qu'une représentation réaliste en mettant en évidence l'information pertinente pour une tâche donnée. Des résultats en vision montrent qu'on peut ainsi catégoriser des images dégradées sans que les parties individuelles soient bien définies (Oliva & Schyns, 1997 ; Oliva & Torralba, 2006). Il existe également en audition des cas de stimuli très dégradés générant de très bonnes performances de reconnaissance. En reconnaissance de la parole par exemple, différents auteurs ont montré que les caractéristiques acoustiques à court-terme de la parole peuvent être dégradées tout en maintenant une bonne reconnaissance grâce à au moins quatre procédés (cf. Figure 15 ; Remez & Thomas, 2013) : (1) *sine-wave speech* : seulement trois sinusoïdes sont placées aux fréquences centrales des premiers formants (Remez et al., 1981, 2001) ; (2) *noiseband vocoded speech* : seulement trois bandes de bruit sont modulées temporellement (Shannon et al., 1995) ; (3) avec un contenu spectral uniformément harmonique (Dorman et al., 1997) ; (4) chimères auditives : le contenu fréquentiel est varié arbitrairement (Smith et al., 2002).

Autrement dit, l'information spectro-temporelle dans la parole est très redondante (Cooke, 2006 ; Varnet et al., 2013). Des patterns dynamiques temporels suffisent à transmettre l'information, malgré une transmission minimale de l'information spectrale. Malgré aussi la perte de la qualité naturelle des sons, ce type de simplifications conserve les combinaisons des composantes naturelles complexes,

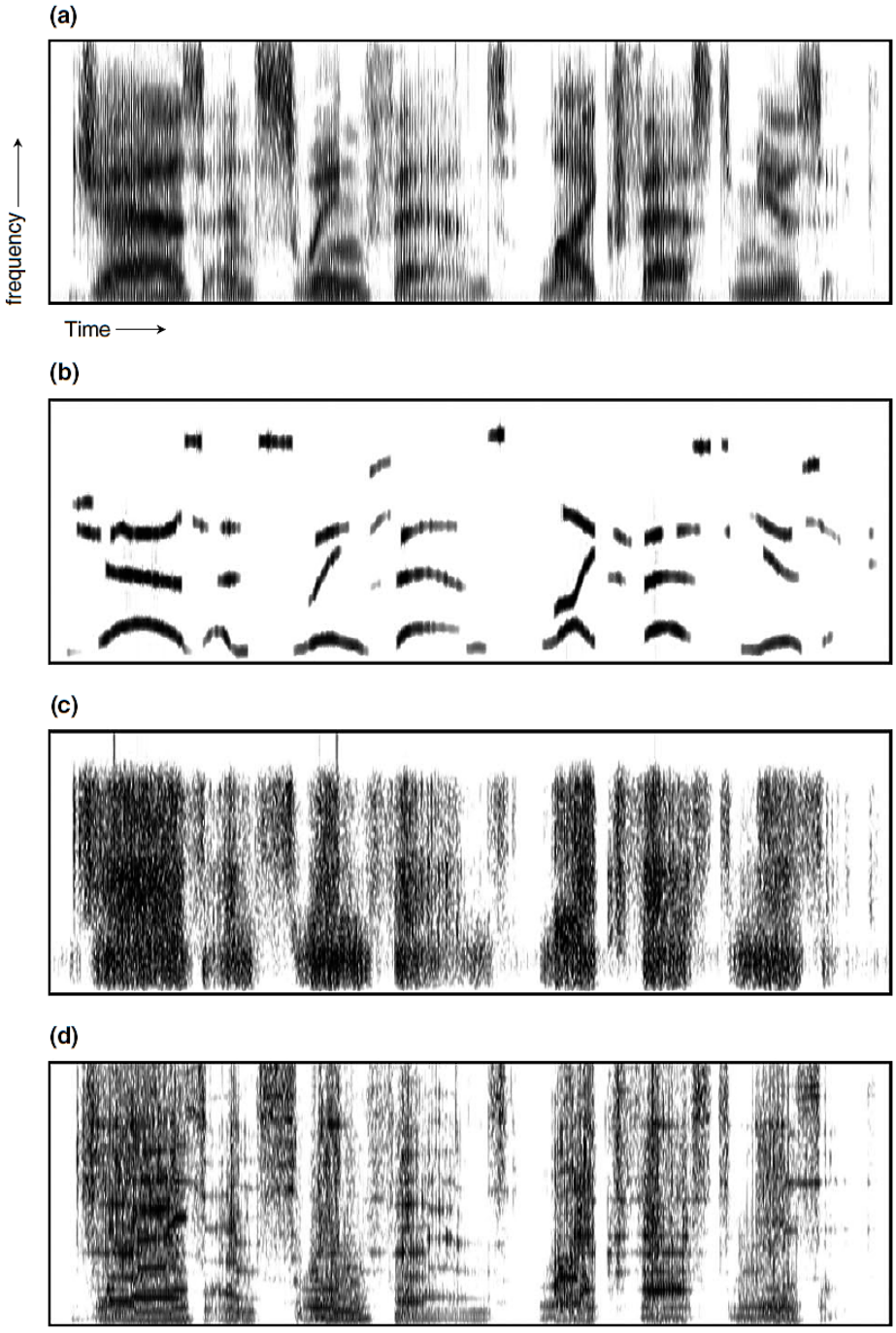


FIGURE 15 – Représentations spectrographiques de quatre variantes de la phrase "Jazz and swing fans like fast music". (a) Parole naturelle ; (b) *sine-wave speech* ; (c) parole vocodée par du bruit de bande ; et (d) chimère parole-musique. Source : Remez & Thomas (2013).

nécessaires pour leur reconnaissance. Gygi et al. (2004) ont appliqué la méthode de réduction de l'information spectrale de Shannon et al. (1995) à des sons environnementaux afin d'identifier les régions fréquentielles utiles pour leur reconnaissance. Des sons avec des indices temporels forts, par exemple courts et répétitifs, mais aussi d'autres types de sons environnementaux, peuvent être identifiés malgré une information spectrale très limitée. Le système auditif est également très performant dans le domaine temporel, avec par exemple une bonne reconnaissance de sons très courts, de l'ordre de quelques millisecondes (Gray, 1942 ; Robinson & Patterson, 1995a,b ; Suied et al., 2014).

Cependant, ces approches ne sont pas particulièrement parcimonieuses et ne peuvent donc pas rendre compte des dimensions utilisées dans le processus de reconnaissance (Varnet et al., 2013). C'est pourquoi Suied et al. (2013b) ont proposé le concept d'esquisse auditive, c'est-à-dire une représentation parcimonieuse d'un son naturel, ou pour reprendre les termes précédents, une concentration de l'essence du stimulus original. Dans le cadre d'une tâche de reconnaissance, leur objectif est de parvenir à isoler les traits parcimonieux contenus dans les sons naturels tout en supprimant le reste de l'information. Nous décrivons dans les paragraphes suivants différentes méthodes de la littérature permettant de mettre en évidence les traits utiles pour la reconnaissance auditive, y compris celle de Suied et al. (2013b) visant explicitement à construire des esquisses auditives parcimonieuses. En effet, les études en reconnaissance auditive ont permis de grandes avancées dans la compréhension du traitement auditif mais ont aussi montré certaines limites (e.g. la dépendance au contexte et aux stimuli). On verra que l'approche de Suied et al. (2013b), basée sur des modèles auditifs, permet d'y apporter des éclairages importants en se rapprochant de façon judicieuse des stratégies potentiellement utilisées par le système auditif pour le traitement des indices présents dans le signal auditif.

## 2.3 Simplification parcimonieuse de sons naturels

### 2.3.1 Informations disponible, représentée, et potentielle

Gosselin & Schyns (2002) ont proposé, dans le contexte de la perception visuelle, un cadre d'étude sur lequel on pourra s'appuyer pour distinguer les dif-

férents types d'informations impliquées dans la reconnaissance d'un stimulus. Ils comptabilisent trois types d'informations :

1. l'information disponible correspond à l'information physique de la cible (une image visuelle ou une onde acoustique) ;
2. l'information représentée correspond à l'estimation, dans la mémoire, de ce qu'est la cible d'après le participant (son image mentale) ;
3. l'information potentielle est l'intersection de l'information disponible et de l'information représentée, c'est-à-dire l'ensemble des caractéristiques physiques du stimulus qui correspondent à la représentation interne que s'en fait le participant. Cette information permet de déterminer le degré de correspondance entre le stimulus et sa représentation mentale a priori.

Comme Gosselin & Schyns (2002) l'ont noté, un observateur idéal utiliserait toute l'information disponible dans une image pour la catégoriser. Mais à cause de biais indépendants, comme celui de la connaissance sur l'information disponible, un observateur humain encode seulement l'information la plus utile pour cette tâche. De ce point de vue, on pourra considérer que le reste de l'information correspond à de l'information non-parcimonieuse. L'efficacité d'un observateur est directement liée au rapport entre information potentielle et information disponible. Plus l'information potentielle augmente, plus le participant peut identifier correctement le stimulus. Si en plus l'information physique disponible est limitée, cela signifie que le participant est performant pour identifier ce stimulus, car il se contente pour cela d'un minimum d'information.

On peut théoriquement généraliser ces trois types d'informations à n'importe quel stimulus à reconnaître. L'information disponible est connue puisqu'il s'agit du stimulus physique lui-même. L'enjeu concerne plutôt les informations représentée et potentielle, pour lesquelles Gosselin & Schyns (2002) distinguent deux méthodologies expérimentales permettant de les révéler, toutes deux issues de la psychophysique dite "moléculaire". La psychophysique moléculaire vise initialement à faire la part des choses entre les facteurs externes et internes impliqués dans la décision d'un participant (Green, 1964). Selon Gosselin & Schyns (2002), la corrélation inverse (*reverse correlation*) permet d'identifier l'information représentée, tandis que la méthode appelée *Bubbles* permet d'identifier l'information

potentielle (ces deux méthodes sont décrites en détail plus bas ; leur contribution respective est discutée par Murray & Gold (2004) et Gosselin & Schyns (2004)). Ces deux types d'informations seraient mis en jeu dans la reconnaissance d'un stimulus et permettraient respectivement de distinguer le stimulus physique de sa représentation perceptive, et de la transformation effectuée pour aller de l'un à l'autre. D'autres méthodes permettant de mettre en évidence plus particulièrement l'information potentielle sont décrites dans le paragraphe suivant.

### 2.3.2 Méthodes de simplification parcimonieuse de sons

**Corrélation inverse.** La corrélation inverse permet d'identifier, pour un participant, l'information représentée d'un objet donné (Gosselin & Schyns, 2002). Cette méthode suit le principe selon lequel une boîte noire peut être analysée grâce à du bruit, la boîte noire étant ici celle des représentations mentales du participant. Elle tient compte des perceptions correctes, mais également des perceptions incorrectes qui sont aussi informatives pour comprendre les processus de décision qui conduisent à la réponse du participant. A l'origine, l'intérêt principal de cette méthode était d'éviter tout biais imputable aux stimuli.

Cette méthode a surtout été utilisée en vision (Neri & Heeger, 2002 ; Gosselin & Schyns, 2003 ; Mangini & Biederman, 2004), bien qu'elle ait initialement été développée pour des tâches auditives (e.g. Ahumada & Lovell, 1971 ; Ahumada et al., 1975 ; Dai & Micheyl, 2010). Dans son principe général, elle consiste à présenter au participant des stimuli de bruit contenant une cible ou non. Il est ensuite possible d'étudier les caractéristiques des images lorsque le participant répond que la cible est présente ou non, en incluant les erreurs. Avec un grand nombre d'essais, cette méthode permet de faire ressortir du bruit les caractéristiques correspondant à la représentation mentale de l'objet à détecter. En pratique, il est possible de doser l'information disponible (le stimulus physique) jusqu'à présenter, à l'extrême, uniquement du bruit à tous les essais. Gosselin & Schyns (2002) citent des exemples en vision qui permettent de comprendre intuitivement le schéma général de cette méthode, comme la détection des contours d'un carré dans du bruit, alors même que ce carré est défini uniquement par ses angles, ou encore la détection d'une lettre 'S' quand bien même cette cible n'est jamais physiquement présentée. La somme des images pour lesquelles les participants perçoivent un carré ou un 'S'

révèle respectivement les bords du carré, ou la lettre 'S' dessinés dans du bruit, correspondant aux représentations mentales des participants.

En audition, Ahumada & Lovell (1971) se sont intéressés aux indices que recherchent les participants dans un stimulus cible (ici un ton de 500 Hz de 100 ms, présent dans la moitié des essais) pour indiquer sa présence, mais aussi son absence, dans un bruit gaussien de même durée construit en ajoutant 32 sinusoïdes entre 350 et 660 Hz. Ils ont pu déterminer la contribution relative des différentes composantes fréquentielles du bruit, en trouvant les combinaisons linéaires des amplitudes de ces composantes qui prédisent le mieux les notations des stimuli. Ces résultats montrent surtout que des modèles simples de détection d'énergie ne peuvent pas expliquer les données et que les participants doivent faire une comparaison entre l'énergie à la fréquence du signal et celle alentour.

Dans une expérience ultérieure, Ahumada et al. (1975) ont repris cette méthode en cherchant à observer si le bruit avant et après l'intervalle où la cible est présentée participe aussi au processus de comparaison. Les participants doivent noter la présence ou l'absence d'un ton de 500 Hz de 100 ms (présent dans la moitié des essais) centré dans un bruit gaussien large bande de 500 ms. Les auteurs expliquent la détection du ton comme une évaluation perceptive d'énergie à travers un filtre centré sur la fréquence du ton dont il faut estimer la largeur (entre 20 et 250 Hz) et le temps d'intégration (entre 50 et 500 ms) en corrélant les réponses des participants avec l'énergie des stimuli correspondant. La meilleure corrélation est obtenue pour une largeur de bande de 40 Hz, avec un temps d'intégration qui dépend de la présence (100 ms) ou de l'absence (300 ms) de la cible. Les auteurs notent aussi que sur les essais où la cible n'est pas présente, les participants cherchent un même pattern de changements temporels et spectraux que lorsque la cible était présente. Autrement dit, les participants écoutent un changement dans le pattern d'énergie, et pas seulement une augmentation générale du niveau sonore.

Ces travaux de corrélation inverse ont des résonances fortes avec ceux menés pour déterminer des images de classification en vision (pour une revue de la littérature, voir Murray (2011)). Il s'agit essentiellement d'étendre le paradigme de détection d'un signal cible vers un paradigme de classification de deux signaux. A nouveau, du bruit ajoute des caractéristiques qui viennent brouiller la perception

du participant et qui permettent de comprendre ce qu'il a écouté pour prendre sa décision. Des travaux en audition se sont depuis insérés dans ce cadre d'étude des images de classification. Par exemple, Varnet et al. (2013) ont pu montrer grâce à ce type de méthodes que la seconde transition formantique est importante pour la classication des phonèmes /b/ et /d/, avec des syllabes /aba/ et /ada/ mélangées à du bruit.

Finalement, cette méthode est astucieuse pour avoir accès à l'information représentée mais présente aussi des limites. La principale étant que les paradigmes expérimentaux sont contraignants, avec plusieurs milliers d'essais généralement requis pour obtenir un bon rapport signal-sur-bruit pour un seul stimulus cible et avec le risque de générer de la fatigue auditive. D'autre part, la distribution statistique du bruit utilisé pour masquer des stimuli naturels doit être prise en compte dans le calcul des images de classification et influe donc sur leurs caractéristiques (Varnet et al., 2013).

**Bubbles.** La méthode *Bubbles* est complémentaire à la corrélation inverse. Elle permet de mettre en évidence l'information potentielle, c'est-à-dire l'information interagissant entre l'information physique disponible dans un stimulus et l'information représentée en mémoire (Gosselin & Schyns, 2002). Plus concrètement, elle a été développée dans le but d'identifier l'information utilisée dans des tâches de catégorisation visuelle (Gosselin & Schyns, 2001 ; Schyns et al., 2002 ; Gosselin & Schyns, 2002, 2005 ; Smith et al., 2005 ; Adolphs et al., 2005).

Cette méthode consiste à isoler des fragments d'une image que les participants doivent ensuite utiliser pour effectuer la tâche de catégorisation. L'espace physique du stimulus est ainsi échantillonné aléatoirement, afin de retenir l'information visuelle ayant conduit à une catégorisation correcte. Pour effectuer cet échantillonnage, on place un masque percé de bulles gaussiennes directement sur l'image à analyser. Une analyse plus fine est possible en adaptant la taille des bulles aux bandes de fréquences de l'image : les basses-fréquences, qui varient plus lentement dans l'espace du stimulus, bénéficient de plus grosses bulles, tandis que les hautes-fréquences bénéficient de plus petites bulles. En-dehors de ces fenêtres, le stimulus original est complètement masqué (Figure 16). A chaque essai, les bulles gaussiennes sont disposées aléatoirement sur l'image. La tâche pour le participant

consiste à reconnaître le stimulus à partir de l'information à laquelle il a accès. On parvient ainsi à identifier les régions de l'image nécessaires pour son identification. Avec suffisamment d'essais, tout l'espace disponible du stimulus est exploré suivant une recherche aléatoire. Enfin, contrairement à la corrélation inverse pour laquelle on s'intéresse aux images qui conduisent le participant à répondre que la cible est présente (même s'il faisait des erreurs), avec la méthode *Bubbles*, on s'intéresse cette fois aux réponses correctes comparées aux réponses incorrectes. L'image de proportion résultante pondère le degré d'informativité perceptive de chaque région de l'espace physique du stimulus (Figure 16).

Cette méthode a été utilisée avec succès en vision pour déterminer l'information potentielle conduisant à l'identification de visages (e.g. genre, expressivité, identité). Par exemple, Gosselin & Schyns (2001) ont montré que la région du contour de la bouche correspond à l'information utilisée pour déterminer l'expressivité du visage, tandis que les yeux et le centre de la bouche permettent de déterminer le genre du visage. De plus, en comparant les résultats humains avec ceux d'un observateur idéal modélisé, les auteurs ont observé que des régions théoriquement saillantes ne sont pas utilisées par les humains. Cela signifie que l'information contenue dans l'image n'est pas utilisée sans transformations de la part de l'observateur humain. Cela signifie aussi que la méthode *Bubbles* est pertinente vis-à-vis de la perception humaine, car elle permet d'identifier la contribution spécifiquement humaine dans le processus d'extraction d'indices.

L'étude d'Adolphs et al. (2005) vient confirmer cette dernière conclusion. En effet, les auteurs ont montré que la reconnaissance de la peur sur des visages s'effectue en regardant les yeux, en testant un patient souffrant de lésions bilatérales de l'amygdale. Celui-ci ne pouvait pas identifier normalement la peur sur les visages tandis que sa perception visuelle était normale, car il avait en fait perdu la capacité d'extraire les indices pour effectuer correctement cette tâche. Les auteurs ont montré avec la méthode *Bubbles* que ce patient ne regardait pas spontanément les yeux pour effectuer un jugement d'émotion. Pour atteindre le même score de reconnaissance qu'un groupe contrôle, fixé à 75% de réponses correctes, il utilisait deux fois plus de bulles que les participants normaux. Par contre, ses mauvaises performances ne différaient pas de celles du groupe contrôle lorsque les yeux sur l'image étaient masqués. De plus, en lui disant explicitement de diriger son at-

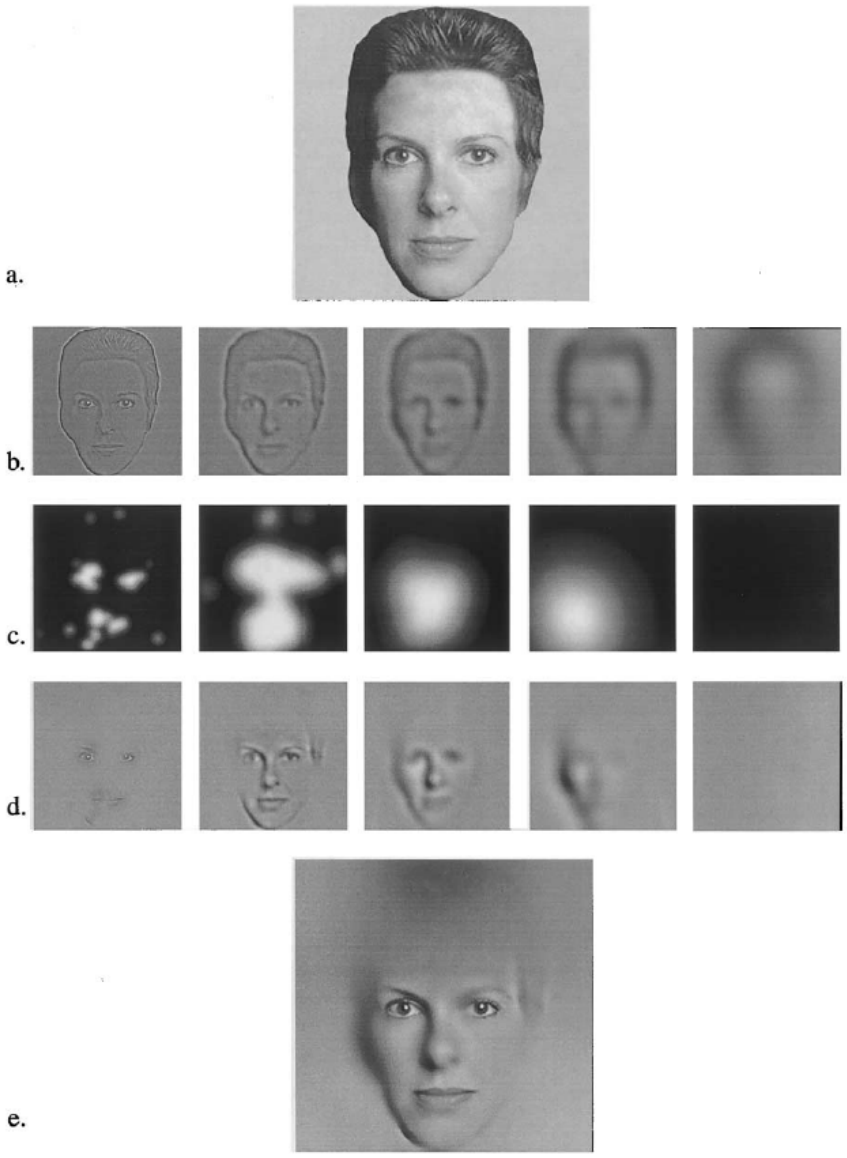


FIGURE 16 – Illustration de l'information potentielle sur un visage dans une tâche de classification avec la méthode *Bubbles*. Dans cette expérience, le nombre de bulles est ajusté pour maintenir une performance constante de 75% correcte. Les images de (b) représentent cinq échelles indépendantes de (a) (90, 45, 22.5, 11.25, et 5.62 cycles). (c) Régions potentielles statistiquement significatives pour chaque échelle spatiale. (d) Produit de (b) et (c). (e) Stimulus potentiel : une illustration de l'information utilisée pour identifier le visage. Source : Gosselin & Schyns (2001).

tention vers les yeux, les performances de reconnaissance de la peur devenaient normales.

La méthode *Bubbles* a été reprise plus tardivement dans la modalité auditive. Mandel (2013) a utilisé des bulles auditives pour prédire l'intelligibilité d'enregistrements sonores bruités. Le bruit, placé sur différentes régions du spectrogramme, correspond au masque, tandis qu'en-dehors de ces régions bruitées, le signal de parole est propre. L'hypothèse est qu'en fonction de leur positionnement sur le spectrogramme, les extraits de parole non-masqués conduisant à une reconnaissance correcte doivent être plus fréquents dans des mélanges intelligibles (peu bruités) que dans des mélanges inintelligibles (bruités). Les participants doivent reconnaître des phonèmes mélangés à du bruit, excepté dans de petits extraits temps-fréquences, le centre de ces bulles étant choisi aléatoirement dans le spectre à chaque essai. A partir des résultats expérimentaux, l'auteur a pu déduire des cartes d'intelligibilité par phonème, c'est-à-dire les régions temps-fréquences nécessaires à leur bonne reconnaissance (e.g. les hautes-fréquences juste avant la voyelle pour les consonnes plosives, les basses-fréquences pour détecter des voisements, etc.).

Dans une étude ultérieure, Mandel et al. (2014) décrivent un modèle computationnel entraîné à identifier les régions importantes pour l'intelligibilité de la parole dans le bruit, en partant de la même méthode de bulles auditives (Figure 17). Leur modèle est notamment capable de prédire l'intelligibilité de nouveaux mélanges de bruits, et de généraliser les résultats à de nouvelles prononciations d'un même mot, prononcé par le même locuteur ou par différents locuteurs. Cette étape de modélisation présente un intérêt non négligeable puisque, comme pour la corrélation inverse, le nombre d'essais nécessaires pour explorer avec suffisamment de détail tout l'espace physique du stimulus et obtenir un résultat convainquant peut être très important.

Enfin, il faut mentionner l'étude très récente de Venezia et al. (2016) qui porte aussi sur l'évaluation de l'intelligibilité de la parole à l'aide de bulles auditives, mais cette fois placées sur le spectre de modulation. Le spectre de modulation étant une représentation à deux dimensions des modulations spectro-temporelles observées sur le spectrogramme, les indices acoustiques de modulation sont sélectionnés sur l'ensemble du son et non pas sur des régions spectro-temporelles

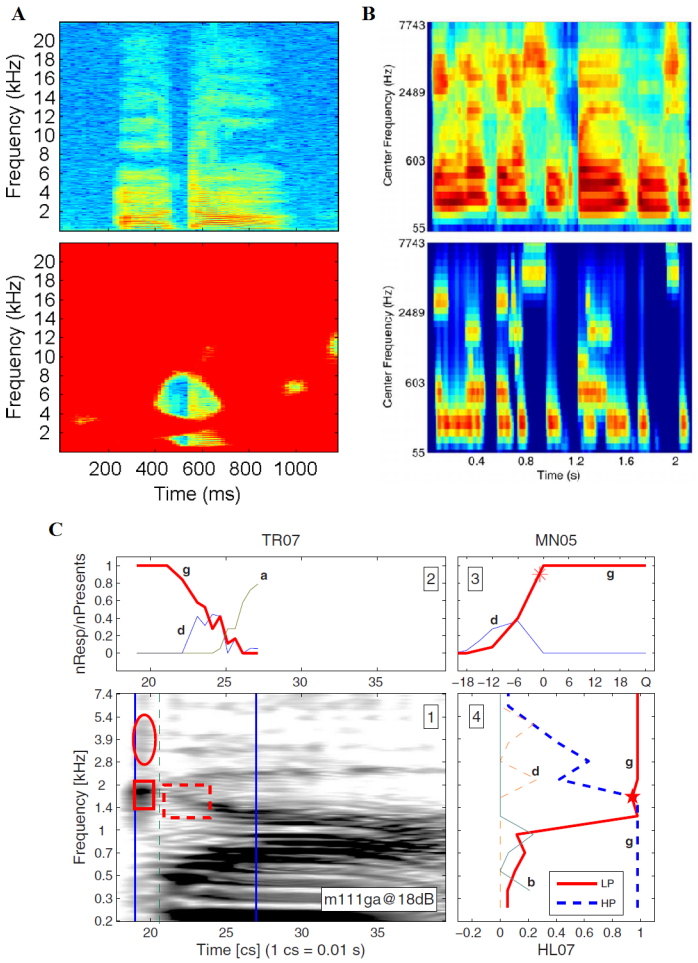


FIGURE 17 – Trois méthodes de sélection parcimonieuse d'indices acoustiques dans des sons de parole : bulles auditives, masques binaires, 3D-Deep Search. A : spectrogramme (en haut) et fonction d'importance temps-fréquence (en bas) du mot /ada/ avec la méthode de bulles auditives (le code couleur représente l'intensité croissante du bleu vers le rouge ; figure adaptée de Mandel, 2014) ; B : cochléogrammes 32 canaux d'une phrase prononcée par un locuteur danois (en haut) et d'un bruit modulé par le masque binaire correspondant obtenu avec 8 canaux (en bas ; figure adaptée de Wang et al., 2008) ; C : (1) AI-gramme (*articulation index*, i.e. visualisation simulant le traitement du système auditif périphérique) de la syllabe /ga/ : la ligne verticale en pointillée indique le début de la voyelle, les carrés en lignes continue et pointillée indiquent les évènements dominant et mineur respectivement, l'ellipse indique un indice provoquant des confusions avec la syllabe /ka/ ; (2) patterns de confusion en fonction du temps de troncation ; (3) patterns de confusion en fonction du SNR ; (4) patterns de confusion en fonction de la fréquence de coupure (figure adaptée de Li et al., 2010).

discrètes du spectre. Cette technique permet néanmoins de révéler des régions de modulation les plus importantes pour l'intelligibilité de différents contenus de parole, étant donné que les caractéristiques de la parole peuvent être efficacement représentées par ses modulations (Singh & Theunissen, 2003).

**Masques binaires.** La méthode des masques binaires présente des similitudes avec *Bubbles*, dans la mesure où elle consiste à isoler et conserver des régions spectro-temporelles discrètes d'un son représenté sous forme d'une matrice à deux dimensions temps-fréquence, et à supprimer le reste de l'information (Wang, 2005 ; Wang et al., 2008 ; Narayanan & Wang, 2010 ; Karadogan et al., 2010). Sur un masque binaire, la valeur 1 indique qu'on retient l'énergie acoustique de l'unité temps-fréquence correspondante, la valeur 0 indique qu'on la supprime (Figure 17). Ce procédé peut augmenter l'intelligibilité de la parole dans du bruit en dirigeant l'attention de l'auditeur (ou par reconnaissance automatique) sur des régions temps-fréquence pertinentes.

Wang et al. (2008) ont construit un masque binaire idéal en comparant, sur un mélange de parole et de bruit, l'énergie de la parole et celle du bruit dans des unités temps-fréquence locales. Le masque binaire idéal prend la valeur 1 lorsque, dans l'unité temps-fréquence correspondante, le rapport signal-sur-bruit (SNR, pour *Signal-to-Noise Ratio*) dépasse un seuil local, et prend la valeur 0 sinon. Dans ce cas, le masque est dit "idéal" car il requiert que parole et bruit soient disponibles séparément avant le mélange. Les auteurs vont plus loin en reprenant l'idée de Shannon et al. (1995) qui consiste à remplacer l'énergie de ces régions temps-fréquence discrètes par du bruit, avec un nombre restreint à la fois de trames temporelles et de canaux fréquentiels. Les participants reconnaissent presque parfaitement le contenu de parole à partir de ce masque qui échantillonne le signal en 16 canaux fréquentiel (banc de filtres gammatones) et avec un taux de trames temporelles de 100 Hz. La forme spectro-temporelle du signal résultant est drastiquement réduite à une variation binaire sans structure harmonique ni structure temporelle fine. Les patterns de variation d'énergie conservent une structure formantique très dégradée dans le contour général des variations spectro-temporelles, avec seulement quelques bandes de bruit modulées binairement par les enveloppes de parole.

En se basant sur ces résultats, Narayanan & Wang (2010) ont conçu et testé un système de reconnaissance automatique de la parole utilisant le même type de patterns binaires. Un réseau de neurones est entraîné sur un ensemble de masques binaires, puis testé sur le reste du corpus. Les performances obtenues sont supérieures à 85% pour toutes les conditions testées (différents types de bruits et de SNRs). Les patterns binaires semblent donc véhiculer de l'information phonétique utile aussi en reconnaissance automatique de la parole.

**3D-Deep Search.** Comme on l'a vu à travers déjà plusieurs études, les indices permettant la reconnaissance de la parole bénéficient d'un intérêt particulier à cause des nombreuses applications auxquelles elles peuvent conduire. Dans leur étude, Li et al. (2010) sont partis du constat que les technologies de synthèse de la parole nécessitent des a priori sur les indices utilisés. Cependant, ils notent que notre connaissance de ces indices est incomplète et inappropriée. Les auteurs ont donc développé une méthode, appelée "3D-Deep Search" (3DDS), permettant d'isoler des régions discrètes du spectrogramme nécessaires à l'identification de consonnes occlusives.

La méthode 3DDS se déroule en trois étapes indépendantes : (1) masquage du signal par du bruit blanc avec différents SNRs, (2) troncation temporelle à partir du début du son, (3) filtrages passe-bas et passe-haut (Figure 17). Pour chacune de ces modifications, les participants doivent identifier les sons présentés à chaque essai. Cette méthode permet ainsi de déterminer la position des indices acoustiques sur les trois dimensions indépendantes (temps, fréquence, intensité), à partir des réponses des participants.

Les auteurs ont mené trois expériences perceptives correspondant à chacune des trois dimensions mentionnées, avec pour stimuli originaux les syllabes /p, t, k, b, d, g/ + /a/ prononcées par plusieurs locuteurs. Pour l'expérience de troncation temporelle, la troncation commence juste avant le début du son et s'arrête à la fin de la consonne. La durée de la consonne est ensuite divisée en intervalles consécutifs et sans recouvrement avec trois durées (5, 10, et 20 ms), pour assigner plus de points où la parole change rapidement et moins de points ailleurs. Pour rendre la parole tronquée plus naturelle et enlever de possibles artefacts autour du début du son, les auteurs ajoutent du bruit blanc avec un

SNR de 12 dB. Pour l'expérience de filtrage fréquentiel, les auteurs ont choisi 19 filtres, dont 9 filtres passe-hauts, 9 filtres passe-bas, et 1 filtre large bande entre 250 et 8000 Hz. Les fréquences de coupure sont choisies de façon à simuler la division du spectre le long de la membrane basilaire. De même que pour l'expérience de troncation temporelle, les auteurs ajoutent du bruit blanc avec un SNR de 12 dB pour s'assurer que la parole filtrée n'avait pas d'autres composantes audibles en-dehors des bandes sélectionnées. Enfin, pour l'expérience de masquage, un bruit blanc est ajouté aux sons avec 8 différentes valeurs de SNR (entre -21 et +12 dB). Pour les trois expériences, à chaque essai, les participants doivent reconnaître par choix forcé la syllabe ou sélectionner l'option "bruit" si elle est trop bruitée.

Finalement, dans cette étude, les auteurs ont mis en évidence les caractéristiques des six consonnes occlusives testées : un court burst (20 ms) qui varie en fréquence centrale et en délai jusqu'au début du voisement (e.g. reconnaissance de la syllabe /ta/ grâce à un burst de 15 ms, hautes-fréquences au-dessus de 3 kHz, et 50-70 ms avant la voyelle). Cette caractérisation est robuste malgré la variabilité des prononciations. Lorsque le burst est supprimé, le score de reconnaissance passe de 100% au niveau de la chance, et les participants reportent alors majoritairement la syllabe /pa/. Les auteurs ont aussi montré de quelle manière des sons peuvent contenir des indices conflictuels (e.g. la syllabe /ka/ contient deux bursts hautes-fréquences du /ta/ et du /pa/) et générer une confusion si l'indice dominant est masqué par du bruit. L'ensemble de ces résultats appuie, au moins en reconnaissance de la parole, la conception discrète des indices de reconnaissance auditive, de façon similaire à la méthode *Bubbles* telle qu'elle a été employée en audition par Mandel (2013). Comme les précédentes méthodes, celle-ci aussi est fastidieuse et son applicabilité à des sons naturels plus divers reste à démontrer.

**Esquisses auditives.** Suied et al. (2013b) ont cherché à cibler directement les composantes spectro-temporelles nécessaires à la reconnaissance auditive de sons naturels en passant par l'utilisation de modèles auditifs. La technique de simplification proposée isole ces indices parcimonieux dans des esquisses auditives, tandis que le reste de l'information est supprimée. De même qu'une esquisse visuelle permet d'identifier un objet visuel naturel complexe, une esquisse auditive doit permettre de reconnaître un son naturel complexe, bien que le processus de

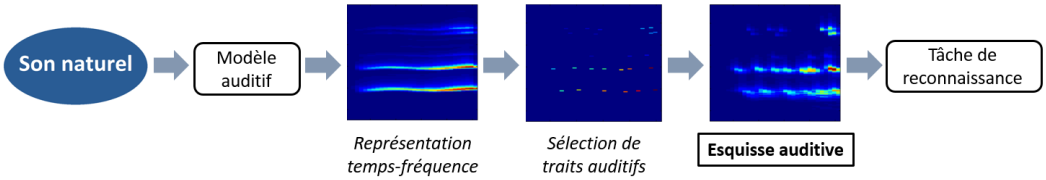


FIGURE 18 – Construction d'une esquisse auditive. Une représentation auditive d'un son naturel est générée, sur laquelle sont retenus seulement quelques composantes spectro-temporelles. L'inversion du modèle permet de resynthétiser l'esquisse auditive. Une évaluation psychophysique avec des participants humains permet d'évaluer l'efficacité des composantes sélectionnées. Figure adaptée de Suied et al. (2013b).

simplification introduise de fortes dégradations acoustiques sur le son original.

Le processus de simplification se déroule en trois étapes principales (Figure 18). Un modèle auditif permet d'obtenir une représentation qui met en valeur l'énergie acoustique effectivement utilisée par le système auditif et donc potentiellement les traits spectro-temporels favorisant la reconnaissance du son. Les auteurs ont testé deux représentations auditives, l'une basée sur un modèle du système auditif périphérique, l'autre incluant un modèle cortical plus complexe. Des pics d'énergie sont ensuite sélectionnés sur chaque représentation testée, sachant que leur nombre détermine le degré de parcimonie de l'esquisse. Celle-ci est finalement obtenue par resynthèse en inversant le modèle auditif utilisé.

Pour tester la validité de leur méthode, Suied et al. (2013b) ont effectué un test perceptif avec des sons de voix. La tâche était de reconnaître, à chaque essai et par choix forcé, l'émotion transmise dans une esquisse auditive de voix. Trois degrés de parcimonie, deux modèles auditifs (périphérique, cortical), et deux algorithmes de sélection de traits auditifs (un algorithme de décomposition parcimonieuse et un algorithme de sélection de maxima locaux) ont été testés. Les résultats montrent que même des esquisses très parcimonieuses peuvent être reconnues lorsqu'elles sont obtenues à l'aide d'un simple algorithme de sélection de maxima locaux sur un spectrogramme auditif (plutôt que sur une représentation corticale).

Cette méthode présente plusieurs avantages sur d'autres méthodes telles que celles que nous avons mentionnées précédemment et qui concernaient en premier lieu la parole. D'abord, elle est modulable (cf. Figure 18), avec la possibilité de choisir la représentation sur laquelle sont sélectionnés les traits auditifs tout comme l'algorithme permettant de les sélectionner. Ensuite, le degré de parcimonie

peut être ajusté en sélectionnant plus ou moins de traits auditifs. Cette flexibilité permet d'établir des points de comparaison directs avec d'autres méthodes, par exemple pour évaluer leur efficacité pour un même taux de simplification (e.g. Lemaitre et al., soumis). Aussi, il s'agit d'une méthode directe puisqu'une esquisse auditive peut être créée pour n'importe quel son sans un ajustement itératif avec un grand nombre d'essais comme le nécessitent les méthodes *Bubbles* ou 3DDS. Cependant, si cette méthode a été présentée sous la forme d'une preuve de concept, il reste à vérifier si elle est effectivement généralisable à n'importe quel type de son naturel avec la même efficacité, en fonction du modèle auditif utilisé pour la sélection des traits auditifs et du degré de parcimonie. C'est pourquoi nous avons repris et adapté cette méthode pour traiter cette question dans la première section de la partie "Contributions expérimentales".

**Bilan sur les méthodes de simplification de sons.** Les méthodes de simplification de sons présentées ici permettent chacune d'identifier des caractéristiques acoustiques discrètes importantes pour la reconnaissance auditive tout en suivant des stratégies différentes. A l'exception de la corrélation inverse qui vise à révéler l'information représentée, toutes les autres méthodes permettent quant à elles de révéler l'information potentielle transmise par le signal acoustique (et utile à la reconnaissance auditive). En majorité, ces études se placent dans le contexte de la reconnaissance de la parole (voir aussi Cooke, 2006 ; Kapoor & Allen, 2012), hormis celle de Suied et al. (2013b) qui s'était toutefois restreinte à la voix. Ces méthodes de simplification de stimuli sonores pourraient s'étendre à d'autres types de sons naturels et venir corroborer les résultats physiologiques et computationnels sur le traitement parcimonieux du système auditif (Lewicki, 2002 ; Olshausen & O'Connor, 2002 ; Olshausen & Field, 2004 ; Smith & Lewicki, 2006).

# 3 La rapidité de la reconnaissance auditive

Certains sons sont plus difficiles que d'autres à reconnaître et peuvent même nécessiter de se remémorer le son entendu une fois celui-ci disparu, indiquant que les indices qu'il contenait étaient ambigus ou en quantité insuffisante. Toutefois, dans la majorité des cas, le processus de reconnaissance auditive est très rapide. Il n'est cependant jamais instantané.

La méthode apparemment la plus simple pour mesurer le temps de reconnaissance serait de demander aux participants d'indiquer le moment à partir duquel ils ont reconnu le son. Ainsi, Donders (1969) a été l'un des premiers à théoriser sur la vitesse des processus mentaux en utilisant des temps de réaction (TRs). Il déduit le temps nécessaire pour aboutir à la conception d'un son connu en soustrayant le TR obtenu pour des sons inconnus avec celui obtenu pour le son connu. Selon lui, ce processus mental prendrait environ 40 ms dans le cas d'une voyelle cible. Cependant, cette méthode soustractive présuppose que chaque étape du traitement (e.g. détection d'indices auditifs dans le signal sonore, reconnaissance d'une cible donnée, action motrice pour valider la réponse) puisse être isolée avec des TRs car elles se succèderaient après une certaine durée dans un ordre sériel. De plus, la reconnaissance auditive a certainement lieu avant d'en arriver à compléter une certaine conception sur la nature du son.

Par la suite, un grand nombre d'auteurs ont pris le parti d'utiliser des paradigmes de masquage de reconnaissance pour s'intéresser spécifiquement au temps de reconnaissance, surtout en vision (e.g. Chun & Potter, 1995 ; Subramaniam et al., 2000 ; Keysers et al., 2001 ; Buffat et al., 2012) mais aussi en audition (e.g. Suied et al., 2013a). Il s'agit de présenter des sons en succession et sans recouvrement dont l'un d'eux est la cible à reconnaître. En partant du principe que chaque son est reconnaissable individuellement mais qu'il nécessite aussi un certain temps de traitement pour être reconnu, si tous les sons se succèdent trop rapidement les uns à la suite des autres, la reconnaissance de la cible devrait diminuer.

Avant de nous intéresser à la reconnaissance d'un son cible au sein d'une séquence de sons distracteurs, nous verrons dans un premier temps quelle est la durée minimale que doivent prendre chacun des sons de la séquence pour être reconnus isolément (paragraphe I.3.1). Après quoi, nous verrons comment le mas-

quage de reconnaissance peut permettre de délimiter une fenêtre de traitement auditif : l'image pré-perceptive (paragraphe I.3.2). Enfin, nous répertorions l'utilisation du masquage de reconnaissance pour l'évaluation du temps de traitement de sons courts (paragraphe I.3.3). Des données expérimentales nous permettront de donner une première estimation des limites du temps de traitement auditif.

## 3.1 Reconnaissance de sons courts

### 3.1.1 Durée minimale du signal sonore

L'étude de la durée de présentation minimale nécessaire pour reconnaître un son permet d'examiner sur quelle durée l'information doit s'accumuler pour parvenir à le reconnaître (Harding et al., 2007). Gray (1942) est l'un des premiers à s'être intéressé à la durée minimale requise pour reconnaître des sons et à montrer que le système auditif humain a la capacité de reconnaître des sons très courts. La procédure consiste à sélectionner aléatoirement des segments de 11 voyelles isolées, et à les faire entendre pour déterminer le seuil temporel d'intelligibilité. Un dispositif mécanique permet de faire varier la durée de l'intervalle, tandis que les voyelles sont prononcées par 3 hommes et 3 femmes au microphone et diffusées dans un haut-parleur. Les participants doivent indiquer l'identité de la voyelle prononcée, avec des extraits de 30 durées allant de 52 à 3 ms et présentées dans un ordre décroissant.

Pour des intervalles supérieurs à 18 ms, presque toutes les voyelles sont correctement identifiées, indépendamment des locuteurs et donc des périodes glottiques. Le nombre de voyelles bien reconnues diminue pour des durées plus courtes. Néanmoins les participants, familiers avec l'alphabet phonétique, peuvent reconnaître la majorité des voyelles dès 5 ms alors même qu'ils ne connaissent pas à l'avance quelles voyelles sont susceptibles d'être prononcées. Au travers des différences interindividuelles observées, l'auteur note que certains participants peuvent identifier des sons de 3 ms et que la reconnaissance est fréquente avec moins d'une période de la fréquence fondamentale (Figure 19).

Powell & Tosi (1970) ont aussi étudié la reconnaissance de sons courts de voyelles, cette fois avec une seule fréquence fondamentale de 125 Hz et 15 durées de 4 à 60 ms présentées en séries croissantes et décroissantes. Le seuil de recon-

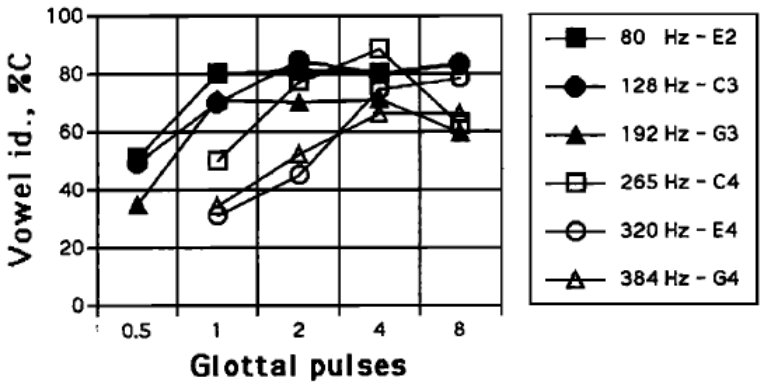


FIGURE 19 – Identification de voyelles en fonction du nombre de pulsations glottiques pour six fréquences fondamentales correspondant à six locuteurs avec les données expérimentales de Gray (1942). Source : Robinson & Patterson (1995a).

naissance pour chaque voyelle et chaque participant est déterminé en fonction de ses réponses correctes et incorrectes avec 10 répétitions par durée, par voyelle, et par série croissante ou décroissante. Le seuil de reconnaissance moyen obtenu est de 15.0 ms mais varie significativement en fonction des voyelles : de 27.2 ms pour /a/ à 9.3 ms pour /u/. La reconnaissance de voyelles ne semble donc pas dépendre simplement du nombre de périodes fondamentales mais plutôt du profil spectral des sons.

Afin d'observer comment le pattern formantique peut contribuer à la reconnaissance de sons courts de voyelles, Suen & Beddoes (1972) ont testé la discrimination de 6 voyelles de durée fixe (10 ms) prononcées par un homme. A titre de comparaison, ils citent notamment les études de Peterson (1939), Gray (1942), et Joos (1948), qui montrent que des voyelles de moins de 10 ms peuvent être reconnues. L'amplitude et la hauteur sont contrôlées en extrayant une période pour la répéter et ainsi resynthétiser chaque voyelle à 131 Hz. Les résultats obtenus montrent une meilleure reconnaissance pour des participants qui ont suivi un apprentissage rapide, pour atteindre une reconnaissance presque parfaite à la fin du test. Les voyelles les moins bien reconnues, /i/ et /u/, ont aussi les fréquences du premier formant les plus basses (260 et 270 Hz), tandis que c'est l'inverse pour les voyelles /a/ et /è/. Les auteurs citent des résultats antérieurs selon lesquels la durée minimale pour reconnaître un ton diminuerait lorsque sa fréquence augmente entre 50 Hz et 2 kHz. Avec des fréquences plus élevées, l'information se

concentre sur des durées plus faibles, et manifestement les participants utilisent cette information pour reconnaître les sons courts de voyelles.

Dans deux études complémentaires, Robinson & Patterson (1995b,a) ont testé respectivement des instruments de musique et des voyelles chantées avec des durées allant de 2.9 à 1952 ms. Cependant, les auteurs ont exprimé leurs résultats en termes du nombre de longueurs d'onde pour avoir des mesures comparables en fonction de la hauteur. Ils ont montré que l'identification d'une voyelle est proche du maximum des performances dès une longueur d'onde, donc même pour des durées inférieures à 10 ms. Les performances sont moins bonnes dans le cas de l'identification d'instruments. De plus, elles dépendent de l'expertise musicale et pas du nombre de longueurs d'ondes présentes dans le son. Autrement dit, le taux d'information minimum nécessaire pour reconnaître un son ne se comptabiliserait pas uniquement en fonction du nombre de longueurs d'ondes (y compris en ce qui concerne les fréquences formantiques des voyelles) mais plutôt en fonction de profils spectraux complexes. Dans le cas des voyelles par exemple, il pourrait s'agir d'un codage combinant les fréquences des pics formantiques (e.g. Formisano et al., 2008).

Plus récemment, Bigand et al. (2011) ont comparé la reconnaissance de sons de voix, d'extraits musicaux, et de sons de l'environnement, avec 5 durées sonores de 20 à 200 ms. Pour choisir les durées à tester, les auteurs prennent pour points de comparaison des tâches de reconnaissance plus complexes (e.g. reconnaître un morceau musical familier) ou des études électrophysiologiques, et n'ont donc pas testé des durées plus courtes. Les trois catégories sonores sont reconnues correctement dès 50 ms tandis que les voix et les instruments peuvent être reconnus dès 20 ms. Pour ces catégories sonores englobant plus de variabilité acoustique que des voyelles isolées, les variations des distributions spectrales, calculées sur les patterns d'excitation des stimuli et comparées dans chaque catégorie et entre les catégories, semblent contribuer aux résultats perceptifs.

Enfin, Suied et al. (2013a) et Suied et al. (2014) ont comparé des enregistrements d'instruments et de voyelles chantées en faisant varier la durée des sons ainsi que la position de début des segments sélectionnés dans les sons originaux. Les auteurs ont montré que leur reconnaissance est possible à des durées très courtes : jusqu'à 2 ou 4 ms en fonction des conditions. De plus, la sélection aléatoire de

la position de début des segments a un effet marginal plutôt que de présenter le début du son. En effet, l'attaque est informative seulement pour les sons de percussions, tandis que pour des sons de voix l'information formantique est présente durant toute la durée du son, et que pour des instruments à cordes frottées l'attaque est bruitée. La reconnaissance des sons est donc globalement possible sans l'attaque. Ces résultats sont cohérents avec ceux de Iverson & Krumhansl (1993) qui ont testé la perception de sons complets, avec, ou sans attaque. Par ailleurs, Suied et al. (2014) ont montré un avantage pour la voix par rapport aux instruments lorsque des stimuli de voix sont comparés à des distracteurs de voix (2 ms) et à des distracteurs de sons d'instruments (4 ms). La reconnaissance d'un son d'instrument est néanmoins aussi possible avec des durées très courtes : 8 ou 16 ms en fonction des conditions.

Globalement, l'ensemble de ces études semblent s'accorder sur la possibilité de reconnaître au-dessus de la chance des sons très courts, c'est-à-dire pour des durées inférieures à une dizaine de millisecondes. Pour des durées aussi courtes, le pattern spectral est nécessairement très impliqué dans la reconnaissance et doit bénéficier d'un codage cérébral très succinct et efficace (Occelli et al., 2015). Par ailleurs, les performances sont proches de la reconnaissance parfaite pour des durées de l'ordre de quelques dizaines de millisecondes. On remarque aussi que le seuil de durée pour reconnaître un son naturel est inférieur à celui reporté pour percevoir une sensation tonale avec des tons avant qu'ils ne soient perçus comme des clics (10 ms ; cf. Creel et al., 1970), ce qui laisse penser que la complexité spectrale des sons naturels contribue à l'efficacité de leur reconnaissance.

### 3.1.2 Durée minimale de percepts auditifs

Un percept visuel ou auditif a une durée minimale qui peut perdurer au-delà de la durée de présentation du stimulus, a fortiori dans le cas de stimuli très brefs. Nous verrons dans les paragraphes suivants que si la durée du percept s'allonge par rapport à la durée du stimulus, c'est certainement parce que la durée du traitement auditif perdure également après la présentation du stimulus pour réaliser une tâche auditive. Efron (1970b) a montré que des stimuli visuels et auditifs avec une durée de présentation inférieure à 120-130 ms produisent des percepts de la même durée que des stimuli avec une durée de présentation égale

à 120-130 ms et qu'il a appelée "durée critique".

Dans une étude complémentaire, Efron (1970a) a cherché à estimer plus précisément la durée de la perception produite par des stimuli de durées courtes. Avec des stimuli visuels et auditifs dont le début était asynchrone, puis en réduisant cette asynchronie, l'auteur a montré que l'erreur du jugement de simultanéité est très faible (de l'ordre de la milliseconde) et n'est pas affectée par la durée des stimuli. Dans une seconde expérience, les participants doivent cette fois juger la simultanéité de la fin d'un premier stimulus avec le début d'un second stimulus, tous deux d'une durée de 500 ms. Le délai obtenu est positif, d'environ 110 ms en visuel et 40 ms en auditif (pour 2 participants).

A partir de ces résultats, la durée critique (i.e. la durée perçue minimale) a été estimée à environ 120-130 ms indépendamment de la brièveté de la présentation (Efron, 1970b). De plus, le délai de perception de fin du premier stimulus a été estimé plus long que celui de début du second stimulus, et au-delà de la durée critique. L'auteur en a déduit que la fenêtre perceptive a une durée minimale de 120 ms, et peut s'étendre en vision à environ 240 ms, et en audition à environ 170 ms, pour des stimuli de durées inférieures à la durée critique.

Autrement dit, si des stimuli de quelques millisecondes contiennent suffisamment d'information pour être reconnus, la durée du percept auditif qui en résulte peut dépasser celle des stimuli. Cela laisse penser que le traitement de la reconnaissance, plutôt que de simplement accumuler continuellement les indices de reconnaissance jusqu'à atteindre un certain seuil où le stimulus serait reconnu, pourrait au contraire s'étaler sur la durée plus longue du percept pour donner sens à l'ensemble des indices tirés du stimulus en prenant un délai supplémentaire.

## 3.2 Stockage pré-perceptif de l'information auditive

### 3.2.1 Le masquage temporel pour l'étude du temps de traitement auditif

Les résultats d'Efron (1970b,a) font partie de ceux qui ont conduit Massaro (1972a,b, 1975, 1977) à proposer la distinction entre l'entrée sensorielle (le stimulus auditif) et son image pré-perceptive, qui perdure pour être traitée après la présentation du stimulus. En effet, d'après Massaro (1972a), le processus de recon-

naissance auditive requiert une analyse et une synthèse de l'information contenue dans l'entrée sensorielle, et par suite de conserver l'information dans une image auditive de façon pré-perceptive même après la fin du stimulus. Se posent les questions de la durée de vie de cette image, de sa vulnérabilité aux stimuli qui précèdent ou qui suivent, et de la possibilité d'un traitement parallèle ou séquentiel du flux d'information auditive.

Pour tester la validité de l'image auditive pré-perceptive, Massaro (1972a) s'appuie sur des expériences de masquage temporel : un masque suit et interfère avec une cible en fonction du délai entre la cible et le masque. En masquage temporel de détection, le participant doit indiquer la présence d'un stimulus ou non, tandis qu'en masquage temporel de reconnaissance, la tâche est d'identifier quelle alternative correspond à la cible. Si un stimulus court produit une image auditive pré-perceptive, un second stimulus décalé dans le temps (avant ou après la cible) doit interférer avec cette image et réduire la quantité d'information qu'elle véhicule. Massaro (1972a) mentionne notamment l'expérience d'Elliott (1967) : un ton de 10 ms est suivi d'un bruit de 100 ms après un intervalle de silence de 10 ou 100 ms. Le masquage augmente avec la diminution de l'intervalle temporel. Cette expérience permet de confirmer simplement que le masquage croît avec la diminution de l'intervalle de silence entre deux sons, et donc que toute l'information sonore n'est pas traitée quand bien même la durée totale du signal sonore serait la même.

Selon Massaro (1972a), il y a masquage car l'image du masque existe et sa durée dépasse celle de sa présentation pour masquer la cible. De plus, l'image dépassant la durée du stimulus ne doit pas différer qualitativement de l'image pendant la présentation du stimulus. Les résultats de masquage indiquent que l'image auditive du masque diminue le SNR de la cible et que le stockage pré-perceptif est facilement perturbé par d'autres entrées auditives. Cependant, ce type de masquage n'empêche jamais complètement la détection, et se révèle donc moins pertinent pour évaluer le décours temporel du traitement de l'information. Pour ces raisons, l'auteur privilégie des tâches de masquage de reconnaissance au masquage de détection (Massaro, 1970, 1971).

### 3.2.2 Durée de l'image pré-perceptive

De façon similaire à l'expérience de masquage de reconnaissance d'Elliott (1967), Massaro (1970) présente à des participants un ton de 20 ms avec une fréquence de 770 ou 870 Hz, suivi par un intervalle de silence entre 0 et 500 ms, lui-même suivi par un ton masquant de 820 Hz d'une durée de 500 ms. Les participants doivent juger si le (premier) ton cible est plus haut ou plus bas en fréquence que le suivant. Les performances augmentent avec l'augmentation de la durée de l'intervalle de silence jusqu'à un plateau à 250 ms. Une image auditive doit donc perdurer au-delà de la durée du ton cible pour améliorer les performances de traitement avec l'augmentation de l'intervalle de silence. Puis le masque termine le traitement de l'image pré-perceptive jusqu'à ce que la performance stagne pour un intervalle de silence de 250 ms. Ces 250 ms semblent correspondre à une estimation large de la durée suffisante pour l'intégration de patterns acoustiques simples sans qu'ils interfèrent entre eux (Massaro, 1972a). Autrement dit, cette fenêtre temporelle donne les limites de l'unité dans laquelle est stockée l'information en tant qu'image pré-perceptive. Si un pattern acoustique consécutif est présenté endessous de cette durée, soit il est intégré au premier pattern acoustique, soit il interfère avec lui.

La durée minimale de traitement auditif pourrait notamment expliquer la durée des voyelles dans la parole normale pour sa bonne intelligibilité. Il ne s'agirait pas tant d'accumuler des indices acoustiques supplémentaires, mais plutôt de laisser le temps à la parole d'être traitée. Dirks & Bower (1970) montrent par exemple que l'intelligibilité de la parole reste bonne même si le signal est interrompu toutes les 5 ou 50 ms par du silence (voir aussi Tallal & Piercy, 1974). A l'inverse, si les voyelles se suivent trop rapidement, l'intelligibilité baisse. Dans une étude non publiée (cf. Massaro, 1972a), Massaro utilise un paradigme de masquage de reconnaissance avec pour cibles les voyelles /i/ (e.g. "heat") et /I/ (e.g. "hit") de 20 ms, tandis que le masque, de 270 ms, est composé des voyelles /a/ (e.g. "hat") et /U/ (e.g. "put") de 45 ms et alternées. La performance d'identification augmente avec l'augmentation de l'intervalle de silence entre la cible et le masque (compris entre 0 et 500 ms), avec une très bonne identification à partir d'un intervalle interstimuli (ISI, *interstimulus interval*) de 80 ms environ, et par ailleurs meilleure que pour l'expérience équivalente avec des tons purs (cf. Massaro, 1970).

Un ISI d'une centaine de millisecondes semble laisser un temps suffisant pour reconnaître correctement un stimulus, en supposant que si cet ISI augmente ou que le stimulus est présenté seul, il sera reconnu parfaitement. Dans le cas contraire d'un ISI inférieur à une centaine de millisecondes, le stimulus est moins bien reconnu, voire pas reconnu du tout avec des réponses données au hasard. Il est cependant difficile de conclure en l'état sur la significativité de l'interaction entre la durée de l'ISI et de celle du stimulus (ou ses caractéristiques spectro-temporelles, i.e. la quantité d'information transmise), sachant qu'une centaine de millisecondes pour le traitement de la reconnaissance auditive laisse une assez grande marge temporelle pour tester des stimuli plus courts et toujours reconnaissables lorsque ceux-ci sont présentés isolément.

## 3.3 Seuils de rapidité du traitement auditif

### 3.3.1 Paradigmes de présentation séquentielle rapide de stimuli

En vision, les limites de l'efficacité temporelle du système visuel ont été étudiées avec des temps de réaction (e.g. Thorpe et al., 1996) et des paradigmes de masquage temporel de reconnaissance (e.g. Rolls et al., 1999). Toutefois, le paradigme de présentation visuelle sérielle rapide (RSVP, *Rapid Serial Visual Presentation*) a été plus spécifiquement élaboré pour étudier les limites temporelles de traitement du système visuel (e.g. Chun & Potter, 1995 ; Subramaniam et al., 2000 ; Keysers et al., 2001 ; Buffat et al., 2012). Le principe général de ce paradigme consiste à présenter une séquence d'items visuels (e.g. des images d'une catégorie distractrice) et à placer une cible (e.g. une image d'une catégorie cible) dans cette séquence, dans la moitié des essais et à une position aléatoire dans la séquence. Les séquences d'images sont présentées à des taux de présentation plus ou moins élevés, ce qui rend la tâche de reconnaissance de la cible plus ou moins difficile. Les performances de reconnaissance de la cible permettent d'en déduire le seuil du temps de traitement nécessaire pour cette reconnaissance, étant données les durées de présentation de chaque image et les catégories visuelles testées (Chun & Potter, 1995). Des équivalents en audition de ce paradigme sont apparus dans le contexte de l'étude de la reconnaissance auditive rapide dans plusieurs cas de figure.

Tout d'abord, la question de la rapidité du traitement auditif a été abordée avec des paradigmes similaires au RSVP pour tenter de détecter et d'expliquer des déficits auditifs plus complexes, tels que la dyslexie. Succinctement, la méthode générale consiste à présenter séquentiellement et rapidement des stimuli auditifs de complexité variable afin d'identifier des différences de temps de traitement auditif (abrégé RAP en anglais, pour *Rapid Auditory Processing*) susceptibles d'affecter le développement du langage (Tallal & Piercy, 1973 ; Benasich et al., 2002 ; Tallal, 2004). Par exemple, Tallal & Piercy (1973) ont présenté à leurs participants des séquences de deux tons complexes successifs imitant les paramètres acoustiques de phonèmes (75 ms chacun, F0 = 54 et 180 Hz) et avec des taux de présentation rapides (ISIs de 8 à 4062 ms). Les tâches proposées sont d'indiquer l'ordre de présentation des tons et s'ils sont les mêmes ou différents. Les participants normaux ont des scores au-dessus de 80% pour tous les ISIs et pour les deux tâches (au-dessus de 90% dès ISI = 15 ms). De plus, ils réalisent la tâche correctement dans le cas de séquences de 3 et de 4 sons. Ce n'est pas le cas des participants dyslexiques, indiquant que les capacités de traitement de caractéristiques acoustiques basiques sont limitées en fonction du taux de présentation, du nombre et de la nature des stimuli.

Ensuite, des auteurs ont proposé des tâches similaires au RSVP pour estimer les limites du traitement auditif en mettant en jeu des contraintes attentionnelles, qui sont à mettre en lien avec les études sur la cécité attentionnelle en vision (e.g. Raymond et al., 1992) ou en audition (e.g. Tremblay et al., 2005 ; Vachon et al., 2010). C'est le cas de Woods & Alain (1993), qui utilisent des tons de 10 ms présentés avec des ISIs allant de 40 à 200 ms. La cible correspond à des stimuli qui combinent les caractéristiques suivantes : en fréquence (250 ou 4000 Hz), en localisation (oreille gauche ou droite), et en intensité (3 à 15 dB plus intenses que les tons standards). Les participants doivent se concentrer sur certaines combinaisons de caractéristiques et désigner la cible en appuyant sur un bouton le plus rapidement possible, tandis que les PEs sont enregistrés pour observer l'effet de l'attention sur les dynamiques cérébrales. Les résultats comportementaux montrent que si les caractéristiques de fréquence et de position sont partagées avec les stimuli distracteurs antérieurs à la cible entre -400 et -40 ms, les réponses sont plus rapides, tandis qu'elles sont plus lentes si ces caractéristiques sont partagées

avec les stimuli distracteurs postérieurs à la cible entre +40 et +250 ms. Cet effet de *priming* semble se réduire dans le cas où le délai séparant le distracteur et la cible est trop important, ou si d'autres stimuli distracteurs ne partageant pas de caractéristiques avec la cible interférent entre temps. Par ailleurs, les formes d'onde des PEs suggèrent que les délais d'attention sélective aux caractéristiques acoustiques peuvent dépasser 300 ms (relevés sur les ondes différentielles, i.e. la soustraction des PEs pour les non-cibles avec les PEs pour les mêmes stimuli lorsque la condition testée les désigne comme cibles). L'estimation de la latence de la prise de décision est du même ordre de grandeur, soit environ 300 ms. La longueur de ces délais indique que les caractéristiques doivent être traitées en parallèle plutôt qu'en série.

Duncan et al. (1997) ont aussi réalisé une tâche attentionnelle avec une méthode s'inspirant explicitement du paradigme RSVP pour comparer les capacités de traitement entre les deux modalités visuelle et auditive (voir aussi Tremblay et al., 2005). Les participants doivent se concentrer sur deux flux de syllabes présentés simultanément soit dans la modalité auditive, visuelle, ou visuo-auditive, avec un taux de présentation fixe, des items de 150 ms (120 ms dans la modalité visuelle pour limiter les performances) séparés de 100 ms. Un ISI compris entre 125 et 1375 ms sépare les cibles présentées respectivement dans chaque flux. Les performances diminuent lorsque la deuxième cible suit de près la première (ISI = 125 ms), uniquement lorsque les deux flux sont présentés dans la même modalité, et non lorsque les deux flux sont respectivement dans les modalités visuelle et auditive. Si l'on élimine des contraintes attentionnelles, un ISI d'une centaine de millisecondes semble largement suffisant pour traiter chaque item d'une séquence. Cependant, encore beaucoup de facteurs acoustiques peuvent jouer sur la reconnaissance rapide de stimuli auditifs, comme la durée de chaque item (s'ils sont tous très longs, il n'y a plus de présentation rapide à proprement parler) ou leur complexité acoustique, et n'ont pas encore été étudiés.

Ces études ont montré la faisabilité de la mesure du seuil de rapidité du traitement auditif, bien qu'elles l'appliquent chacune avec leur propre méthode et des stimuli basiques ou avec peu de variabilité acoustique, ou encore dans un contexte attentionnel. Suied et al. (2013a) ont proposé le paradigme de présentation séquentielle audio rapide (RASP, pour *Rapid Audio Sequential Presentation*), analogue

direct du RSVP en audition, afin d'étudier plus systématiquement les seuils de rapidité du traitement auditif dédié à la reconnaissance auditive. Le paradigme RASP consiste à présenter rapidement des séquences de sons courts d'une catégorie donnée après avoir vérifié qu'ils peuvent être reconnus lorsqu'ils sont présentés isolément. Dans la moitié des essais, une cible d'une catégorie sonore différente des distracteurs est placée parmi eux, à une position aléatoire dans la séquence (Figure 20). Les auteurs ont utilisé des stimuli naturels de voix comme cibles et d'instruments comme distracteurs, tous égalisés en hauteur, durée, et niveau sonore. Restent donc uniquement les indices de timbre pour réaliser correctement la tâche. Les résultats obtenus montrent une diminution des performances de reconnaissance de la cible pour des taux de présentation élevés, ainsi que lorsque les sons de la séquence étaient plus courts (16 vs. 32 ms). Pour les deux durées sonores testées, les performances de reconnaissance de la cible sont au-dessus de la chance à un taux de présentation de 30 Hz, ce qui laisse un ISI = 17 ms dans le cas de sons de 16 ms, et un ISI = 1 ms dans le cas de sons de 32 ms. La difficulté de la tâche est donc liée à la fois à la durée qui sépare les stimuli mais aussi à la durée des items sonores. De plus, pour ces durées, les différences de reconnaissance de sons isolés sont encore assez importantes entre des sons de 16 et 32 ms. Dans le paragraphe 3.3.2, nous utilisons ces données pour estimer le temps de traitement auditif de sons naturels en tenant compte de l'ensemble de ces facteurs. Par ailleurs, nous avons approfondi expérimentalement la question de l'implication, dans le traitement auditif rapide, de la nature des stimuli auditifs (cf. la deuxième section de la partie "Contributions expérimentales").

On peut aussi mentionner le paradigme RSAP (pour *Rapid Serial Auditory Presentation*) proposé indépendamment par différents auteurs en perception de la parole, mais avec par conséquent des items sonores relativement longs (200 ms ; Yund et al., 1999 ; Franco et al., 2015). De plus, Yund et al. (1999) ont comparé les performances entre une condition de flux monaurale et une condition de flux dichotique, comme pour d'autres tâches d'attention sélective (e.g. Vachon et al., 2010). Dans l'étude de Franco et al. (2015), les auteurs ont testé spécifiquement l'apprentissage statistique des régularités entre des syllabes présentées rapidement, d'après leur temps de détection en fonction de leur position dans un mot. Néanmoins, dans tous les cas, les auteurs pointent justement la polyvalence

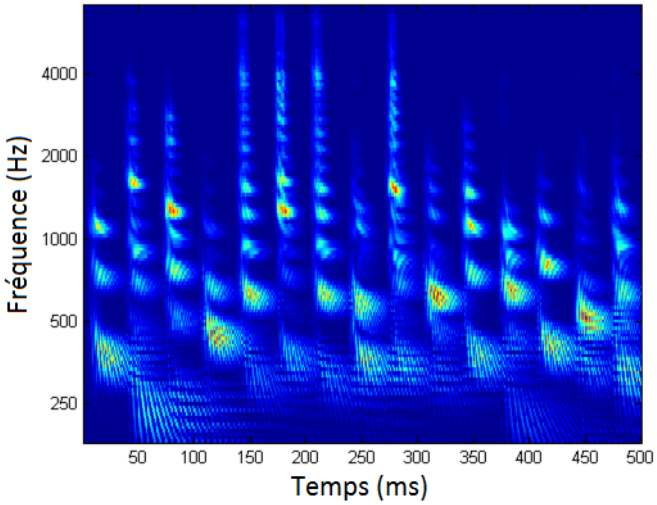


FIGURE 20 – Paradigme de présentation séquentielle audio rapide. Spectrogramme d'une séquence de sons d'instruments présentés rapidement et parmi lesquels pouvait être placée une cible, i.e. un son de voix, à une position aléatoire dans la séquence.

de la méthode de présentation rapide de stimuli auditifs, résultant surtout du choix des stimuli laissé aux expérimentateurs en fonction des objectifs de l'étude. On profitera de cette polyvalence dans le cadre de nos expériences.

### 3.3.2 Estimation quantitative du temps de traitement auditif de sons naturels

Les paradigmes expérimentaux impliquant une mesure du temps de traitement auditif diffèrent d'une étude à l'autre suivant leur problématique spécifique : implications attentionnelles dans le traitement auditif, déficits sensoriels, reconnaissance de la parole, reconnaissance de caractéristiques acoustiques basiques ou complexes. Finalement, les données expérimentales concernant le temps de traitement de sons naturels, sans l'implication d'autres facteurs (e.g. attentionnels), sont assez restreintes. On peut citer celle de Massaro (cf. Massaro, 1972a) : des voyelles de 20 ms peuvent être reconnues si elles sont séparées d'un masque distant d'au moins 20 ms ; et celle de Tallal & Piercy (1973) : deux phonèmes de synthèse successifs de 75 ms sont reconnus correctement dès le plus petit ISI testé, de 8 ms. Les données de Suied et al. (2013a), qui sont en accord avec les précédentes, sont aussi plus étoffées avec l'utilisation de sons naturels (voix et instruments) et

en fonction de plusieurs conditions, incluant la durée des sons dans la séquence et le taux de présentation. Ces données nous permettent d'estimer les latences du traitement pré-perceptif.

Avant cela il faut préciser que des données physiologiques indiquent que l'information auditive est très rapidement transmise jusqu'à A1. Chez l'humain, l'ordre de grandeur mentionné dans la littérature du délai avec lequel le signal sonore atteint A1 est d'environ 15 ms (e.g. Liegeois-Chauvel et al., 1994), bien que le décours temporel de cette propagation dans d'autres aires auditives puisse dépendre du type de sons d'après des expériences chez le singe (Lakatos et al., 2005a).

D'après Suied et al. (2013a), une cible peut être reconnue dans une séquence de sons de 32 ms présentés à 30 Hz (soit 1 son toutes les 33 ms). Or, lorsqu'un son de 32 ms est transmis à t = 0 ms dans les voies auditives, la fin du son atteint A1 à 32 + 15 = 47 ms (Figure 21). Avant qu'il n'ait fini d'être transmis à A1, le son suivant a déjà commencé à être émis, et ce, depuis l'instant t = 33 ms. Donc pendant 47 - 33 = 14 ms, soit presque la moitié de la durée sonore (délai *a* sur la Figure 21), les deux sons successifs ont transité simultanément dans les voies auditives antérieures à A1. On en déduit que des informations auditives de natures différentes peuvent transiter avant A1 et commencer à subir des pré-traitements dans les voies auditives sans se perturber mutuellement à cause d'un recouvrement potentiel de l'information, donc que la transmission auditive jusqu'à A1 semble relativement continue. Pour simplifier la suite des estimations, on considèrera que le traitement avant A1 ne génère pas de délai supplémentaire au temps de transmission.

Ensuite, les résultats de Suied et al. (2013a) indiquent qu'une cible peut être reconnue dans une séquence de sons de 16 ms présentés à 30 Hz, mais pas à 42.4 Hz (soit 1 son toutes les 24 ms). Dans le cas du taux de présentation de 30 Hz, il n'y a pas de recouvrement entre la fin du son qui atteint A1, à 16 + 15 = 31 ms, et le début d'émission du son suivant, à 33 ms. Par contre, dans le cas du taux de présentation de 42.4 Hz, la fin du premier son atteint A1 à 16 + 15 = 31 ms, tandis que le son suivant est émis dès l'instant t = 24 ms. Depuis donc 31 - 24 = 7 ms, presque la moitié du son suivant a déjà été émis sans avoir encore atteint A1 (comme précédemment, il s'agit du délai *a* sur la Figure 21). On a vu que dans le cas de sons de 32 ms, ce recouvrement avant A1 ne semble pas

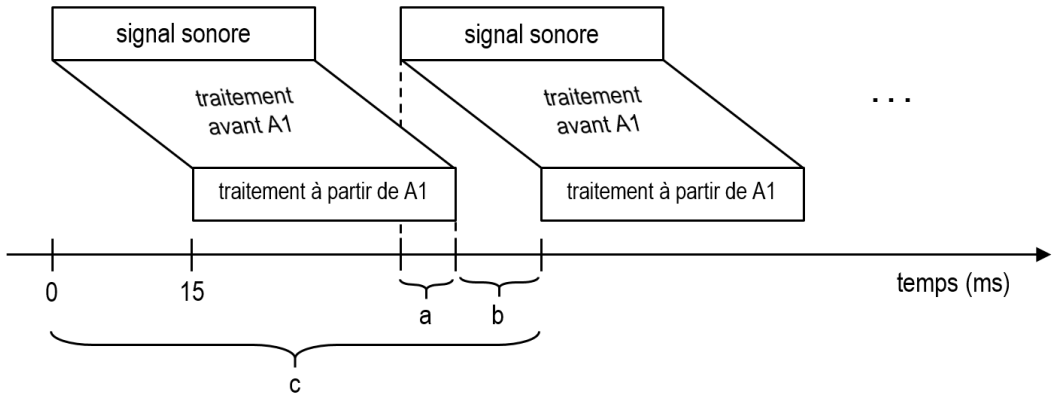


FIGURE 21 – Décours temporel du traitement de séquences auditives présentées rapidement. On pose que 15 ms sont nécessaires au signal pour atteindre A1 (voir texte). Le délai *a* indique le délai entre le moment où un premier signal auditif complet est arrivé dans A1 et le début d'émission d'un deuxième stimulus acoustique. Le délai *b* indique le délai entre le moment où un premier signal auditif complet arrive dans A1 et le moment où un deuxième signal auditif entre dans A1. Le délai *c* indique le temps de traitement total disponible pour traiter un signal auditif (y compris avant A1) avant qu'un signal auditif consécutif n'entre dans A1.

affecter la reconnaissance de la cible, on suppose que c'est également le cas avec des sons de 16 ms et que ce n'est donc pas la raison pour laquelle des sons de 16 ms présentés à 42.4 Hz ne sont pas reconnus (pour rappel, il n'est pas possible de tester des séquences de sons de 32 ms avec un taux de 42.4 Hz). Il y a pourtant eu un recouvrement de l'information qui a généré cette baisse des performances. Il ne semble pas avoir eu lieu avant A1 (d'après les données obtenus avec des sons de 32 ms). Les données obtenues avec des sons de 16 ms nous incitent donc plutôt à penser que le traitement auditif doit se prolonger dans ou au-delà de A1 pour parvenir à la reconnaissance des sons.

Le temps de traitement total de chaque son correspond (en supposant un masquage nul avant A1) à la durée avant que le son suivant dans la séquence n'apparaisse dans A1, soit, pour le taux de traitement de 30 Hz pour lequel des sons de 16 et 32 ms sont reconnus, à $33 + 15 = 48$ ms (délai *c* sur la Figure 21). En réalité, cette estimation est certainement surévaluée car il n'est pas possible de tester des sons de 32 ms à des taux de présentation supérieurs à 30 Hz, et des sons de 16 ms peuvent certainement être reconnus à des taux de présentation compris entre 30 et 42.4 Hz (les scores commencent à s'éloigner du niveau de chance dès

42.4 Hz ; à 42.4 Hz, le temps de traitement total est de 24 + 15 = 39 ms).

En d'autres termes, une fenêtre d'une cinquantaine de millisecondes semble suffire à reconnaître un son naturel, bien que cette estimation dépende du type de sons testés et de la quantité d'information auditive transmise par chacun des sons de la séquence (plutôt à surévaluer dans le cas de sons simples par rapport à des sons complexes ; Massaro, 1972a ; Lakatos et al., 2005b). De plus, le temps de traitement minimal ne semble pas pouvoir descendre bien en-dessous de 25-30 ms. Cette estimation de la durée de la fenêtre de traitement auditif (25-50 ms) reste toutefois très inférieure à des estimations obtenues par enregistrements cérébraux, bien que certains auteurs observent des réponses différenciées en fonction des catégories sonores à des durées parfois inférieures à 100 ms (cf. paragraphe I.1.4.3 ; Murray et al., 2006 ; Charest et al., 2009). Les différenciations cérébrales plus tardives pourraient s'expliquer par d'autres traitements des catégories sonores, ultérieurs à l'étape de reconnaissance auditive.

## 3.4 Bilan sur les temporalités auditives

L'ensemble des données analysées dans cette section permettent de discerner le décours temporel de l'information auditive lorsqu'elle est soumise à plusieurs contraintes temporelles complémentaires. Tout d'abord, le système auditif permet de reconnaître des sons complexes courts isolés, d'une durée de moins de 10 ms. Autrement dit, des sons de cette durée contiennent suffisamment d'information pour être reconnus. Pour autant, le système auditif nécessite un certain temps pour les traiter.

Si l'information auditive circulait en continu et bénéficiait d'un traitement instantané, les caractéristiques auditives ne se rencontreraient jamais et il n'y aurait par conséquent pas de recouvrement de l'information susceptible de générer du masquage de reconnaissance. Ce n'est pas le cas : pour des taux de présentation trop élevés, des latences de traitement semblent entraîner ce masquage de reconnaissance. Les résultats obtenus avec des paradigmes de présentation séquentielle audio rapide semblent aller dans le sens des hypothèses de fenêtres de traitement pré-perceptif de Massaro (1972a), qui suggèrent que le signal auditif perdure pour allonger la durée de son traitement au-delà de la durée du stimulus.

Plus particulièrement, des données expérimentales nous ont permis d'évaluer la rapidité du traitement auditif dans le cas de la reconnaissance du timbre de sons naturels (Suied et al., 2013a). Ces données indiquent que des sons courts peuvent être reconnus si un intervalle de silence suffisamment long les suit ; pour des taux de présentation de séquences sonores trop élevés, le son suivant de la séquence interfère avec le précédent qui n'est pas reconnu ; enfin, les performances pour un même taux de présentation sont meilleures si les sons de la séquence sont plus longs. Ces données nous ont permis d'estimer à environ 25-50 ms le temps de traitement de sons naturels courts (16-32 ms), ce qui est plus rapide que des estimations obtenues avec d'autres paradigmes de masquage de reconnaissance (cf. Massaro, 1972a) ou à partir d'enregistrements de l'activité cérébrale (e.g. Murray et al., 2006 ; Charest et al., 2009). En outre, cette efficacité de la reconnaissance auditive semble bénéficier des pré-traitements dans les voies auditives périphériques, puis, à partir de A1, d'une réanalyse des caractéristiques auditives pendant une durée dépassant celle du signal auditif jusqu'à parvenir à la reconnaissance du son.

Malgré ces résultats, la question de l'influence de la nature des stimuli sur leur temps de traitement demeure. Dans le cas de sons naturels comme des sons de voix ou d'instruments, les caractéristiques acoustiques transmises dans des extraits sonores courts semblent à première vue équivalentes et devraient donc être traitées avec la même efficacité. Nous nous sommes intéressés expérimentalement à la rapidité du traitement auditif de sons courts de voix et d'instruments (cf. la deuxième section de la partie expérimentale). Nous verrons qu'elle diffère entre les deux catégories sonores, et favorise (en termes de temps de traitement) la reconnaissance des voix.

# II Contributions expérimentales

## 1 Esquisses auditives : reconnaissance de sons parcimonieux

Ce travail a fait l'objet de communications orales :

- lors de la journée des doctorants de l'Institut de Recherche Biomédicale des Armées (IRBA) : "Esquisses visuo-auditives : représentations parcimonieuses d'objets bimodaux basées sur des modèles perceptifs", le 17 décembre 2013 ;
- lors de la journée des doctorants de la Direction Générale de l'Armement (DGA) : "—", le 30 janvier 2014 ;
- lors de la journée des doctorants de l'Ecole Doctorale Cerveau, Cognition, Comportement (ED3C) : "Visuo-auditory sketches : sparse representations of bimodal objects based on perceptive models" (poster), le 12 mars 2014 ;
- lors de la 6$^{e}$ biennale de la recherche du Service de Santé des Armées (SSA) : "—" (poster), le 26 juin 2014 ;
- lors de la Journée des Jeunes Chercheurs en Audition, Acoustique musicale et Signal Audio (JJCAAS) : "Esquisses visuo-auditives : représentations parcimonieuses d'objets bimodaux basées sur des modèles perceptifs" (poster), le 2 juillet 2014 ;
- lors de la journée des doctorants de l'ED3C : "Reconnaissance d'esquisses auditives : approche psychophysique et modélisation" (poster), le 11 mars 2015 ;
- lors de la 170$^{e}$ conférence de l'*Acoustical Society of America* (ASA) : "Acoustic and auditory sketches : recognition of severely simplified natural sounds by human listeners", le 5 novembre 2015.

Il a par ailleurs donné lieu à la publication suivante :

- Vincent Isnard, Marine Taffou, Isabelle Viaud-Delmon, Clara Suied, *Auditory sketches : very sparse representations of sounds are still recognizable*, PLoS ONE, 11(3), 2016.

## 1.1 Résumé

Des sons très divers peuvent être reconnus sur la base d'une quantité limitée d'indices, tels que ceux contenus dans le timbre. L'objectif de cette étude est de préciser quels sont les caractéristiques spectro-temporelles sous-jacentes à la reconnaissance auditive. Pour cela, un grand nombre de sons très divers ont été simplifiés en esquisses acoustiques et auditives, de façon à évaluer jusqu'à quel point il est possible de supprimer de l'information tout en maintenant la reconnaissance auditive au-dessus de la chance.

Le procédé de simplification consiste à sélectionner des pics d'énergie sur un spectrogramme acoustique ou auditif, avec trois niveaux de simplification (faible, moyen, élevé). L'information non-parcimonieuse restante est supprimée. Les sons originaux proviennent de 4 catégories sonores : instruments, oiseaux, véhicules, et voix. La tâche des participants est d'indiquer, à chaque essai, à quelle catégorie appartient le son, suivant une procédure 4-AFC (*Alternative Forced-Choice*).

Pour analyser les résultats, nous avons adapté un modèle récent de la théorie de détection du signal, afin de dissocier la sensibilité (scores $d'$) du biais pour chaque catégorie sonore (les tenants et les aboutissants de cette démarche sont décrits en compléments d'analyses, cf. paragraphe II.1.3). Globalement, des sons très simplifiés peuvent encore être reconnus au-dessus de la chance par les participants. Les performances des participants sont par ailleurs fortement corrélées avec les distances auditives, calculées entre des représentations auditives des sons simplifiés (STEPs ; Moore, 2003), décrivant l'appartenance des stimuli à une catégorie sonore comparée aux autres catégories (voir aussi les compléments d'analyses, cf. paragraphe II.1.4).

L'ensemble des résultats obtenus dans cette étude suggèrent que la reconnaissance auditive est un processus perceptif très robuste qui se base sur des indices spectro-temporels parcimonieux et dont la variabilité peut être capturée par les distances auditives pour expliquer cette reconnaissance robuste.

## 1.2 "Auditory sketches : very sparse representations of sounds are still recognizable"

Auteurs : Vincent Isnard, Marine Taffou, Isabelle Viaud-Delmon, Clara Suied.
Article publié dans la revue PLoS ONE, 11(3) : e0150313, 2016.

PLOS ONE

RESEARCH ARTICLE

# Auditory Sketches: Very Sparse Representations of Sounds Are Still Recognizable

**Vincent Isnard[1,2]*, Marine Taffou[1], Isabelle Viaud-Delmon[1], Clara Suied[2]***

**1** Espaces Acoustiques et Cognitifs, Sorbonne Universités, UPMC Univ Paris 06, CNRS, IRCAM, STMS, Paris, France, **2** Département Action et Cognition en Situation Opérationnelle, Institut de Recherche Biomédicale des Armées, Brétigny-sur-Orge, France

* vincent.isnard@ircam.fr (VI); clara.suied@irba.fr (CS)

CrossMark
click for updates

OPEN ACCESS

**Citation:** Isnard V, Taffou M, Viaud-Delmon I, Suied C (2016) Auditory Sketches: Very Sparse Representations of Sounds Are Still Recognizable. PLoS ONE 11(3): e0150313. doi:10.1371/journal.pone.0150313

**Editor:** Trevor Bruce Penney, National University of Singapore, SINGAPORE

**Received:** August 7, 2015

**Accepted:** February 11, 2016

**Published:** March 7, 2016

**Copyright:** © 2016 Isnard et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Funding:** This study was supported by project DGA-PDH-1-SMO-3-0808 and French program DEFISENS from the CNRS MI, project Supplé-Sens.

**Competing Interests:** The authors have declared that no competing interests exist.

## Abstract

Sounds in our environment like voices, animal calls or musical instruments are easily recognized by human listeners. Understanding the key features underlying this robust sound recognition is an important question in auditory science. Here, we studied the recognition by human listeners of new classes of sounds: acoustic and auditory sketches, sounds that are severely impoverished but still recognizable. Starting from a time-frequency representation, a sketch is obtained by keeping only sparse elements of the original signal, here, by means of a simple peak-picking algorithm. Two time-frequency representations were compared: a biologically grounded one, the auditory spectrogram, which simulates peripheral auditory filtering, and a simple acoustic spectrogram, based on a Fourier transform. Three degrees of sparsity were also investigated. Listeners were asked to recognize the category to which a sketch sound belongs: singing voices, bird calls, musical instruments, and vehicle engine noises. Results showed that, with the exception of voice sounds, very sparse representations of sounds (10 features, or energy peaks, per second) could be recognized above chance. No clear differences could be observed between the acoustic and the auditory sketches. For the voice sounds, however, a completely different pattern of results emerged, with at-chance or even below-chance recognition performances, suggesting that the important features of the voice, whatever they are, were removed by the sketch process. Overall, these perceptual results were well correlated with a model of auditory distances, based on spectro-temporal excitation patterns (STEPs). This study confirms the potential of these new classes of sounds, acoustic and auditory sketches, to study sound recognition.

## Introduction

Although human listeners can apparently recognize very easily and with no effort very diverse sound sources in their surrounding environment, the literature focusing on the recognition of natural sounds and on the features used by the listeners to recognize them is relatively scant (e.g. [1,2]). Yet, as it has been argued for a long time in vision [3], natural stimuli may recruit

specific mechanisms derived from adaptation to natural environments. The specificity of natural sounds has recently been highlighted: they can capture attention in an auditory-visual setting [4], and a few milliseconds are enough to recognize them [5–7]. The majority of studies focusing on the features used by the auditory system for the representation of natural sounds comes from brain imaging techniques. Until recently, a fairly accepted model of cortical processing of natural sounds implied a hierarchical temporal stream, from the encoding of low-level features to a high-level and more abstract category encoding [8,9]. It has been shown and developed for voice sounds [10], tool vs. animal sounds [11], or songbirds, animal sounds, speech and musical instruments [12]. Taking carefully into account some low-level acoustic features, other models have been proposed, involving distributed neural representations in the entire human auditory cortex for both low-level features and abstract category encoding [13–15]. They also showed that a complex spectro-temporal pattern of features represents more accurately the auditory encoding of natural sounds than a purely spectral or temporal approach (see [16] for animal sounds only; [13,17]; see also [18] for a computational and psychophysical approach). In particular, Moerel et al. [14] found that the voices and speech regions also responded to low-level features, with a bias toward low-frequencies that are characteristic of the human voices. This result is coherent with the theoretical approach proposed by Smith and Lewicki [19], which shows that the auditory code is optimum for natural sounds and especially suggests that the acoustic structure of speech could be adapted to the physiology of the peripheral auditory system.

As evidenced in this theoretical approach [19], or in physiological studies [20], not all information in a sound is useful to encode natural sounds: sparse coding based on the time/frequency properties of the auditory system is a highly efficient coding strategy. In perceptual studies, this is a well-known fact, not all information is useful for a given listening task. As primarily shown in speech studies, the auditory signal can be drastically distorted, or modified, and still be recognizable [21,22]. More recently, similar noise-band vocoder method as the one used by Shannon et al. [21], which removed most of the fine frequency information, has been applied to environmental sounds [2]. Although the effect is less spectacular with environmental sounds than with speech, the authors also showed that environmental sounds are resilient to a large amount of distortions. However, all of these transformations are not particularly sparse.

Recently, Suied et al. [23] have tackled the question of the sounds features that carry the most substantial information for a listener by introducing a new behavioral method: auditory sketches. Sketches are sparse representations of sounds that are severely impoverished, but still afford good performance on a given perceptual task, for example, recognition (see Fig 1 for an

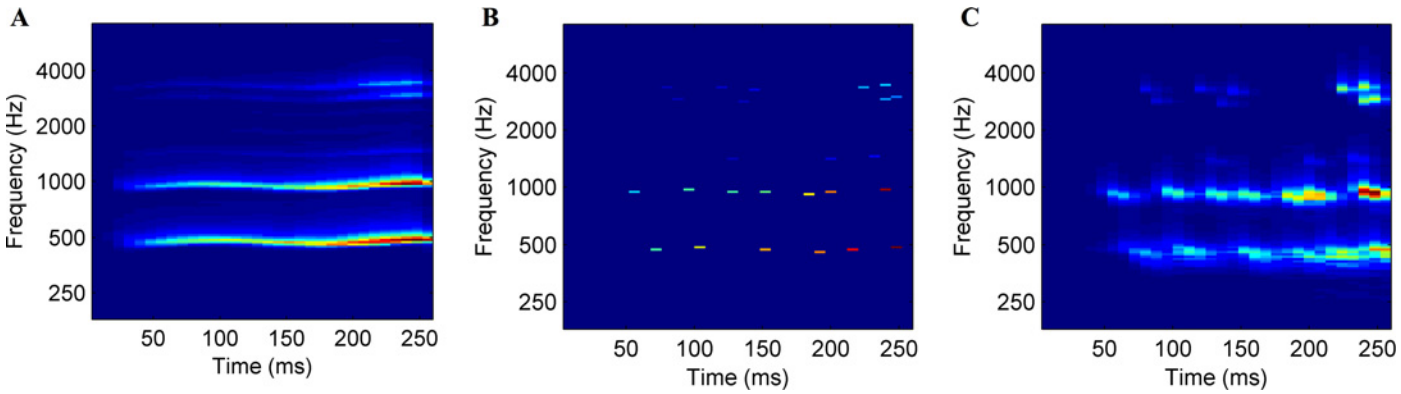


**Fig 1. The sketch process.** (A) The panel shows the first step, i.e. the time-frequency representation of a sound; here, the auditory spectrogram of the original sound (a voice sound of a female alto singer singing an /a/ on a B4). (B) The panel represents the sparsification algorithm: the 25 highest peaks in the signal are selected, corresponding to the 100 feat./s sparsification level. Based on this sparse representation, a sketch sound is then resynthesized. (C) The panel displays the auditory spectrogram of this sketch sound.

doi:10.1371/journal.pone.0150313.g001

illustration of the sketch process). In order to create an auditory sketch, the first necessary step is to choose the appropriate representation of the sound. Because auditory representations are inspired by the physiology of the auditory system, they should contain the features relevant to perception. The second step consists in selecting sparse features on these representations. Again, if the auditory representation is efficient, the selection mechanism should be drastic. Finally, the representation was then inverted to give rise to a new sound: the auditory sketch. A proof of concept of the auditory sketch process was done in a first study [21], by studying the recognition of different emotions in voices. For this particular task, good recognition of the auditory sketches was observed. Nevertheless, the hypothesis that biologically plausible representations are better suited for efficient sketches than simple time-frequency representations remains to be tested. In addition, an extension of these results to more various sound sources is needed in order to generalize the sketch process. Finally, no attempt to model the acoustic features present in the sketches, which enable a good recognition, had been made.

The aim of the present study was to test the auditory sketches idea with a large, diverse, but still controlled set of sounds, and, by this mean, try to untangle the acoustic features used by the listeners to recognize these sounds. Listeners were presented with sketch sounds and had to recognize them, by indicating the category to which they belong. Four sound categories were used: voices, instruments, birds, and vehicles sounds. The original sounds of two of these four categories (voices and instruments) were equalized in pitch, loudness, and duration. At least for these two categories, listeners were thus left with only timbre cues to perform the task (see [24–26]). Adding the other two categories (birds and vehicles) ensured a sufficiently heterogeneous set of sounds in order to test the generality of the sketch process, but still controlled, by measuring some of the classical timbre cues highlighted in previous studies on timbre perception, like spectral centroid (see [27]), or in previous imaging studies on the encoding of natural sounds by the auditory cortex, like the Harmonic-to-Noise Ratio, or HNR (e.g. [13,15]). To test whether biologically grounded representations lead to more efficient sketches, we compared two time-frequency representations: an auditory spectrogram (see [28]), and a classical acoustic spectrogram, which performs a Fourier transform. The two resulting classes of sketches sounds will be referred to as 'auditory sketches' and 'acoustic sketches'. The same sparsification levels as in the first study (10, 100, and 1000 features/second—the 'features' being here the energy peaks) were also tested, in order to explore how the recognition evolves (positively, we hypothesize) with the increase of the number of features.

## Experiment

### Methods

**Participants.** Fourteen individuals (6 men and 8 women; mean age 24.4 ± 2.7) took part in this experiment. None of the individuals reported having hearing problems. They all provided written informed consent to participate in the study. The Institutional Review Board of the French Institute of Medical Research and Health ethically approved this specific study prior to the experiment (opinion n°14–151). All participants were paid for their participation.

**Original and sketch sounds.** 120 original sounds were used, equally divided into four categories: instruments, birds, vehicles, and voices (30 different sound exemplars in each category). These original sounds were selected in the Sound Ideas database (vehicle and bird sounds), in the Vienna Library database (instrument sounds), and in the RWC database (voice sounds). As in the Giordano et al.'s study [15], the sound set was characterized in terms of pitch, HNR and spectral centroid. Ten different instruments were selected: celesta, bassoon, flute, harp, clarinet, marimba, oboe, trumpet, cello, and vibraphone. Each instrument was played at 3 different pitches (F4, G#4, and B4), leading to 30 instrument sounds. For the voices,

5 different vowels were chosen (/a/, /e/, /i/, /o/, and /u/), each sung by a male tenor singer or a female alto singer. Vowels were sung at the same pitches as the instrument sounds (F4, G#4, and B4). Their Harmonic-to-Noise Ratio (HNR) was estimated using Praat software [29]. The HNR measures the ratio of the periodic and aperiodic (noisy) components of a signal. The mean (± standard deviation, SD) HNR value for the voices was 22.3 dB ± 9.1 dB; for the instruments, it was 26.6 dB ± 5.9 dB. A one-way ANOVA was run to compare the mean HNR of the four categories, and it revealed a significant effect of the HNR [$F(3,116) = 87,35$; $p<0.0001$; see below for the other two categories]. Tukey-HSD post-hoc tests showed that there was no statistical difference between the HNRs of the voices and the instruments [$p = 0.07$]. The 30 bird sounds were composed of a variety of birds: e.g. blue jay, crow, eagle, flycatcher. The pitch values for each of these 30 sounds were estimated using Praat, by means of an autocorrelation method. The pitch estimates ranged from 308.9 Hz to 582.8 Hz with a mean of 491.0 Hz (± 76.2 Hz). The average pitch for birds was slightly higher than the pitches of voices and instruments [$t = 4.03$; $p<0.0002$]. The mean HNR (± SD) for bird sounds was 11.5 dB (± 6.0 dB), lower than the HNR for voices and instruments, but higher than the HNR for vehicles (see below) [Tukey HSD post-hoc tests: $p<0.0002$ in all cases]. The vehicle sounds were running engine sounds. Because of their noisy nature, pitch could not be estimated for the vehicle sounds; with the exception of a few of them, for which the pitch obtained was apparently lower than the pitch of the other categories (around 180 Hz, compared to an average around 450 Hz for the three other categories). The HNR estimate of the vehicle sounds had a mean of 0.3 dB ± 6.2 dB. It was lower than for all the other categories [Tukey-HSD post-hoc tests: $p<0.0002$ in all cases]. Instruments and voices were comparable in terms of spectral centroid because of their similar harmonic structure (respectively: 955 Hz ± 495 Hz and 943 Hz ± 545 Hz), whereas it was much higher for birds [Tukey-HSD post-hoc tests: $p<0.0002$ in all cases], although with an important variance from one sound to the other (3122 Hz ± 1193 Hz). Spectral centroid of the vehicles was comparable to that of instruments and voices, with a mean of 719 Hz ± 449 Hz. Finally, all 120 sounds were equalized in duration (250 ms). The vehicle sounds were almost stationary and we arbitrarily chose 250-ms excerpts in the sounds (with 10-ms fades in and out to prevent clicks). For the bird, instrument, and voice sounds, the first 250 ms of the signal were kept, thus preserving their natural attack. 10-ms fade-outs were applied to these sounds. The sounds had a sampling frequency of 16 kHz, so there was no energy above 8 kHz.

These 120 original sounds were then transformed in acoustic and auditory sketches, following the method outlined in the Introduction: peaks were selected on an acoustic or auditory time-frequency representation of the sound, and then resynthesized in a new sound: the sketch (see Fig 1). Six sketch conditions were created: two representations (acoustic and auditory), and three levels of sparsity (10, 100, and 1000 features/second), leading to a total of 720 sketch sounds. A set of sound examples is available at: https://hal.archives-ouvertes.fr/hal-01250175.

The acoustic spectrogram was performed with fast Fourier transform on 8-ms Hanning windows. The auditory spectrogram mimics the frequency decomposition performed by the cochlea. It was obtained with 128 overlapping constant band-pass filters with center frequencies uniformly distributed along a logarithmic frequency axis, followed by spectral sharpening simulating lateral inhibition (1st order derivative and half-wave rectifier) [28]. The original programs are freely available online as the "NSL toolbox" (http://www.isr.umd.edu/Labs/NSL/Software.htm). Temporal integration was performed with 8-ms time windows. The resulting matrices for both representations had a similar size of 128 frequency bins x 32 temporal samples. The selection of features performed on these representations was based on the peak-picking of local maxima (see [23]). Here, the features were energy peaks. First, all local maxima were identified. Then, they were sorted by decreasing order, and only the n largest were kept; n

corresponding to the sparsification level. This algorithm, by using a simple local maxima detection, tended to select relatively distant energy peaks, as high-energy areas in the original time-frequency representations could be summarized in one peak (for an illustration of the peak-picking effect, see also Fig 2). The same three levels of sparsification as in the previous study [23] were chosen: 10, 100 and 1000 features per second. For the 250-ms stimuli of the current experiment, it means that 3, 25 or 250 energy peaks were kept. However, for some of the sounds, the total number of peaks was smaller than 250 (more precisely, for 39% of the sounds, for which the mean number of peaks was 167±48). This was the case for all the auditory sketches of the instrument sounds (M = 155±39), 4 acoustic sketches of the instrument sounds (M = 213±30), 24 auditory sketches of the bird sounds (M = 205±29), 4 auditory sketches of the vehicle sounds (M = 232±9), all auditory sketches of the voice sounds (M = 130±36), and 2 acoustic sketches of the voice sounds (M = 227±3).

Then, the sparse representation was inverted back to give rise to the sketches. For the acoustic sketches, reconstruction was possible by simple inverse fast Fourier transform. The selected acoustic features were converted back to the amplitudes and phases stemming from the analyses of the original signals, to be resynthesized in acoustic sketches. For the auditory sketches, because of the nonlinear processing (lateral inhibition, thresholding), direct reconstruction could not be achieved. Similarly as in [23], we used the method developed by Yang et al. [30], which provides reconstruction for auditory spectrogram that are perceptually similar to the original, whenever there is no specific treatment on the auditory spectrogram. The algorithm estimates the phases thanks to an iterative procedure, which starts with a Gaussian distributed white noise and reconstructs the time waveform by inverse filtering.

Finally, the 840 (720 sketches and 120 original) stimuli were normalized by the root-mean-square level. Examples of original sounds and sketches, from each category, are represented in Fig 2.

**Apparatus.** The stimuli were presented diotically through a Sennheiser HD 250 Linear II headphone, connected to a RME Fireface 800 digital-to-analog converter, at a 16 kHz sampling rate. They were presented at around 68 dBA. The experimental session was run using a Matlab R2008b interface on an Apple Mac Pro. The participants were tested individually in a double-walled IAC sound-proof booth. They provided their response using the computer mouse, by clicking on the corresponding button on a computer screen.

**Procedure.** A four-alternative forced choice (4-AFC) paradigm was used. On each trial, participants heard a single sound, which could be either an instrument, a bird, a vehicle or a voice sound. They had to classify the sound they just heard into one of the four categories. No feedback was provided during the test sessions, only during the short training blocks.

For the main experiment, only sketch sounds were used. We carefully avoided familiarizing the participants with the original sounds at the beginning, to ensure that the first encounter with each sound was with its sketch version. For each of the six sketch conditions (2 time-frequency representations x 3 sparsification levels), and for each of the categories, 30 repetitions (each corresponding to a different sound; see Stimuli section) were collected. These 720 trials were presented in a randomized manner. Breaks were possible every 180 trials. Before data collection began, participants performed a short training block. The training block contained one example of a sketch sound for each category and for each condition, leading to 24 stimuli in total. Sounds for the training were not included in the main experiment stimulus dataset. At the end of the main experiment, a control block was run on the original stimuli alone, to ensure that the original sounds were well recognized. The order of presentation of the 120 original stimuli was randomized within a unique block of 120 stimuli. The four original sounds that were used to generate the sketches sounds of the first training session were also presented at the

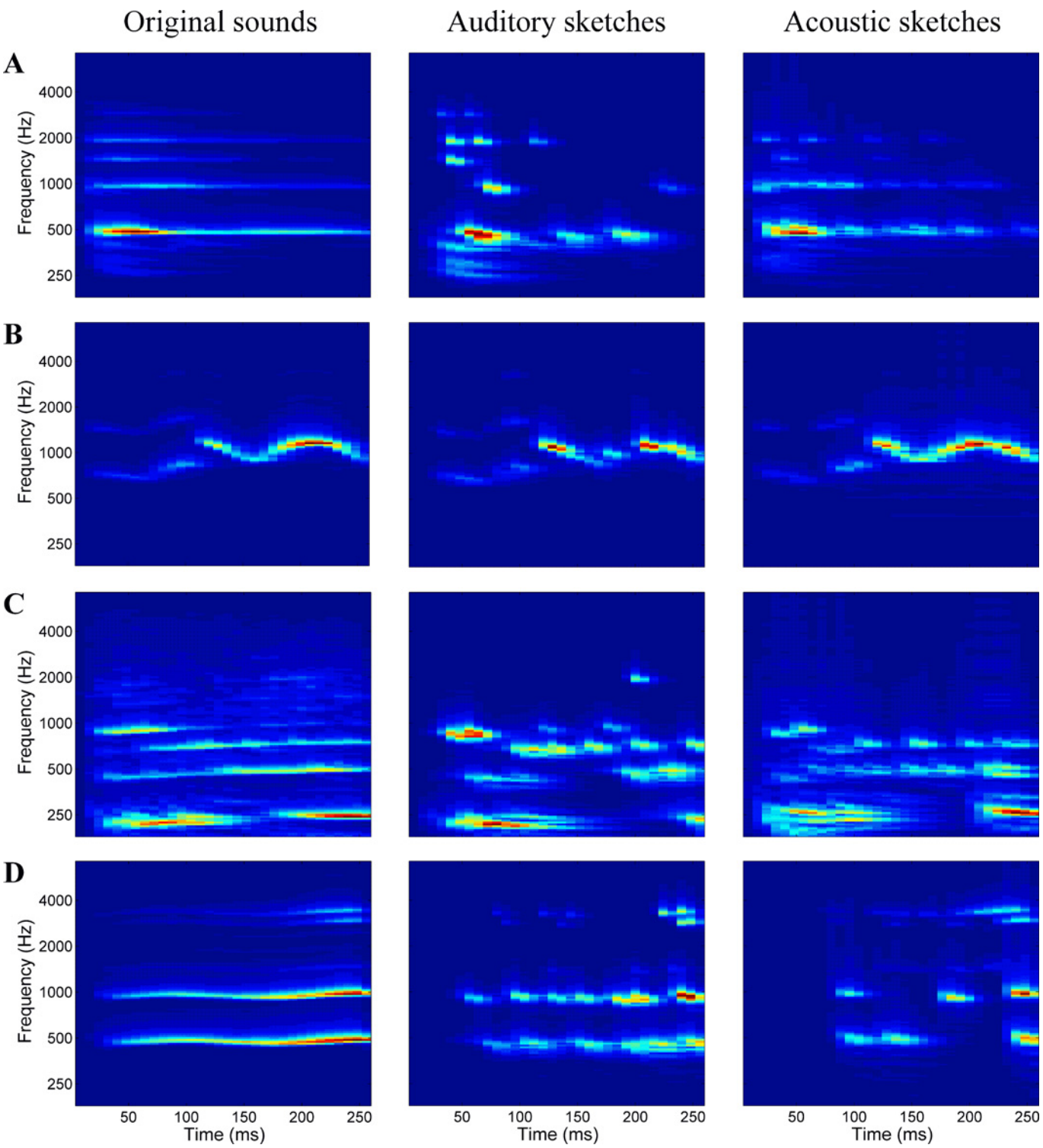


**Fig 2. Auditory spectrograms of original and sketch stimuli.** All panels are auditory time-frequency representations (Chi et al., 2005; see Original and sketch sounds section) of original and sketch stimuli. Left: original sounds; middle: auditory sketches (100 feat./s); right: acoustic sketches (100 feat./s). The sound examples are from the categories: (A) instruments: a harp playing a B4, (B) birds: a loon vocalization, (C) vehicles: a motorcycle, (D) voices: a female voice, singing the vowel /a/, B4.

doi:10.1371/journal.pone.0150313.g002

beginning of the second block, as a small training. The total experiment lasted about two hours.

**Statistical analyses: signal detection model.** To evaluate performance, the d' statistic of signal detection theory (SDT) was used. However, because the original theory has been developed for tasks with only 2 possible responses (yes/no, or 2-AFC; see [31]), we had to extend the theory to a 4-AFC case.

The traditional approach to apply SDT to m-AFC tasks (here, m = 4) assumes that there is no response bias [31]. However, it has been shown that this can affect the d' computations [32]. The method described by DeCarlo [32] takes into account the biases for each possible choice, in order to compute a global d' score for m-AFC tasks. It means that, although biases are computed for each choice (here, the four categories), only an average d' would be available. We thus extended DeCarlo's method to compute d' scores for each category, and in each sparsification condition, while still taking the biases into account.

The decision rule here was the same as in DeCarlo [32], whereas the structural model, differing from DeCarlo [32], included the d' scores for each category as variables. We derived from the decision rule and the structural model a set of three equations which constitutes our 4-AFC model with bias in a normal theory version,

$$p(Y = 1|X_1, X_2, X_3) = \int_{-\infty}^{\infty} \Phi(d_1X_1 - d_2X_2 + b_1 - b_2 + e_1)\Phi(d_1X_1 - d_3X_3 + b_1 - b_3 + e_1)\Phi(d_1X_1 - d_4X_4 + b_1 + e_1)f(e_1)d(e_1),$$

$$p(Y = 2|X_1, X_2, X_3) = \int_{-\infty}^{\infty} \Phi(d_2X_2 - d_1X_1 + b_2 - b_1 + e_2)\Phi(d_2X_2 - d_3X_3 + b_2 - b_3 + e_2)\Phi(d_2X_2 - d_4X_4 + b_2 + e_2)f(e_2)d(e_2),$$

$$p(Y = 3|X_1, X_2, X_3) = \int_{-\infty}^{\infty} \Phi(d_3X_3 - d_1X_1 + b_3 - b_1 + e_3)\Phi(d_3X_3 - d_2X_2 + b_3 - b_2 + e_3)\Phi(d_3X_3 - d_4X_4 + b_3 + e_3)f(e_3)d(e_3).$$

To limit the number of variables in our set of equations, the bias scores computed with DeCarlo's method [32] were used as inputs in our version of the model. The four d' scores, corresponding to the four categories, were the variables. Both models were fitted thanks to OpenBUGS programs. They were run for each dataset with 4000 burnins and 16000 iterations. With this amount of iterations, the Monte Carlo errors were less than 5% of the posterior standard deviation, which is the criterion suggested for convergence (cf. [32]).

All the statistical tests (repeated-measures ANOVA) were conducted on these d' scores. Chance level corresponds to a d' of 0, while near perfect recognition (here, proportion correct of 97%) corresponds to a d' of 3.2.

## Results

**Outlier sounds and participants.** The sounds, in their original version, were overall well recognized by participants (97.2% ± 8.4%) except for three sounds (two bird sounds and one vehicle sound), which were misidentified by more than 30% of participants. The results for the original and simplified versions of these sounds were excluded from the analyses.

For each participant, a recognition score for the original stimuli was computed. The mean recognition score of the 14 participants with the original stimuli, excluding the three outlier sounds, was 98.3% (SD = 4.7%). We also computed recognition scores for each participant on the 702 remaining simplified stimuli (M = 51.4% ± 5.7%). One participant had a particularly

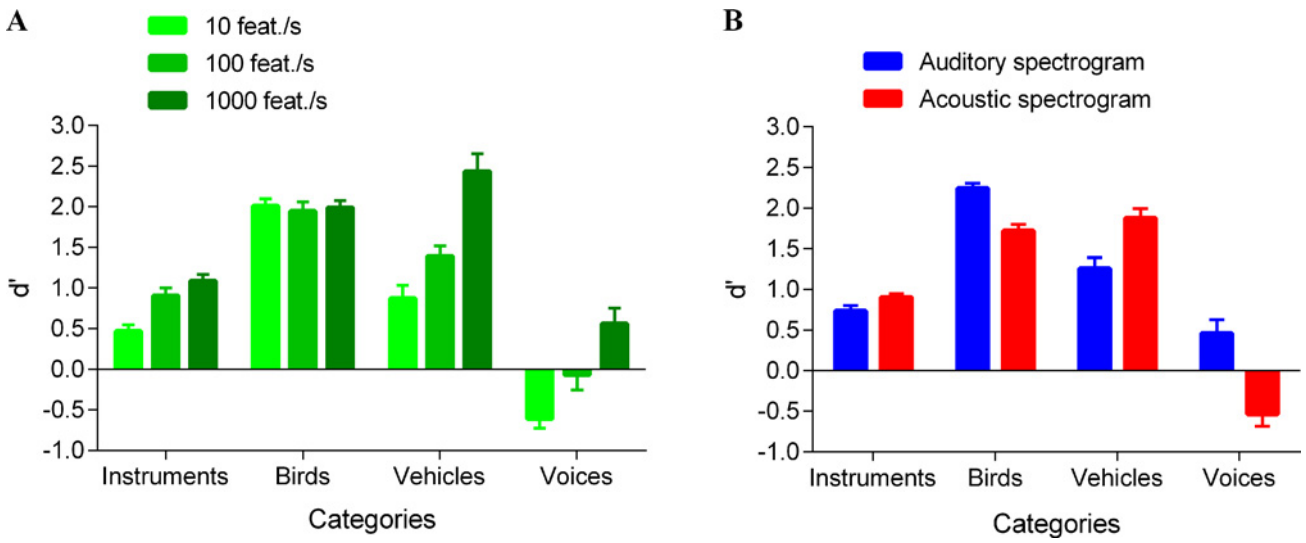


**Fig 3. Recognition performance.** (A) For each category, performance (as measured by d') is displayed at each sparsification level. With the exception of voice sounds, performance was well above chance even at the highest sparsification level, 10 feat./s. For voice sounds, performance was at chance or even lower (negative d'), meaning that participants responded systematically anything but voices for these voice sounds. (B) Performance is displayed for auditory sketches and acoustic sketches. For bird and voice stimuli, performances were higher with auditory sketches, whereas for vehicles, the reverse pattern emerged. No differences were observed for instrument sounds. Error bars correspond to the standard error of the mean.

doi:10.1371/journal.pone.0150313.g003

low recognition score with the simplified stimuli (< mean– 2xSD) and was thus excluded from the following analyses. The mean recognition score on the 13 remaining participants for the simplified stimuli was 52.4% (SD = 4.3%).

**Global recognition of the simplified stimuli.** A repeated-measures ANOVA with the sparsification level, the category, and the time-frequency representation as within-subjects variables was performed on the d' scores. Significant effects were further analyzed with Tukey-HSD post-hoc tests. Fig 3 displays the recognition performances for each category and each sparsification level.

It first revealed a significant main effect of the sparsification level [$F(2,24) = 31.447$; $p<0.00001$; $\eta_p^2 = 0.724$]. Participants better recognized sounds with low level (1000 feat./s) than medium level (100 feat./s) of sparsification [$p<0.0006$]. They also better recognized sounds with medium level (100 feat./s) than high level (10 feat./s) of sparsification [$p<0.007$].

The repeated-measures ANOVA also revealed a significant main effect of the time-frequency representation variable [$F(1,12) = 7.188$; $p<0.03$; $\eta_p^2 = 0.375$]. Recognition performances were better with auditory sketches than with acoustic sketches. However, this difference [$p<0.03$] was very small ($\Delta d' = 0.2$).

A significant main effect of the category was also found [$F(3,36) = 108.810$; $p<0.00001$; $\eta_p^2 = 0.901$]. Recognition performances were higher for the bird sounds than for the vehicle sounds [$p<0.008$]; higher for the vehicle sounds than for the instrument sounds [$p<0.0002$]; and higher for the instrument sounds than for the voice sounds [$p<0.0002$].

The ANOVA exhibited a significant two-way interaction between category and time-frequency representation variables [$F(3,36) = 23.552$; $p<0.00001$; $\eta_p^2 = 0.663$]. Participants better recognized the auditory sketch versions of bird and voice sounds than the acoustic sketch versions [respectively: $p<0.03$ and $p<0.0002$]. In contrast, they better recognized acoustic sketches of vehicle sounds [$p<0.004$]. For the instrument sounds, recognition performances were similar in both sketch conditions [$p = 0.941$].

Finally, the ANOVA revealed a significant two-way interaction between category and sparsification level [$F(6,72) = 8.023$; $p<0.00001$; $\eta_p^2 = 0.401$] (see Fig 3). For the vehicles, the recognition performances were better with a low level (1000 feat./s) than with medium level (100 feat./s) and high level (10 feat./s) of sparsification [$p<0.0002$ in both cases]. Similarly, for the voices, recognition performances were better with a low level (1000 feat./s) than with a medium level (100 feat./s) or high level (10 feat./s) of sparsification [respectively: $p<0.04$ and $p<0.0002$]. For the instruments, the recognition performances were better with low level (1000 feat./s) than with high level (10 feat./s) of sparsification [$p<0.04$]. For the birds, recognition performances were not influenced by the sparsification levels [$p = 1.000$ in all cases].

The effect of the two-way interaction between time-frequency representation and sparsification level variables was not significant [$F(2,24) = 0.971$; $p = 0.393$; $\eta_p^2 = 0.075$], nor was the three-way interaction [$F(6,72) = 1.321$; $p = 0.259$; $\eta_p^2 = 0.099$].

Finally, to investigate the sparsity levels for which recognition was above chance, we performed one-sample t-tests testing d' against 0 (chance level) for all conditions. Instrument and bird sketches were all significantly recognized above chance [one-sample t-tests: $p<0.002$ in all cases]. For vehicle sketches, they were also all recognized significantly above chance [$p<0.00001$ for all cases except the auditory sketches at 10 feat./s; $p = 0.05$]. Finally, for the voices, recognition performance was significantly above chance only for auditory sketches with a low sparsity level (1000 feat./s) [$p<0.0003$]. Unexpectedly, recognition was significantly below chance for acoustic sketches at 10 feat./s [$p<0.00001$]. This means that participants classified systematically the voice stimuli in any other category but not the voice one.

## Acoustic Analyses: Auditory Distance Model

To understand the possible acoustical bases of the perceptual results described above, we derived a new model of auditory similarity, based on the model developed by Agus et al. [33]. The original model computes auditory distances between two different sound categories. The model is based on the time-frequency distribution of energy for each sound, estimated using spectro-temporal excitation patterns (STEPs; [34]), which simulate peripheral auditory filtering. Auditory distances are then computed between pairs of STEPs, using a dynamic time-warping algorithm, to minimize the possible misalignment between features. Several behavioral results emerged from our data: we thus computed several auditory distances to investigate their auditory bases.

Firstly, in order to evaluate the impact of the sparsification level on the auditory distances, auditory distances between the original version of the sound and each of the sparsification level were computed. This was done for both acoustic and auditory sketches. The auditory distance between a sketch and its original version increased with the degree of sparsity: the higher the degree of sparsity, the larger the distance was [$F(2,478) = 256.32$; $p<0.00001$; $\eta_p^2 = 0.51$]. The mean distances were $0.21 \pm 0.07$ at 1000 feat./s; $0.29 \pm 0.12$ at 100 feat./s, and $0.37 \pm 0.16$ at 10 feat./s. This result mirrors the behavioral effect: the closer (in auditory distance terms) a sound was to its original version, the easier it was to recognize.

Secondly, we evaluated the effect of the time-frequency representation used as a basis for sparsification. We thus computed the distances between acoustic and auditory sketches, for each category. The auditory distance between an acoustic sketch and an auditory sketch depended on the category [$F(3,267) = 30.061$; $p<0.00001$; $\eta_p^2 = 0.25$], with slightly (as revealed by the small $\eta_p^2$) lower auditory distances for the birds than for the three other categories [Tukey-HSD post-hoc test: $p<0.00001$]. The mean distances were $0.20 \pm 0.06$ for instruments, $0.17 \pm 0.07$ for birds, $0.26 \pm 0.05$ for vehicles, $0.21 \pm 0.07$ for voices. Overall, no large differences between acoustical and auditory sketches emerged, as a function of the category. There was an

acoustical difference between both types of sketches, but this auditory model distance could not explain the pattern of results observed in the perceptual results (see Fig 3B).

Finally, in order to model the entire set of data, we computed distances between categories, for each sparsification level and each type of sketch. This model is probably the most accurate to model our data, because it allows us to compare the auditory distances between categories (as a function of sparsification level and time-frequency representation) with the d' scores, which are themselves 'perceptual' distances between the categories, for the 4-AFC task. Each category was considered successively as a target category (30 sounds), while the three remaining categories were considered as distractor categories (90 sounds). For each condition (2 time-frequency representations x 3 sparsification level), we computed the auditory distance Ad between each stimulus and stimuli from the other categories with the following equation: $Ad(i) = \mu_{distr}(i) - \mu_{targ}(i), i = 1 \ldots 30$, where $i$ is a stimulus of the target category, $\mu_{distr}$ is the mean of the distances between the target stimulus and each stimulus of the distractor categories (90 distances for each target stimulus), and $\mu_{targ}$ is the mean of the distances between the target stimulus and the other stimuli of the target category (29 distances for each stimulus). The six distance matrices are represented in Fig 4.

To compare the auditory distance matrices with the perceptual results (d'), we computed, for each category, the mean of the auditory distances Ad on all the stimuli of the category. Fig 5 displays the d' scores for each condition as a function of the corresponding auditory distances for these conditions. The perceptual results were overall well correlated with the auditory

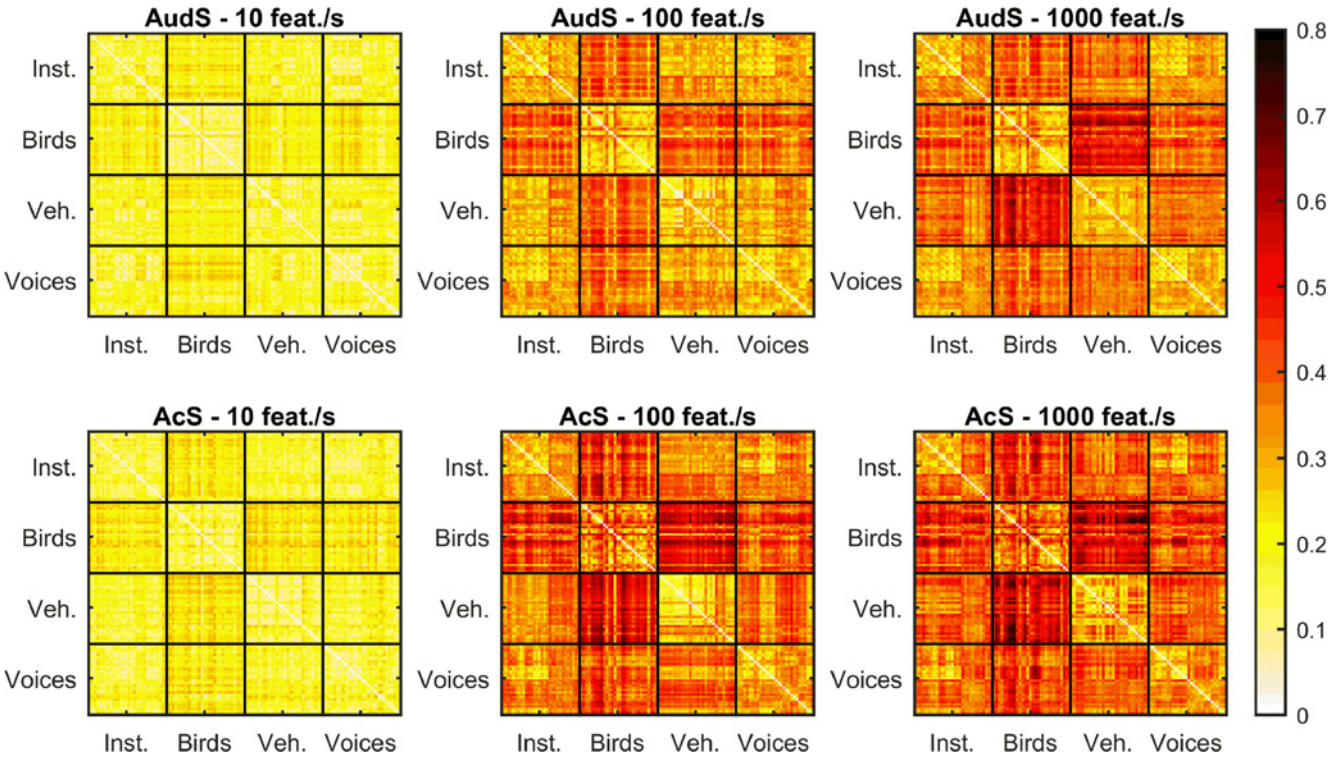


**Fig 4. Auditory distance model.** For each time-frequency representation (AudS: auditory spectrogram, and AcS: acoustic spectrogram) and each sparsification level (10, 100, and 1000 features per second), an auditory distance dissimilarity matrix is plotted (see [33]). The mean absolute distance between STEPs [34] is represented for each sound pair of each category (Inst. for musical instruments, Birds, Veh. for vehicle engine sounds, and Voices). With the high level of sparsity (10 feat./s), sounds are more similar between them than with the low level of sparsity (1000 feat./s). No obvious differences emerged between the two auditory or acoustic time-frequency representations.

doi:10.1371/journal.pone.0150313.g004

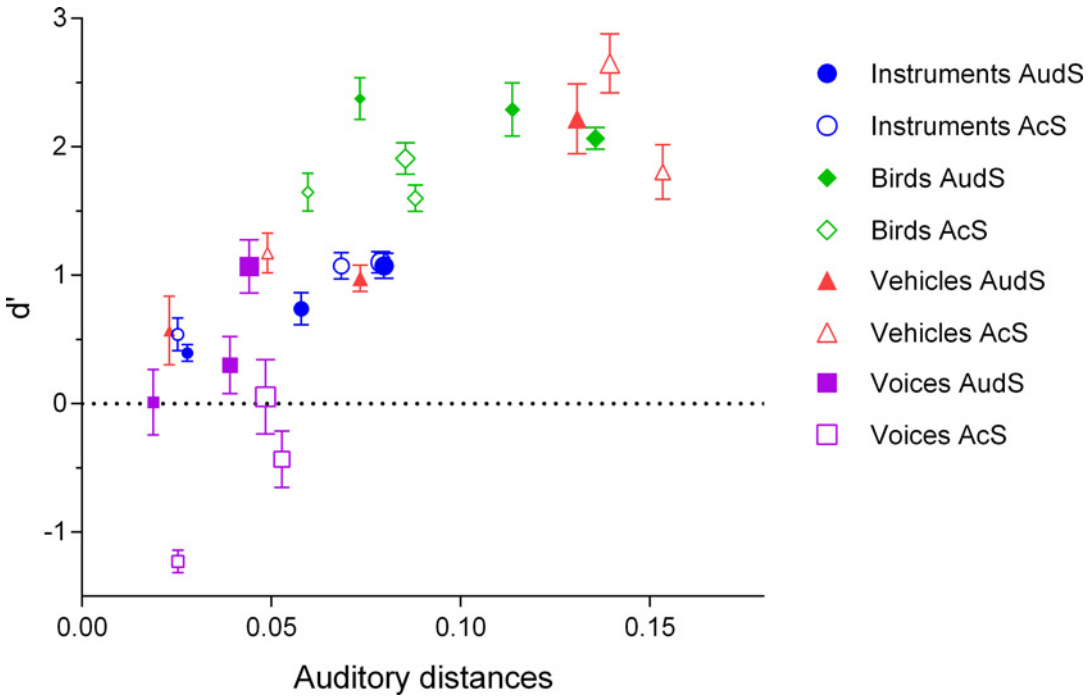


**Fig 5. The perceptual results (d') plotted as a function of the auditory distance values.** The filled symbols represent sketches based on the auditory spectrogram (AudS) representation; the open symbols are for the acoustic spectrogram (AcS) representation. The size of the symbols corresponds to the level of sparsity: small symbols for 10 feat/s; medium symbols for 100 feat/s; large symbols for 1000 feat/s. Error bars correspond to the standard error of the mean. A good correlation is exhibited between the model and the data.

doi:10.1371/journal.pone.0150313.g005

distances: the more sounds were similar (smaller auditory distances), the more they were difficult to recognize (Spearman's correlation: $\rho = 0.721$, $p<0.00001$).

## Discussion

We have studied acoustic and auditory sketches, new classes of sounds based on sparse representations, which are severely impoverished versions of original sounds. Salient features of the sounds were kept by means of a simple peak-picking algorithm that was performed on a time-frequency representation. Two representations were compared: an auditory spectrogram, i.e. a biologically grounded representation, and an acoustic spectrogram. Three levels of 'sparsification', i.e. the number of energy peaks kept in the representation, were also contrasted. To test the sketch process, we conducted an experiment that assessed the recognition by human listeners of acoustic and auditory sketches. We found that: (1) with the notable exception of voice sounds (see (3)), all sounds were recognized above chance even when they were drastically impoverished, i.e. with a very low number of features (10 feat./s). In addition, recognition performance increased with the increase of the number of features; (2) in contrast to our hypothesis, no clear differences in recognition were observed between the acoustic and the auditory sketches; (3) voice sounds followed a very different response pattern than all the other categories, with performance being at chance, or even below chance, meaning that participants systematically recognized voice sounds as any other categories except voice itself; (4) a model based on auditory distances between spectro-temporal excitation patterns (STEPs) exhibited a good correlation with the perceptual data.

In our experiment, for all sounds except voice sounds, extreme sparsification could be applied (only 3 peaks for one sound were kept in the most drastic conditions) while keeping

recognition above chance. As has been shown previously for speech [21,22] and environmental [2] sounds, auditory recognition can be very robust to sound distortions and modifications. It is also worth noting that for the bird sound category, recognition plateaued already at 10 feat./s (with a relatively high d', around 2). The few key features, probably located in the upper part of the spectrum (higher spectral centroid), were selected by the peak-picking algorithm even at the highest level of sparsification (10 feat./s). Within this set of sounds where the birds stand out with respect to these high-frequency features, adding more peaks did not add any necessary information for the listener. For all the other categories, as expected, as the number of features increased, so did recognition performances. Altogether, these results confirm that (very) sparse representations of sounds can produce perceptually relevant results.

However, the results of our experiment did not support the hypothesis that, for any type of natural sound, sparse representations would lead to better results if they are implemented on a biologically grounded representation of sounds, like an auditory spectrogram. For some categories (voices and birds), the behavioral advantage, evidenced by better performance, was indeed observed for the auditory sketches. However, no differences were found for the instrument sounds; and the reverse pattern—with higher performance for the acoustic sketches—appeared for the vehicle sounds. No simple explanation for this surprising interaction can be given based on the basic timbre features computed on the sounds (see Methods for the differences between the categories in the spectral centroid and the HNR). One of the limitations of our experiment is the short duration of the sound used (250 ms). With this duration, a possible difference between the two representations would not arise. The probability to have more potential features useful for recognition (and thus more important differences between the features kept in the two representations) is indeed higher with longer sounds, at least for sounds that are not stationary. It is interesting to note that the only category, for which performance was actually worse with the auditory sketches than with the acoustic sketches, is the only one that contained stationary sounds (vehicles). This is in accordance with a result obtained in the audio signal processing community (see [35]). They found that, whereas for typical steady-state signals the Fourier representation is sufficient to provide a good representation, for sounds with onsets and transients, like voices, animal calls, or music, a 'union of bases' composed of both a Modified Discrete Cosine Transform basis and a Wavelet basis, is needed to have a better sparse representation of the signal. Another way to interpret this result is in terms of a dichotomy between living and non-living sounds. Living sound (voices and birds) were better recognized when presented as auditory sketches, whereas non-living sounds did not show any advantage (or even show a disadvantage) when presented as auditory sketches. This interpretation remains to be confirmed and extended in future experiments.

The results obtained for the voice sounds, i.e. recognition at chance level or even negative d', can be seen as another behavioral evidence that voice is special (see [36] for the evidence for speech; [7,33] for behavioral evidences for voices; [37] for a review on the selectivity for the human voice observed in fMRI studies). The cues useful for voice recognition were completely removed when subjected to the sketch process. In Agus' study [33], using chimeric sounds in which the temporal structure from one sound (e.g. instrument) is mixed with the spectral structure of another (e.g. voice), they showed that only natural voices could elicit special behavioral advantage for voices; in their case, this advantage was provided by faster reaction times. In the present study, we showed that a large and diverse set of sounds could be simplified with only a few peaks, while still being recognized well above chance. The noticeable exception of voice sounds may suggest that for recognition of voices to be effective, complex spectro-temporal patterns might be needed.

Finally, whatever the features used to recognize sounds sparsified on different representations and with different levels of sparsification, the perceptual results were relatively well

correlated with a model based on auditory distances of STEPs (see [33]). The larger the distances between one category and the other three, the better the recognition was. This new auditory distance model would probably be useful in future studies, using various techniques such as psychophysical methods and/or brain imagery, in a further attempt to equalize different classes of stimuli along some relevant acoustic or auditory dimensions (see for example [13,15]).

## Supporting Information

**S1 Dataset. D-primes for all participants.**
(XLSX)

**S2 Dataset. Auditory distances between each sound for each sparsification condition.**
(XLSX)

## Acknowledgments

We are very grateful to Lawrence T. DeCarlo and Trevor Agus for sharing their codes and for their help. We would like to thank Véronique Chastres for her help with the statistical analyses on a previous version of this document.

## Author Contributions

Conceived and designed the experiments: VI CS IVD. Performed the experiments: VI. Analyzed the data: VI MT CS. Contributed reagents/materials/analysis tools: VI CS. Wrote the paper: VI MT IVD CS.

## References

1. Ballas JA. Common factors in the identification of an assortment of brief everyday sounds. Journal of experimental psychology: human perception and performance. 1993; 19(2):250. PMID: 8473838
2. Gygi B, Kidd GR, Watson CS. Spectral-temporal factors in the identification of environmental sounds. The Journal of the Acoustical Society of America. 2004; 115(3):1252. PMID: 15058346
3. Felsen G, Dan Y. A natural approach to studying vision. Nature neuroscience. 2005; 8(12):1643–6. PMID: 16306891
4. Suied C, Viaud-Delmon I. Auditory-visual object recognition time suggests specific processing for animal sounds. PloS one. 2009; 4(4):e5256. doi: 10.1371/journal.pone.0005256 PMID: 19384414
5. Robinson K, Patterson RD. The stimulus duration required to identify vowels, their octave, and their pitch chroma. The Journal of the Acoustical Society of America. 1995; 98(4):1858–65.
6. Robinson K, Patterson RD. The duration required to identify the instrument, the octave, or the pitch chroma of a musical note. Music Perception. 1995:1–15.
7. Suied C, Agus TR, Thorpe SJ, Mesgarani N, Pressnitzer D. Auditory gist: recognition of very short sounds from timbre cues. J Acoust Soc Am. 2014; 135(3):1380–91. doi: 10.1121/1.4863659 PMID: 24606276
8. Romanski LM, Tian B, Fritz J, Mishkin M, Goldman-Rakic PS, Rauschecker JP. Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nature neuroscience. 1999; 2 (12):1131–6. PMID: 10570492
9. De Lucia M, Clarke S, Murray MM. A temporal hierarchy for conspecific vocalization discrimination in humans. The Journal of Neuroscience. 2010; 30(33):11210–21. doi: 10.1523/JNEUROSCI.2239-10.2010 PMID: 20720129
10. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. Nature. 2000; 403(6767):309–12. PMID: 10659849
11. Lewis JW, Brefczynski JA, Phinney RE, Janik JJ, DeYoe EA. Distinct cortical pathways for processing tool versus animal sounds. The Journal of neuroscience: the official journal of the Society for Neuroscience. 2005; 25(21):5148–58.

12. Leaver AM, Rauschecker JP. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. The Journal of neuroscience: the official journal of the Society for Neuroscience. 2010; 30(22):7604–12.
13. Staeren N, Renvall H, De Martino F, Goebel R, Formisano E. Sound categories are represented as distributed patterns in the human auditory cortex. Current biology: CB. 2009; 19(6):498–502. doi: 10.1016/j.cub.2009.01.066 PMID: 19268594
14. Moerel M, De Martino F, Formisano E. Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. The Journal of Neuroscience. 2012; 32 (41):14205–16. doi: 10.1523/JNEUROSCI.1388-12.2012 PMID: 23055490
15. Giordano BL, McAdams S, Zatorre RJ, Kriegeskorte N, Belin P. Abstract encoding of auditory objects in cortical activity patterns. Cereb Cortex. 2013; 23(9):2025–37. doi: 10.1093/cercor/bhs162 PMID: 22802575
16. Altmann CF, Doehrmann O, Kaiser J. Selectivity for animal vocalizations in the human auditory cortex. Cerebral Cortex. 2007; 17(11):2601–8. PMID: 17255111
17. Santoro R, Moerel M, De Martino F, Goebel R, Ugurbil K, Yacoub E, et al. Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS computational biology. 2014; 10(1):e1003412. doi: 10.1371/journal.pcbi.1003412 PMID: 24391486
18. Patil K, Pressnitzer D, Shamma S, Elhilali M. Music in our ears: the biological bases of musical timbre perception. PLoS computational biology. 2012; 8(11):e1002759. doi: 10.1371/journal.pcbi.1002759 PMID: 23133363
19. Smith EC, Lewicki MS. Efficient auditory coding. Nature. 2006; 439(7079):978–82. PMID: 16495999
20. Hromadka T, Zador AM. Representations in auditory cortex. Current opinion in neurobiology. 2009; 19 (4):430–3. doi: 10.1016/j.conb.2009.07.009 PMID: 19674890
21. Shannon RV, Zeng F-G, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. Science. 1995; 270(5234):303–4. PMID: 7569981
22. Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech perception without traditional speech cues. Science. 1981; 212(4497):947–9. PMID: 7233191
23. Suied C, Drémeau A, Pressnitzer D, Daudet L. Auditory sketches: Sparse representations of sounds based on perceptual models. From Sounds to Music and Emotions: Springer; 2013. p. 154–70.
24. Grey JM. Multidimensional perceptual scaling of musical timbres. The Journal of the Acoustical Society of America. 1977; 61(5):1270–7. PMID: 560400
25. McAdams S, Winsberg S, Donnadieu S, De Soete G, Krimphoff J. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. Psychological research. 1995; 58(3):177–92. PMID: 8570786
26. Elliott TM, Hamilton LS, Theunissen FE. Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. J Acoust Soc Am. 2013; 133(1):389–404. doi: 10.1121/1.4770244 PMID: 23297911
27. Krimphoff J, McAdams S, Winsberg S. Caractérisation du timbre des sons complexes.II. Analyses acoustiques et quantification psychophysique. Le Journal de Physique IV. 1994; 04(C5):C5-625–C5-8.
28. Chi T, Ru P, Shamma SA. Multiresolution spectrotemporal analysis of complex sounds. The Journal of the Acoustical Society of America. 2005; 118(2):887. PMID: 16158645
29. Boersma P, Weenink D. Praat: doing phonetics by computer [Computer program]. Version 5.4.14, retrieved 24 July 2015 from http://www.praat.org/. 2015.
30. Yang X, Wang K, Shamma SA. Auditory representations of acoustic signals. Information Theory, IEEE Transactions on. 1992; 38(2):824–39.
31. Macmillan N, Creelman C. Detection Theory: A User's Guide Lawrence Erlbaum Associates. New York. 2005.
32. DeCarlo LT. On a signal detection approach to -alternative forced choice with bias, with maximum likelihood and Bayesian approaches to estimation. Journal of Mathematical Psychology. 2012; 56(3):196–207.
33. Agus TR, Suied C, Thorpe SJ, Pressnitzer D. Fast recognition of musical sounds based on timbre. J Acoust Soc Am. 2012; 131(5):4124–33. doi: 10.1121/1.3701865 PMID: 22559384
34. Moore BCJ. Temporal integration and context effects in hearing. Journal of Phonetics. 2003; 31(3–4):563–74.
35. Plumbley MD, Blumensath T, Daudet L, Gribonval R, Davies ME. Sparse representations in audio and music: from coding to source separation. Proceedings of the IEEE. 2010; 98(6):995–1005.
36. Liberman AM, Mattingly IG. A specialization for speech perception. Science. 1989; 243(4890):489–94. PMID: 2643163

37. Belin P. Voice processing in human and non-human primates. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2006; 361(1476):2091–107. PMID: 17118926

## 1.3 Compléments d'analyses (1/2) : théorie de détection du signal

Les paragraphes qui suivent sur la Théorie de Détection du Signal (TDS) expliquent la démarche que nous avons adoptée pour quantifier les performances des participants. Pour comprendre l'enjeu de cette démarche, il est nécessaire de reprendre quelques notions de base de la TDS.

### 1.3.1 Introduction

Pour quantifier l'intensité d'une excitation en fonction de la réponse du participant, les mesures psychophysiques de seuils sont les plus directes en posant, par exemple, la question au participant : "Percevez-vous un son ?". Pourtant, une simple réponse à cette question ne suffit pas pour en déduire l'intensité de l'excitation produite par le stimulus : à la fois la méthode du test perceptif et le traitement des réponses doivent tenir compte d'une variabilité propre au participant, à cause de facteurs tels que ses capacités perceptives, son attention, ses motivations, etc.

La TDS a pour objectif de prendre en compte la variabilité du participant dans ses prises de décision, lorsqu'il répond correctement mais aussi lorsqu'il répond incorrectement (Macmillan & Creelman, 2005). Le cadre de la TDS a permis de définir et mesurer des phénomènes perceptifs, malgré leur variabilité, et en s'affranchissant d'évènements non-observables de la cognition. Elle a donc apporté plusieurs contributions majeures, en premier lieu une théorisation de la façon dont les décisions sont prises dans des contextes variables et incertains, induisant des biais dans les réponses perceptives (Yost, 2015). Cette théorisation a permis d'aboutir au calcul de la sensibilité indépendamment du biais de réponse.

Le matériau de départ pour évaluer les performances d'un participant sont ses réponses. D'une manière générale, celles-ci seront classées en fonction de leur valeur de vérité. Il se présente deux cas de figure où le participant peut donner une réponse correcte : il peut répondre qu'il perçoit quelque chose alors que le stimulus était présent (détection correcte), et inversement, qu'il ne perçoit rien alors que le stimulus n'était pas présent (rejet correct). Tandis qu'il donnera une réponse incorrecte s'il répond qu'il perçoit quelque chose alors que le stimulus n'était pas

présent (fausse alarme), et inversement, s'il répond qu'il ne perçoit rien alors que le stimulus était présent (omission).

La TDS fait la distinction entre la sensibilité et le biais de réponse à partir de l'ensemble de ces réponses brutes correctes et incorrectes (Stanislaw & Todorov, 1999). La sensibilité correspond à la capacité du participant à discerner le stimulus du bruit. Une sensibilité élevée correspond à des taux élevés à la fois de détections correctes et de rejets corrects. Le biais correspond au seuil d'intensité du stimulus à partir duquel le participant répondra qu'il l'a perçu. On dira d'un participant avec un biais (ou critère) élevé qu'il est plutôt conservateur, car l'excitation nécessaire pour qu'il réponde qu'il a perçu le stimulus est élevée, et il répondra moins souvent qu'il l'a perçu, limitant ainsi les fausses alarmes. Tandis qu'un participant avec un biais faible est dit plutôt libéral, car l'excitation nécessaire pour qu'il réponde qu'il a perçu le stimulus est faible, générant davantage de fausses alarmes. Les discussions qui suivent concernent la méthode employée pour obtenir ces réponses, le nombre de choix dont dispose le participant, et les approximations faites sur les mesures de sensibilité et de biais.

### 1.3.2 Procédure un-intervalle à choix forcé

La procédure un-intervalle permet de faire une mesure de discrimination en présentant un seul stimulus à chaque essai (Macmillan & Creelman, 2005). Il peut s'agir par exemple de reconnaître à chaque essai à quelle catégorie, entre plusieurs, appartient le stimulus présenté. Cette terminologie n'implique pas le nombre de réponses à donner, ni le nombre de stimuli dans le corpus. La mention de "choix forcé" indique cependant que le participant doit choisir sa réponse parmi plusieurs proposées (m-AFC). Le Tableau 2 répertorie les différentes capacités évaluées avec la procédure un-intervalle (cf. Macmillan & Creelman, 2005).

La dénomination de la capacité évaluée dans une tâche peut légèrement varier en fonction des études, avec des termes plus génériques comme "reconnaissance" pour "identification". On peut citer également le terme de "catégorisation", censé désigner la capacité évaluée dans le cas où le nombre de réponses à donner est strictement inférieur au nombre de classes de stimuli. Les cas des procédures à deux-intervalles (ou plus) relèvent eux de la "comparaison".

| Capacité évaluée | Classes de stimuli | Tâche |
|---|---|---|
| Détection | Cible vs. bruit | Oui/Non |
| Reconnaissance | 2 cibles | 2-AFC |
| Identification | $>$ 2 cibles | m-AFC |
| Classification | $\geq$ 2 cibles | Classer les stimuli |

Tableau 2 – Les différentes capacités évaluées avec la procédure un-intervalle, en fonction du nombre de classes de stimuli et de la tâche à effectuer.

### 1.3.3 Analyse des données perceptives : illustration avec une tâche Oui/Non (ou 2-AFC)

**Tâche.** La tâche Oui/Non classique consiste à présenter aléatoirement et sur un grand nombre d'essais, soit du bruit seul, soit du bruit auquel s'ajoute un signal cible. A chaque essai, la tâche du participant est de répondre "Oui" ou "Non" (ou, de façon équivalente, choisir la classe du stimulus dans le cas d'une tâche 2-AFC ; Macmillan & Creelman, 2005), selon qu'il pense avoir perçu la cible ou non.

**Répartition des données correctes et incorrectes.** S'il s'agit de discriminer une cible dans du bruit à du bruit seul, on parle de détection, tandis que s'il n'y a jamais de stimulus de bruit seul, on parle de reconnaissance. Mais dans les deux cas les méthodes d'analyses sont les mêmes.

Dans un premier temps, on répartit dans un tableau à double entrée les réponses des participants en fonction de la présence réelle, ou de l'absence, du signal cible à chaque essai (Tableau 3). Quatre variables sont présentées dans ce tableau, mais en réalité seules deux suffisent pour représenter les données complètes du participant (une valeur pour les stimuli cibles et une valeur pour les stimuli de bruit), dès lors que les nombres de stimuli cibles et de bruits sont fixés. Ces deux valeurs indépendantes sont suffisantes pour comprendre l'ensemble des données perceptives dans ce type de procédures, en revanche les deux sont nécessaires. En effet, si l'on s'intéresse seulement à l'une des deux valeurs, l'interprétation sera faussée. Par exemple, si un participant répond à tous les essais qu'il perçoit la cible, son taux de détections correctes sera maximal, et pourtant il n'aura pas effectué la tâche correctement. Il faut donc tenir compte des deux valeurs.

| | | Stimulus | |
|---|---|---|---|
| | | Cible + bruit | Bruit seul |
| Réponses | Oui | Détections correctes : $P(s\|S)$ | Fausses alarmes : $P(s\|B)$ |
| | Non | Omissions : $P(b\|S)$ | Rejets corrects : $P(b\|B)$ |
| Total | | $P_{tot} = P(s\|S) + P(b\|S)$ $P_{tot} = 1$ | $P_{tot'} = P(s\|B) + P(b\|B)$ $P_{tot'} = 1$ |

Tableau 3 – Répartition des réponses du participant en fonction des stimuli. La lettre 's' indique 'signal', tandis que la lettre 'b' indique 'bruit', avec une minuscule en référence à la réponse du participant, et une majuscule en référence au stimulus physique.

**Proportion de réponses correctes.** La proportion de réponses correctes donne une première information globale sur les performances du participant. Elle correspond à la proportion de détections et rejets corrects :

$$p(c) = \frac{détections\ correctes + rejets\ corrects}{nombre\ total\ d'essais}. \quad (1)$$

D'après cette mesure, plus le participant détecte correctement la cible tout en limitant les fausses alarmes, plus la proportion de réponses correctes augmente. Cependant, cette mesure seule est insuffisante pour conclure sur les mécanismes perceptifs impliqués dans la prise de décision. Par exemple, si un participant répond qu'il perçoit tout le temps la cible, ou si au contraire il répond qu'il ne la perçoit jamais, les proportions de réponses correctes seront identiques, et pourtant les comportements sont opposés. La TDS vient combler ces lacunes en posant des hypothèses sur l'intensité perceptive produite par les stimuli et en reconstruisant deux valeurs indépendantes plus interprétables que des taux de réponses bruts.

**Sensibilité et biais.** L'hypothèse principale de la TDS permettant de calculer la sensibilité et le biais consiste à assimiler chaque état d'observation des différentes classes de stimuli à une distribution gaussienne de même variance le long d'un axe correspondant à l'intensité de l'excitation [13] (Figure 22). Ces distributions reflètent la variabilité de la décision due aux stimuli présentés et à des facteurs

13. Il existe des mesures non-paramétriques de sensibilité et de biais dans les cas où l'on ne suivrait pas cette hypothèse (cf. Stanislaw & Todorov, 1999).

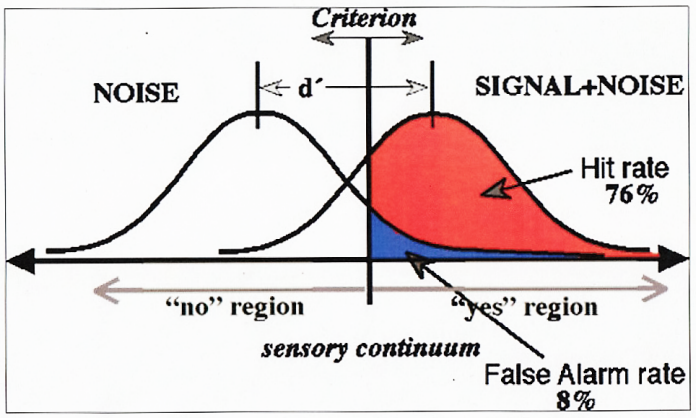


FIGURE 22 – Calcul de la mesure de sensibilité d' : la distance normalisée entre les moyennes des deux distributions, d'après la TDS. Le critère du participant pour répondre "Oui" ou "Non" si le signal est présent ou non génère les taux de détections correctes et de fausses alarmes à partir des distributions, posées comme normales, de la perception du signal dans du bruit et du bruit seul le long du continuum sensoriel. Source : Yost (2015).

internes comme le bruit cérébral ou l'attention. A chaque essai, les stimuli de chaque classe produiront une excitation variable répertoriée par sa distribution correspondante.

La distance entre les deux distributions quantifie la sensibilité du participant. Quant au critère de réponse, c'est-à-dire le biais, il est supposé fixe. Si l'excitation dépasse ce critère, alors le participant répond que la cible était présente dans le stimulus, sinon il répond qu'elle ne l'était pas. L'aire sous la distribution de la cible à partir du critère de réponse (en rouge sur la Figure 22) correspond au taux de détections correctes, tandis que l'aire sous la distribution du bruit à partir du critère de réponse (en bleu sur la Figure 22) correspond au taux de fausses alarmes. Les portions d'aires restantes correspondent aux taux d'omissions et de rejets corrects.

Ces deux indicateurs, la sensibilité et le biais, apportent des informations plus interprétables que les taux de détections correctes et de fausses alarmes en les combinant tous les deux. En effet, la sensibilité $d'$ quantifie la capacité du participant à discriminer la cible du bruit seul. Par exemple, dans le cas de performances parfaites, le participant présenterait un taux de détections correctes de 1 et un taux de fausses alarmes de 0. Tandis que le cas d'un participant ne pouvant dis-

tinguer la cible du bruit et répondant au hasard se traduit par le recouvrement des deux distributions, c'est-à-dire les deux taux mentionnés égaux à 0.5.

Le biais permet quant à lui de quantifier le fait qu'un participant est plutôt libéral (critère décalé vers la gauche) ou plutôt conservateur (critère décalé vers la droite). Si le participant est plutôt libéral, alors il répond souvent qu'il perçoit la cible, et ses taux de détections correctes et de fausses alarmes sont élevés. Dans le cas d'un participant plutôt conservateur, c'est l'inverse. Cette mesure complète celle de sensibilité $d'$ en indiquant si un biais dans le processus décisionnel (et non plus perceptif), interne au participant ou expérimental, l'incitait à donner plus souvent une réponse plutôt que l'autre.

Comme pour les taux de réponses correctes, pour lesquels deux taux indépendants sont nécessaires pour décrire complètement les données, les deux mesures de sensibilité et de biais décrivent complètement les données tout en permettant de les interpréter plus explicitement. Elles permettent également de comparer les données de différents participants, et même pour des tâches avec davantage d'alternatives. Par exemple, pour un 2-AFC, si un participant répond au hasard, son taux de proportion correcte sera de 0.5, alors qu'il sera de 0.33 pour un 3-AFC, etc. Tandis que le niveau de chance en terme de $d'$ sera toujours de 0. Green & Swets (1966) ont montré que cette correspondance pour différents nombres d'alternatives est validée expérimentalement. Un exemple est représenté sur la Figure 23, correspondant à une expérience de perception auditive où le nombre d'alternatives varie de deux à huit. La sensibilité est calculée pour chaque cas à l'aide de la table d'Elliot (1964) supposant un biais nul.

### 1.3.4 Méthodes de calcul de la sensibilité et du biais pour une tâche Oui/Non (ou 2-AFC)

**Calcul de la sensibilité.** La mesure de la sensibilité $d'$ est déterminée par la répartition des distributions de probabilité pour chaque classe de stimuli sur l'axe différenciant ces distributions. Connaissant les taux de détections correctes et de fausses alarmes obtenus expérimentalement, c'est-à-dire des portions d'aires en rouge et en bleu sur la Figure 22, il est possible d'en déduire la distance entre les deux distributions. Pour cela, on passe par les z-scores des deux taux qui donnent l'éloignement, en unités d'écart-types, du critère de réponse par rapport

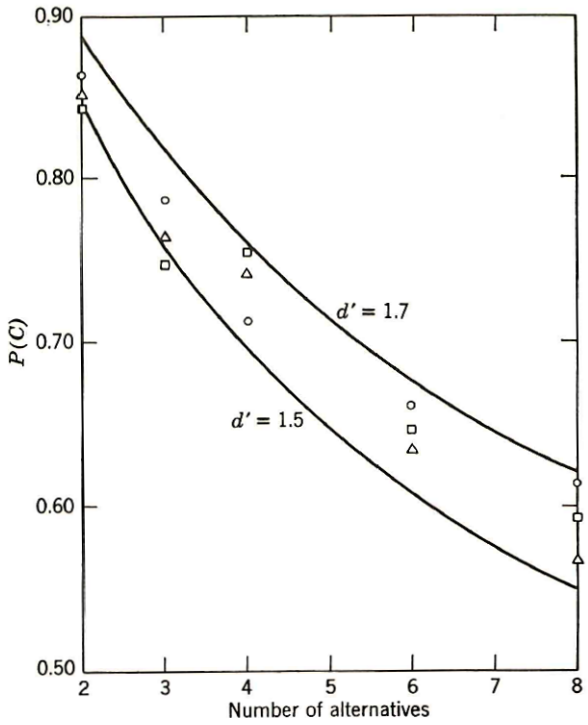


FIGURE 23 – Proportion correcte dans des tâches à choix forcé pour différents nombres d'alternatives. Les résultats de trois participants à une expérience de perception auditive sont représentés en fonction du nombre d'alternatives. Le nombre d'observations varie avec le nombre d'alternatives : 2-AFC–300, 3-AFC–500, 4-AFC–600, 6-AFC–900, et 8-AFC–1200. Les deux courbes théoriques sont des courbes de d'constant. Source : Green & Swets (1966).

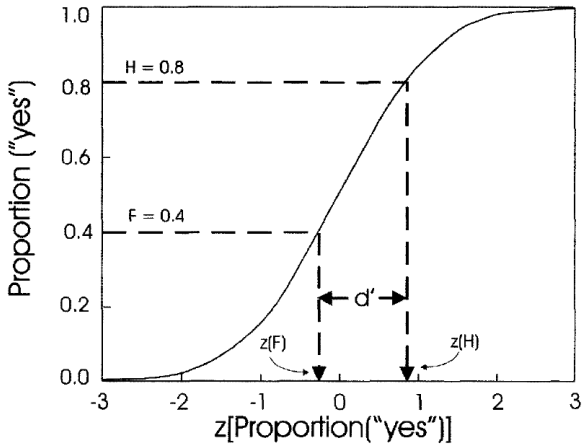


FIGURE 24 – Fonction de répartition de la loi normale centrée réduite. Cette fonction inverse peut être utilisée pour transformer les proportions de détections correctes et de fausses alarmes en z-scores, et la sensibilité est la différence entre les z-scores des détections correctes et des fausses alarmes. Source : MacMillan & Creelman (2005).

à la moyenne de la distribution concernée :

$$z = \frac{x - \mu}{\sigma}, \tag{2}$$

avec $x$ la variable dont on veut mesurer l'éloignement de la distribution (ici, le critère de réponse), $\mu$ la moyenne de la distribution, et $\sigma$ son écart-type. Les valeurs de z-scores sont données directement par la fonction de répartition de la distribution de probabilité (Figure 24). Par exemple, un z-score nul indique que le critère de réponse est confondu avec la moyenne de la distribution et donc que l'aire prend exactement la moitié de cette distribution, tandis qu'il est positif si l'aire dépasse la moitié de la distribution, et négatif sinon.

En l'appliquant aux réponses perceptives, le z-score du taux de détections correctes correspond à la distance (positive) entre le critère de réponse et la moyenne de la distribution du signal mélangé à du bruit, tandis que le z-score du taux de fausses alarmes correspond à la distance (négative) entre le critère de réponse et la moyenne de la distribution du bruit seul. La différence entre ces deux distances donne la distance entre les moyennes des deux distributions, c'est-à-dire la valeur de $d'$ (Figure 24).

Pour reprendre, avec $c$ correspondant à la valeur du critère de réponse, $\mu_1$ à

la moyenne de la distribution du signal mélangé à du bruit, $\mu_2$ à la moyenne de la distribution du bruit seul, et un écart-type $\sigma = 1$ pour les deux distributions, on peut vérifier que la sensibilité $d'$ est bien indépendante du critère de réponse :

$$d' = (c - \mu_1) - (c - \mu_2) = \mu_2 - \mu_1. \tag{3}$$

De façon plus générale, on écrit [14] :

$$d' = zscore(détections\ correctes) - zscore(fausses\ alarmes). \tag{4}$$

A nouveau, cette mesure permet d'observer que plus le participant détecte correctement la cible tout en limitant les fausses alarmes, plus la sensibilité augmente. Par ailleurs, dès que les taux de détections correctes et de fausses alarmes s'annulent, la sensibilité devient nulle, et ce, indépendamment du biais (i.e. on peut avoir des taux de détections correctes et de fausses alarmes équivalents, qu'ils soient élevés ou faibles). Un $d' = 1$ donne un autre point de repère, avec une proportion correcte de 70% environ. Qualitativement, le signe de $d'$ indique si le participant a suivi les bonnes consignes. Si le taux de détections correctes est plus grand que celui de fausses alarmes, $d'$ est positif, ce qui traduit le fait que le participant fait correctement la tâche, avec plus ou moins de difficultés. Par contre, un $d'$ négatif traduit quant à lui une tendance à confondre les stimuli ou les réponses.

On peut aussi constater que la mesure $d'$ diverge théoriquement vers l'infini, lorsque la proportion de réponses correctes tend vers 1 (ou 0). Ces valeurs peuvent apparaître pour des signaux très différents du bruit, si peu d'essais sont présentés, ou si le participant adopte un critère très libéral ou conservatif. Pour éviter cela, on ajuste généralement les performances parfaites ou nulles en les remplaçant respectivement par $1 - 1/(2N)$ et $1/(2N)$, avec $N$ le nombre d'essais. De cette façon, la mesure $d'$ est généralement bornée entre -4 et 4 environ. Stanislaw & Todorov (1999) proposent également d'autres solutions comme quantifier la sensibilité avec des mesures non-paramétriques, ou combiner les données de plusieurs participants avant de calculer les taux de détections correctes et de fausses alarmes plutôt que

14. Sur Matlab, on peut utiliser la fonction 'norminv' qui permet d'inverser la fonction de répartition de la loi normale.

de le faire a posteriori.

On peut noter aussi qu'il est possible de donner une estimation de la valeur de sensibilité $d'$ à partir de la seule mesure de proportion correcte. Pour cela, il faut faire l'hypothèse que le taux de détections correctes et de rejets corrects sont égaux, et donc que le biais est nul. Dans ce cas :

$$zscore(rejets\ corrects) = -zscore(fausses\ alarmes) \tag{5}$$

A partir de l'Equation 4, on peut écrire :

$$d' = 2 \cdot zscore[p(c)] \tag{6}$$

La différence entre les deux équations peut s'observer sur la Figure 25 en fonction des taux de détections correctes et de fausses alarmes. Elle est nulle lorsque les taux de détections et rejets corrects croissent conjointement de 0 à 1. Mais comme l'avaient déjà noté Macmillan & Creelman (2005), la sensibilité $d'$ est sous-estimée lorsque le biais n'est pas pris en compte dans les autres cas, et ce de façon particulièrement marquée lorsque l'un des deux taux, de détections correctes ou de fausses alarmes, prend des valeurs extrêmes proches de 0 ou 1 (sur les bords de la surface rouge, cf. Figure 25). Par exemple, pour un taux de détections correctes de 0.95 et un taux de fausses alarmes de 0.50, la différence de $d'$ en tenant compte ou non du biais atteint 0.45. Cela est dû au fait que la proportion correcte fait perdre les deux informations indépendantes des taux de détections correctes et de fausses alarmes.

**Calcul du biais.** Le participant peut avoir tendance à donner plus souvent une réponse qu'une autre, alors même que la probabilité qu'il s'agisse de la bonne réponse soit plus faible que pour l'autre. Il ne s'agit plus cette fois strictement des capacités perceptives du participant, mais de son niveau de confiance pour donner un jugement qui peut varier à cause de facteurs externes à la perception. On peut observer par exemple que selon la tâche, avec le même ensemble de stimuli de départ mais chacun étant présenté dans des proportions variables, la sensibilité du participant ne changera pas, contrairement à sa stratégie de réponse.

Le critère $c$ permet de quantifier ces tendances (Tableau 4). Il prend une valeur

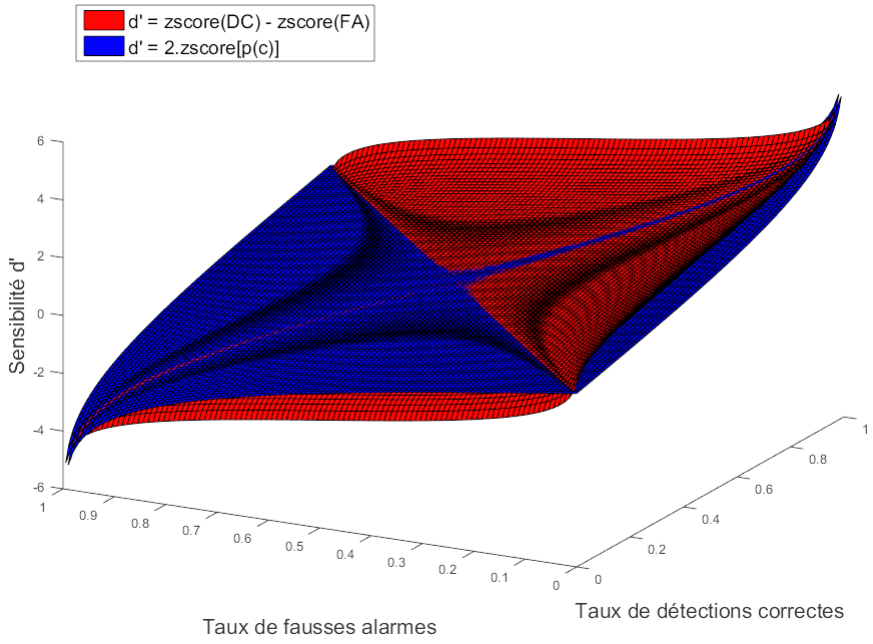


FIGURE 25 – Différences de sensibilités d' en fonction de la prise en compte ou non du biais. Les taux de détections correctes et de fausses alarmes varient entre 0 et 1. L'estimation de d' à partir de la proportion correcte p(c) néglige le biais, sachant qu'en réalité le biais augmente en négatif lorsque les taux de détections correctes (DC) et de fausses alarmes (FA) augmentent simultanément, tandis qu'il augmente en positif lorsque ces deux taux diminuent.

| Mesures de biais | Ecriture mathématique | Utilisation |
|---|---|---|
| Critère $c$ | $c = -\frac{zscore(DC)+zscore(FA)}{2}$ | Position du critère de réponse par rapport au point de croisement des distributions des deux classes de stimuli. |
| Critère relatif $c'$ | $c' = \frac{c}{d'}$ | Comparaison du biais entre des participants avec des sensibilités différentes. Comparaison du biais d'un même participant avant et après un entraînement. |
| Rapport de vraisemblance $\beta$ | $RV(x) = \frac{f(x\|S)}{f(x\|B)}$, soit, pour des distributions normales :$\beta = e^{cd'}$ | Autre mesure de biais interprétable sur la courbe ROC (pente de la courbe). |

Tableau 4 – Les différentes mesures de biais en TDS. DC : taux de détections correctes, FA : taux de fausses alarmes.

nulle quand le taux de détections correctes est égal à celui de rejets corrects (le point de croisement des deux distributions sur la Figure 22). Il est négatif quand le participant répond plus souvent qu'il entend la cible (participant libéral), tandis qu'il est positif dans le cas contraire (participant conservateur). Comme la mesure de sensibilité $d'$, le critère $c$ est borné entre -4 et 4 environ.

D'autres mesures de biais existent en TDS, elles sont mentionnées dans le Tableau 4. Le critère relatif $c'$ normalise le critère $c$ par la sensibilité $d'$, afin de pouvoir comparer les biais obtenus pour des participants de sensibilités différentes. Le rapport de vraisemblance, quant à lui, correspond au rapport entre les hauteurs des distributions sur l'axe d'intensité de la stimulation (Figure 22), en un point de l'axe donné. Dans le cas de distributions normales, il s'écrit $\beta$ et est fonction uniquement du critère $c$ et de la sensibilité $d'$. Si $c = 0$ (à l'intersection des distributions), alors $\beta = 1$ ; si $c > 0$ (participant conservateur), alors $\beta > 1$ ; et si $c < 0$ (participant libéral), alors $0 < \beta < 1$.

**Courbes ROC.** Avec un ensemble de stimuli donné, on peut supposer que les participants auront une sensibilité constante pendant toute la tâche de reconnais-

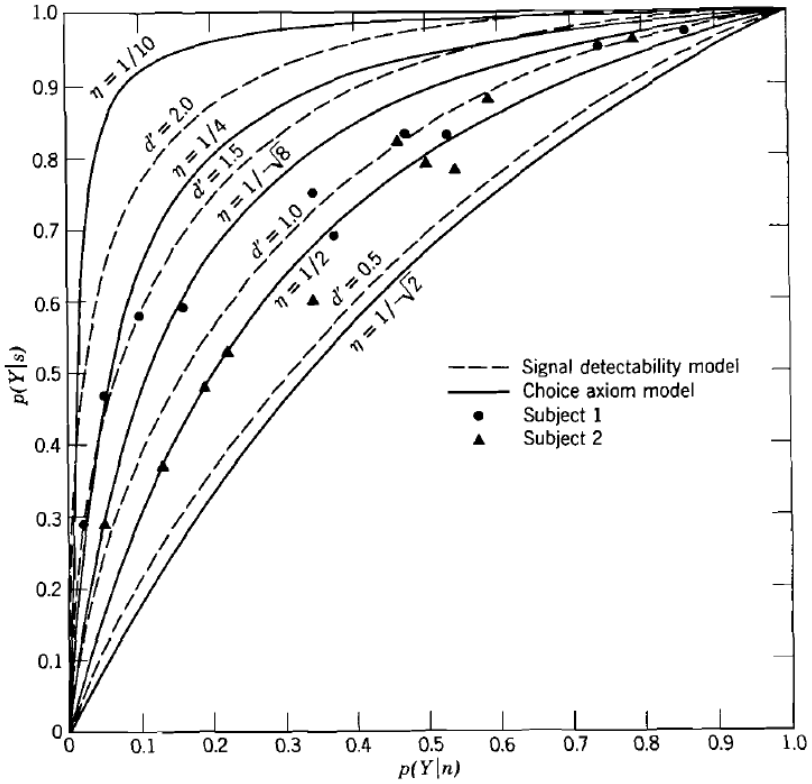


FIGURE 26 – Courbes d'isosensibilité pour une tâche Oui/Non. Les lignes en pointillés correspondent au modèle classique de la TDS, tandis que les lignes continues correspondent à un modèle alternatif de la Théorie du Choix, où $\eta$ est une mesure perceptive de similarité entre les stimuli. Les points ont été obtenus en présentant des tons purs dans du bruit, avec une probabilité variant entre 0.1 et 0.9 par pas de 0.2. Source : Luce (1963).

sance. Cependant, il est possible d'introduire un biais artificiellement en faisant varier les taux de présentation de la cible et du bruit. Les participants produiront donc plusieurs couples (taux de détections correctes, taux de fausses alarmes) à chaque étape du test, avec une sensibilité constante mais un biais variable. Ces couples permettent de construire des courbes d'isosensibilité (courbes ROC, pour *receiver operating characteristic*) dans un graphique avec le taux de fausses alarmes en abscisses et celui de détections correctes en ordonnées (Figure 26).

L'aire sous la courbe donne la sensibilité. Le niveau de la chance ($d' = 0$) est représenté par la diagonale principale. Pour une sensibilité positive, les points sont représentés au-dessus de cette diagonale, sinon en-dessous. On peut égale-

ment visualiser le biais du participant. Les points pour lesquels le biais est nul correspondent à ceux situés sur la diagonale $y = 1 - x$, c'est-à-dire là où les taux de détections et de rejets corrects sont égaux. Sous cette courbe, le participant est plutôt conservateur, au-dessus il est plutôt libéral. De façon équivalente aux courbes d'isosensibilité (cf. Figure 26), il est possible de tracer des courbes d'isobiais dans le même repère.

### 1.3.5 Méthodes de calcul de la sensibilité et du biais pour une tâche m-AFC

**Tables de sensibilité négligeant le biais.** Le calcul de la sensibilité et du biais est relativement aisé dans le cas de tâches 2-AFC mais peut s'avérer plus complexe dans le cas de tâches m-AFC (m $>$ 2). Hacker & Ratcliff (1979) ont proposé une table permettant d'estimer la valeur de la sensibilité $d'$ connaissant seulement la valeur de proportion correcte et le nombre d'alternatives dans la tâche. Il s'agissait d'une amélioration des valeurs proposées par Elliot (1964), dont la table était très utilisée à l'époque de leur article (cf. Hacker & Ratcliff, 1979), et qui l'est encore à l'heure actuelle (e.g. Seymour et al., 2009).

La table proposée par Hacker & Ratcliff (1979) répertorie les valeurs de $d'$ pour des proportions correctes allant de 0.01 à 0.99 (par pas de 0.01) et pour un nombre d'alternatives allant de 2 à 1000 (16 valeurs présentées). Cependant, comme dans l'approximation de la sensibilité à partir de la proportion correcte de Macmillan & Creelman (2005) (Equation 6), Hacker & Ratcliff (1979) se sont vus contraints de négliger le biais. Pour obtenir ces valeurs de $d'$, les auteurs ont estimé l'équation suivante :

$$p(c) = \int_{-\infty}^{+\infty} \phi(x - d') \cdot \Phi(x)^{M-1} \cdot dx \tag{7}$$

avec $p(c)$ la proportion correcte, $\phi$ et $\Phi$ respectivement l'ordonnée et l'aire sous une distribution normale réduite, et $M$ le nombre d'alternatives orthogonales [15].

15. L'hypothèse sous-jacente permettant d'établir la règle de décision, lorsque le nombre d'alternatives augmente en comparaison au cas Oui/Non, est décrite par Green & Swets (1966) : un vecteur d'observations contient M composantes indépendantes gaussiennes et de même variance dont la moyenne est nulle s'il s'agit d'un bruit, ou de valeur $d'$ s'il s'agit du signal cible. La

L'intégrale était approximée par un polynôme quadratique avec la méthode de Simpson entre -4.0 et 10.0 et sur 175 intervalles. D'autres algorithmes ont par la suite été proposés pour optimiser les calculs (e.g. Smith, 1982), ou pour d'autres types de tâches m-AFC (e.g. Craven, 1992). Les moyens de calcul actuels permettent de reproduire ces résultats avec plus de précision et de les étendre à n'importe quel nombre d'alternatives dans la tâche, ce qui devrait encourager l'application de la TDS selon Stanislaw & Todorov (1999) (un programme Matlab reproduisant le calcul de Hacker & Ratcliff (1979) est disponible en Annexe A.1).

Cependant, dans le cas de l'équation proposée par Hacker & Ratcliff (1979), l'hypothèse d'un biais nul reste contestable, comme elle l'était déjà avec l'équation de Macmillan & Creelman (2005). Dans la réalité, il y a de fortes chances que les taux de détections et de rejets corrects soient différents.

**Estimations jointes de la sensibilité et du biais.** L'estimation de la sensibilité en tenant compte du biais pour des tâches 2-AFC est possible dès lors que les taux de détections correctes et de fausses alarmes sont disponibles. Cela peut s'avérer plus compliqué pour des tâches m-AFC, avec m $>$ 2, pour lesquelles négliger le biais permet de simplifier le modèle pour estimer la sensibilité. Pourtant, les auteurs ont reconnu l'importance du biais pour ne pas fausser l'estimation de la sensibilité (e.g. Luce, 1963). DeCarlo (2012) a proposé une nouvelle méthode de calcul de la sensibilité avec deux approches computationnelles, incluant le biais dans son modèle et pour n'importe quel nombre d'alternatives.

**Modèle Oui/Non.** Le modèle de DeCarlo (2012) est composé d'une règle de décision et d'un modèle structurel. L'auteur part du cas le plus simple : une tâche Oui/Non. L'effet de la présentation d'un stimulus sur l'observateur est représenté par la variable aléatoire $\Psi$. La décision $Y$ de l'observateur est prise en fonction de cette variable, d'après la règle de décision suivante :
$Y = 1$ si $\Psi > c$,
$Y = 0$ si $\Psi \leq c$.
La position du critère de décision $c$ reflète le biais autour de la réponse 'Oui'

règle de décision consiste à choisir l'alternative dont l'intensité d'excitation prend la plus grande valeur.

ou 'Non'. Le modèle structurel, quant à lui, lie la variable psychologique $\Psi$ à l'évènement présenté $X$ :
$\Psi = d \cdot X + \varepsilon$,
avec $\varepsilon$ la variation aléatoire de la perception ($\varepsilon \sim \mathcal{N}(0,1)$). Donc la moyenne de la variable psychologique est nulle si le stimulus est du bruit ($X = 0$), ou égale à la valeur $d$ s'il s'agit du signal cible ($X = 1$). On déduit de la règle de décision et du modèle structurel l'équation suivante :

$$p(Y = 1|X) = \Phi(d \cdot X - c) \tag{8}$$

avec $\Phi$ la fonction de répartition de la distribution normale (qu'on pourrait remplacer par une autre distribution).

Avec les deux paramètres $c$ et $d$, le modèle TDS Oui/Non classique est complètement identifié :
$d = \Phi^{-1}[p(Y = 1|X = 1)] - \Phi^{-1}[p(Y = 1|X = 0)]$,
$c = -\Phi^{-1}[p(Y = 1|X = 0)]$,
où $d$, avec des distributions normales, est la mesure classique de $d'$.

DeCarlo (2012) décrit une méthode consistant à approcher le modèle TDS classique pour obtenir en plus les variances des paramètres estimés. Il écrit aussi la démonstration pour montrer de quelle manière le modèle est sous-identifié lorsqu'une seule observation est disponible (la proportion correcte p(c)), et pourquoi il est alors nécessaire de poser un biais nul.

**Modèle 3-AFC sans biais.** On peut généraliser la méthode précédente à des tâches m-AFC en négligeant le biais, en posant que les observateurs ont une information d'amplitude perceptive et qu'ils choisissent l'alternative associée à l'amplitude la plus grande (voir aussi Green & Swets, 1966). La règle de décision s'écrit alors :
$Y = 1$ si $\Psi_1 > \max(\Psi_2, \Psi_3)$,
$Y = 2$ si $\Psi_2 > \max(\Psi_1, \Psi_3)$,
$Y = 3$ si $\Psi_3 > \max(\Psi_1, \Psi_2)$.
Et le modèle structurel :
$\Psi_1 = d \cdot X_1 + \varepsilon_1$,
$\Psi_2 = d \cdot X_2 + \varepsilon_2$,

$\Psi_3 = d \cdot (1 - X_1 - X_2) + \varepsilon_3$.

En combinant la règle de décision et le modèle structurel, on obtient le modèle TDS 3-AFC sans biais :

$$\begin{aligned} p(Y = 1|X_1, X_2) &= p[\max(\Psi_2, \Psi_3) < \Psi_1] \\ &= p[(\Psi_2 < \Psi_1) \cap (\Psi_3 < \Psi_1)] \\ &= p[(d \cdot X_2 + \varepsilon_2 < d \cdot X_1 + \varepsilon_1) \cap (d \cdot (1 - X_1 - X_2) + \varepsilon_3 < d \cdot X_1 + \varepsilon_1)] \\ &= p[(\varepsilon_2 < d \cdot (X_1 - X_2) + \varepsilon_1) \cap (\varepsilon_3 < d \cdot (2 \cdot X_1 + X_2 - 1) + \varepsilon_1)]. \end{aligned}$$

Dans la probabilité conditionnelle, les évènements sont indépendants pour une valeur $\varepsilon_1 = e_1$ :

$$\begin{aligned} p(Y = 1|X_1, X_2, \varepsilon_1 = e_1) &= p[(\varepsilon_2 < d \cdot (X_1 - X_2) + e_1) \cap (\varepsilon_3 < d \cdot (2 \cdot X_1 + X_2 - 1) + e_1)] \\ &= F(d \cdot (X_1 - X_2) + e_1) \cdot F(d \cdot (2 \cdot X_1 + X_2 - 1) + e_1). \end{aligned}$$

En intégrant, on obtient :

$p(Y = 1|X_1, X_2) = \int_{-\infty}^{+\infty} F(d \cdot (X_1 - X_2) + e_1) \cdot F(d \cdot (2 \cdot X_1 + X_2 - 1) + e_1) \cdot f(e_1) \cdot de_1$.

Et, par analogie :

$p(Y = 2|X_1, X_2) = \int_{-\infty}^{+\infty} F(d \cdot (X_2 - X_1) + e_2) \cdot F(d \cdot (2 \cdot X_2 + X_1 - 1) + e_2) \cdot f(e_2) \cdot de_2$,

$p(Y = 3|X_1, X_2) = \int_{-\infty}^{+\infty} F(d \cdot (1 - X_2 - 2 \cdot X_1) + e_3) \cdot F(d \cdot (1 - X_1 - 2 \cdot X_2) + e_3) \cdot f(e_3) \cdot de_3$.

Deux des trois dernières équations suffisent à constituer le modèle TDS 3-AFC sans biais complet, car la somme des possibilités de réponses est fixe. On vérifie également l'intégrale de Hacker & Ratcliff (1979) (Equation 7) en calculant la proportion correcte à partir des taux de détections correctes :

$$\begin{aligned} p(c) &= \tfrac{1}{3}[p(Y = 1|X_1 = 1, X_2 = 0) + p(Y = 2|X_1 = 0, X_2 = 1) + p(Y = 3|X_1 = 0, X_2 = 0)] \\ &= \tfrac{1}{3}[\int_{-\infty}^{+\infty} F(d + e_1) \cdot F(d + e_1) \cdot f(e_1) \cdot de_1 + \int_{-\infty}^{+\infty} F(d + e_2) \cdot F(d + e_2) \cdot f(e_2) \cdot de_2 + \int_{-\infty}^{+\infty} F(d + e_3) \cdot F(d + e_3) \cdot f(e_3) \cdot de_3] \\ &= \int_{-\infty}^{+\infty} [F(x)]^2 \cdot f(x - d) \cdot dx. \end{aligned}$$

En généralisant à un modèle m-AFC, on retrouve bien la formulation de Hacker & Ratcliff (1979) (Equation 7).

**Modèle 3-AFC avec biais.** Des biais de réponse peuvent être introduits avec des paramètres de biais dans la règle de décision (DeCarlo, 2012). Pour trois classes de stimuli, deux paramètres suffisent, la troisième classe de stimuli servant de référence pour les deux autres. La règle de décision s'écrit cette fois :

$Y = 1$ si $\Psi_1 + b_1 > \max(\Psi_2 + b_2, \Psi_3)$,
$Y = 2$ si $\Psi_2 + b_2 > \max(\Psi_1 + b_1, \Psi_3)$,
$Y = 3$ si $\Psi_3 > \max(\Psi_1 + b_1, \Psi_2 + b_2)$.
Le modèle structurel est le même que celui pour le modèle sans biais. De façon similaire aux calculs précédents, on obtient le modèle TDS 3-AFC avec biais suivant :
$p(Y = 1|Z) = \int_{-\infty}^{+\infty} F(b_1 - b_2 + d \cdot Z + e_1) \cdot F(b_1 - d \cdot Z_1 + e_1) \cdot f(e_1) \cdot de_1$,
$p(Y = 2|Z) = \int_{-\infty}^{+\infty} F(b_2 - b_1 - d \cdot Z + e_2) \cdot F(b_2 - d \cdot Z_2 + e_2) \cdot f(e_2) \cdot de_2$,
avec $Z = X_1 - X_2$, $Z_1 = 1 - 2 \cdot X_1 - X_2$, $Z_2 = 1 - X_1 - 2 \cdot X_2$.

**Approximations computationnelles des modèles m-AFC avec biais.** DeCarlo (2012) donne deux méthodes computationnelles pour approcher les modèles m-AFC avec biais : une méthode par maximum de vraisemblance et une méthode par statistiques bayésiennes avec des algorithmes de chaînes de Markov Monte-Carlo (MCMC). L'auteur fournit des programmes informatiques en annexe de son article pour le cas 3-AFC (voir aussi Annexe A.2.1 pour le cas 4-AFC). Nous abordons ici seulement l'approche par estimation bayésienne. Pour cela, nous devons spécifier des a priori sur les paramètres du modèle, c'est-à-dire dans le cas 3-AFC avec biais : $d$, $b_1$, et $b_2$. Les $Y_j$ suivent des distributions de Bernouilli (à valeurs dans {0,1}), et les $\varepsilon_j$ suivent des distributions $\mathcal{N}(0, 1)$ indépendantes. L'avantage de la méthode bayésienne est d'introduire des connaissances préalables dans le calcul.

Comme on l'avait mentionné plus haut, si le biais expérimental est non-nul mais qu'il est négligé dans l'estimation de la valeur de $d'$, celle-ci est sous-estimée. Les simulations de DeCarlo (2012) le montrent avec des données de la littérature : pour une expérience 3-AFC, dans une condition, l'estimation de $d'$ avec la table de Hacker & Ratcliff (1979) est sous-estimée par rapport à la valeur obtenue par estimation bayésienne, alors que les valeurs de biais sont importantes, tandis que dans une autre condition où le biais est proche de zéro, les deux estimations de $d'$ sont équivalentes (cf. Annexe A.2.2). La méthode de DeCarlo (2012) permet en outre de donner les écart-types des estimations de la sensibilité et du biais.

**Modèle 4-AFC avec sensibilités et biais par catégorie.** Dans notre modèle TDS 4-AFC avec biais (Isnard et al., 2016), nous avons étendu le calcul de DeCarlo (2012) pour estimer une valeur de sensibilité par classe de stimuli. La règle de décision était obtenue avec la méthode proposée par DeCarlo (2012) :
$Y = 1$ si $\Psi_1 + b_1 > \max(\Psi_2 + b_2, \Psi_3 + b_3, \Psi_4)$,
$Y = 2$ si $\Psi_2 + b_2 > \max(\Psi_1 + b_1, \Psi_3 + b_3, \Psi_4)$,
$Y = 3$ si $\Psi_3 + b_3 > \max(\Psi_1 + b_1, \Psi_2 + b_2, \Psi_4)$,
$Y = 4$ si $\Psi_4 > \max(\Psi_1 + b_1, \Psi_2 + b_2, \Psi_3 + b_3)$,
Par contre, le modèle structurel incluait des sensibilités par catégorie :
$\Psi_1 = d_1 \cdot X_1 + \varepsilon_1$,
$\Psi_2 = d_2 \cdot X_2 + \varepsilon_2$,
$\Psi_3 = d_3 \cdot X_3 + \varepsilon_3$,
$\Psi_4 = d_4 \cdot (1 - X_1 - X_2 - X_3) + \varepsilon_4$.
Ce qui permet d'obtenir les équations du modèle TDS 4-AFC avec sensibilités et biais par catégorie :
$p(Y = 1|X_1, X_2, X_3) = \int_{-\infty}^{\infty} F(d_1 \cdot X_1 - d_2 \cdot X_2 + b_1 - b_2 + e_1) \cdot F(d_1 \cdot X_1 - d_3 \cdot X_3 + b_1 - b_3 + e_1) \cdot F(d_1 \cdot X_1 - d_4 \cdot X_4 + b_1 + e_1) \cdot f(e_1) \cdot d(e_1)$,
$p(Y = 2|X_1, X_2, X_3) = \int_{-\infty}^{\infty} F(d_2 \cdot X_2 - d_1 \cdot X_1 + b_2 - b_1 + e_2) \cdot F(d_2 \cdot X_2 - d_3 \cdot X_3 + b_2 - b_3 + e_2) \cdot F(d_2 \cdot X_2 - d_4 \cdot X_4 + b_2 + e_2) \cdot f(e_2) \cdot d(e_2)$,
$p(Y = 3|X_1, X_2, X_3) = \int_{-\infty}^{\infty} F(d_3 \cdot X_3 - d_1 \cdot X_1 + b_3 - b_1 + e_3) \cdot F(d_3 \cdot X_3 - d_2 \cdot X_2 + b_3 - b_2 + e_3) \cdot F(d_3 \cdot X_3 - d_4 \cdot X_4 + b_3 + e_3) \cdot f(e_3) \cdot d(e_3)$.

Comme trois nouvelles variables sont venues s'ajouter aux quatre variables du modèle initial de DeCarlo (2012), nous estimons d'abord la valeur des trois biais avec le modèle initial, sans inclure les sensibilités par catégorie. Puis, dans un deuxième temps, nous estimons la valeur des sensibilités par catégorie en fixant celle des biais, que nous avons obtenues avec la première estimation. Les programmes OpenBUGS correspondant sont donnés en Annexe A.2.1 [16]. Nous présentons également en Annexe A.2.2 les valeurs de sensibilités et de biais obtenues avec différentes méthodes dont celles présentées précédemment.

16. Sachant que les estimations des paramètres avec OpenBUGS peuvent prendre parfois plusieurs dizaines de minutes pour un seul participant et une seule condition (dans notre cas, environ 30 min en 4-AFC), il peut être préférable de faire communiquer OpenBUGS avec Matlab à l'aide de l'interface 'matbugs', afin de mettre en forme les données dans Matlab avant de les lancer dans une routine incluant tous les participants et toutes les conditions.

Ce modèle a été développé en vue d'obtenir une mesure de la performance des participants par catégorie auditive (i.e. par classe de stimuli), dans un cas bien précis où il était intéressant de la comparer avec une mesure de distance auditive calculée sur les signaux. L'évolution des modèles de la TDS doit notamment aux ressources computationnelles au départ limitées. Néanmoins, à l'heure actuelle, certaines approximations (e.g. négliger le biais) n'ont plus véritablement lieu d'être et devraient être mises à jour. En revanche, il est vrai que la pluralité des modèles et des théories complémentaires ou concurrentes peut complexifier leur lecture.

## 1.4 Compléments d'analyses (2/2) : modèle de distance auditive

### 1.4.1 Représenter des distances perceptives

Dans un grand nombre d'études perceptives, l'objectif est de faire la correspondance entre deux mesures de distances : celle calculée entre des stimuli, et celle mesurée par des moyens expérimentaux pour évaluer des capacités cognitives (e.g. mesures comportementales, mesures cérébrales). Dans le cas du timbre d'instruments de musique par exemple, les auteurs ont débattu sur l'ensemble des meilleurs corrélats acoustiques pouvant expliquer les données perceptives représentées dans des espaces de timbre multidimensionnels (e.g. Grey, 1977).

L'analyse des mesures expérimentales (comportementales, cérébrales) sont généralement l'objet central d'une étude, à travers l'observation de différences significatives entre des données expérimentales quantifiées sur des échelles prédéfinies. En revanche, la caractérisation précise des stimuli est laissée davantage à la guise des expérimentateurs. Il arrive même que les stimuli soient si bien égalisés qu'on ne trouve pas de dimensions acoustiques les séparant, alors qu'on observe pourtant des différences comportementales ou cérébrales significatives, ce qui peut seulement laisser suggérer un traitement auditif "complexe" (e.g. Murray et al., 2006). Néanmoins, les stimuli restent à l'origine des résultats expérimentaux et doivent nécessairement présenter des différences acoustiques pour générer ces différences perceptives (en faisant l'hypothèse que les autres conditions expérimentales sont contrôlées par ailleurs).

Deux points clés permettent donc d'expliquer a posteriori les résultats expéri-

mentaux : la caractérisation des stimuli, et le lien entre cette caractérisation et les données expérimentales. Calculer des distances (qui peuvent s'exprimer en unités arbitraires) entre ces données permet de les comparer indépendamment de leur provenance (voir aussi Kriegeskorte et al., 2008). On décrit ci-dessous plusieurs stratégies pour calculer des distances entre des stimuli, ainsi que pour faire le lien entre des distances de différentes natures (e.g. entre des stimuli et des données expérimentales ou issues de modélisations).

**Calculs de distances entre des stimuli sonores.**

**Distances entre percepts auditifs dans un espace de timbre multidimensionnel.** Les études de timbre classiques révèlent des corrélats acoustiques de dimensions d'un espace perceptif multidimensionnel où sont représentés les sons testés, distants les uns des autres en fonction de leur similarité perçue. Misdariis et al. (1998) ont proposé de résumer les corrélats acoustiques du timbre dans une seule mesure de distance perceptive. Pour cela, les auteurs ont utilisé le CGS (SC, en Hz), le temps d'attaque logarithmique (LT, en $\log(s)$), l'irrégularité de l'enveloppe spectrale (SI, en dB), et le flux spectral (SF, sans unité). La distance perceptive DIST résulte du calcul de la distance euclidienne entre les stimuli perçus et répartis dans l'espace multidimensionnel obtenu expérimentalement :

$$DIST = \sqrt{3.5385 \cdot 10^{-5} \cdot SC^2 + 15.5236 \cdot LT^2 + 0.01188 \cdot SI^2 + 2728.7 \cdot SF^2}.$$

Elle permet notamment d'observer le poids que prend chaque corrélat acoustique dans la perception de la similarité entre paires de stimuli. Par exemple, le poids du CGS est très faible comparé à celui du flux spectral. Cette mesure semble découler assez naturellement des résultats des études de timbre et pourrait permettre d'estimer la reconnaissabilité d'un ensemble de stimuli auditifs. Elle ne semble pourtant pas avoir été proposée ailleurs, probablement car ces estimations restent fortement dépendantes des stimuli utilisés dans le test perceptif, ainsi que des propositions de corrélats acoustiques qui varient entre les études sur le timbre.

**Distances entre représentations acoustiques.** Pour contrôler les caractéristiques acoustiques de leurs stimuli, Murray et al. (2006) réalisent des tests

de Kolmogorov-Smirnov entre les paires de bins temps-fréquence (86 Hz×5.8 ms) des spectrogrammes moyens de chacune des deux catégories sonores testées (objets vivants et objets fabriqués). Les seules différences significatives sont sur des durées courtes, 125 ms après le début du son, et pour des fréquences supérieures à 4 kHz. Dans leurs analyses acoustiques, les auteurs complètent ce test par un test sur le HNR moyen de chaque catégorie, qui n'est pas significatif. Selon les auteurs, ces analyses de caractéristiques bas-niveaux ne suffisent pas à expliquer les différences de traitements cérébraux de chacune des deux catégories testées. Des analyses comparables ont également été utilisées dans d'autres études (e.g. Charest et al., 2009 ; De Lucia et al., 2010).

Cependant, les conclusions données à partir de tests statistiques comparant des spectrogrammes acoustiques sont parfois quelque peu hâtives, puisque les seuils perceptifs pour différencier des caractéristiques acoustiques ne relèvent pas de seuils statistiques prédéfinis (e.g. valeur $p < 0.05$). D'autant moins que les représentations acoustiques utilisées pour réaliser ces tests ne tiennent pas compte des transformations non-linéaires du traitement auditif, susceptibles d'élargir les différences entre les représentations des sons, qui pourraient dès lors devenir significatives avec le même critère statistique choisi.

**Distances entre représentations auditives.** Des mesures de distances auditives, sur des représentations reproduisant des traitements auditifs, devraient permettre d'approcher davantage les résultats perceptifs. Pour identifier les mécanismes cérébraux sous-jacents à la discrimination des sons naturels, Woolley et al. (2005) ont comparé les propriétés d'accord spectro-temporel de neurones auditifs chez une espèce d'oiseaux en fonction du contenu statistique de sons naturels. Les vocalisations, en particulier, permettent de communiquer efficacement et se distinguent des autres sons. Pour faciliter cette discrimination auditive, les neurones auditifs de haut-niveau semblent maximiser les différences acoustiques entre les sons.

Pour valider cette hypothèse, les auteurs ont quantifié la discriminabilité cérébrale d'après la répartition des modulations spectro-temporelles sur un spectre de modulation. Ces réponses cérébrales sont obtenues pour des segments de sons de 100 ms en convoluant chaque stimulus avec un ensemble de STRFs et com-

parées entre elles par des distances euclidiennes. Les auteurs ont montré que les distances calculées sont discriminantes entre les réponses cérébrales de différentes catégories sonores (chants d'oiseaux, parole humaine, sons environnementaux) et aussi entre celles pour différents sons d'une même catégorie, grâce à un mécanisme d'extension des différences acoustiques entre les catégories sonores à travers les modulations fréquentielles, tandis que l'information redondante est supprimée.

Dans le cas de la perception auditive humaine et pour modéliser des performances issues de tâches auditives complexes, certains auteurs préfèrent utiliser des représentations de la périphérie auditive pour lesquelles les résultats de la littérature ont été répliqués et validés. Par exemple, Giordano et al. (2010) ont comparé des jugements de dissemblances entre des sons avec des distances euclidiennes calculées entre des spectrogrammes auditifs (30 bandes fréquentielles en tiers d'octaves réparties sur une échelle logarithmique). Les deux mesures corrèlent significativement entre elles, bien que faiblement (environ 0.3), mais plus qu'avec une mesure de distance sémantique, sachant que c'est cette comparaison entre distances acoustique et sémantique qui intéresse les auteurs.

Pour modéliser la perception de sons courts, Bigand et al. (2011) ont réalisé une analyse sur les patterns d'excitation (i.e. puissance RMS calculée sur 80 bandes fréquentielles) des stimuli par catégorie sonore (voix, musique, sons environnementaux). Après quoi, une analyse en composantes principales leur a permis de repérer les bandes fréquentielles contribuant le plus aux différences auditives calculées (celles centrées sur 247 et 1000 Hz), puis de répartir les sons dans l'espace des composantes principales. A la fois les distances entre les sons d'une même catégorie (plutôt faibles) et les distances entre les sons de différentes catégories (plutôt élevées) contribuent à la modélisation des données perceptives.

Agus et al. (2012) ont également proposé une mesure de distance auditive sur la base d'une représentation auditive périphérique (STEPs ; Moore, 2003). Les stimuli, des sons de voix et d'instruments, sont égalisés pour ne laisser que les indices du timbre pour réaliser la tâche de reconnaissance auditive proposée. Les voix sont reconnues plus rapidement que les instruments. Les auteurs n'en restent pas à l'observation de ces différences perceptives en fonction des catégories sonores malgré l'égalisation des stimuli. En effet, leur calcul de distance auditive permet de constater que les temps de réaction pour la voix sont toujours rapides, et ce, quelle

que soit la distance auditive par rapport aux distracteurs, tandis que les temps de réaction pour des sons d'instruments dépendent de cette distance. Les conclusions sont plus convaincantes lorsque sont proposées des origines possibles ayant pu induire les différences perceptives, plutôt que lorsqu'elles suggèrent seulement un traitement auditif "complexe".

**Correspondance entre différents types de distances expérimentales.** L'intérêt des mesures de distances auditives est particulièrement manifeste dans le cas des études de neuroimagerie ou de neurophysiologie cherchant à faire le lien entre des stimuli ou des mesures comportementales et des mesures d'activité cérébrale (e.g. Halpern et al., 2004 ; Zatorre et al., 2004 ; Formisano et al., 2008 ; Kriegeskorte et al., 2008 ; Staeren et al., 2009 ; Patil et al., 2012 ; Giordano et al., 2013). L'objectif de ces études est généralement de déterminer quelle information est encodée dans les patterns d'activité cérébrale. La tâche des participants se résume généralement à écouter passivement les sons, donc les résultats dépendent principalement des sons (en comparaison à des études mettant en œuvre des tâches comportementales plus élaborées).

Kriegeskorte et al. (2008) ont proposé un cadre d'analyse assez général, l'analyse de similarité représentationnelle (RSA, pour *representational similarity analysis*), pour faire la correspondance entre des mesures d'activité cérébrale, des mesures comportementales, et des modélisations computationnelles. La RSA met donc en commun différentes modalités de mesures, mais aussi différents participants ou espèces animales. Elle passe par l'abstraction de l'information d'une représentation donnée dans des matrices de dissimilarité représentationnelle (RDM, pour *representational dissimilarity matrix*).

Les auteurs ont appliqué la RSA à des mesures issues de données IRMf ainsi que de modèles computationnels, et correspondant à des représentations d'objets visuels. Cette technique permet ainsi de comparer des patterns d'activité cérébrale à travers une RDM en fonction des conditions testées (Figure 27). Différentes matrices peuvent ensuite potentiellement être comparées, par exemple à l'aide de corrélations, pour différents types de mesures de l'activité du système nerveux central (IRMf, EEG, MEG, etc.) ou issues de modèles computationnels.

La RSA a notamment été reprise par Giordano et al. (2013) pour évaluer l'ap-

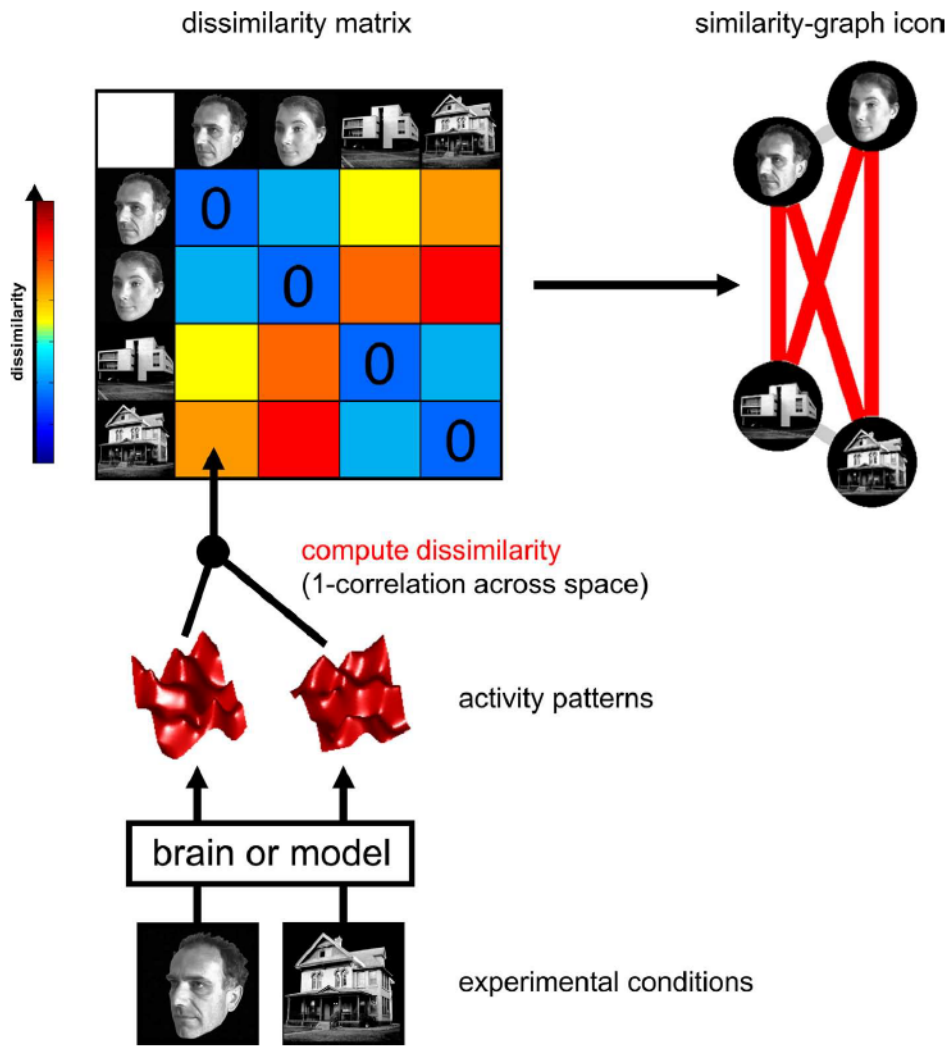


FIGURE 27 – Calcul de la matrice de dissimilarité représentationnelle (RDM). Pour chaque paire de conditions expérimentales, les patterns d'activité associés (mesurés à partir de l'activité cérébrale ou issus d'un modèle) sont comparés par corrélations spatiales. Les mesures de dissimilarités entre toutes les paires sont rassemblées dans la RDM. Un graphique de similarité peut permettre de visualiser la représentation d'un petit nombre de conditions (en haut à droite). Source : Kriegeskorte et al. (2008).

partenance cérébrale de percepts auditifs à des catégories sonores (voir aussi Formisano et al., 2008 ; Staeren et al., 2009). Les auteurs mesurent la différenciation des représentations corticales en réponse à des catégories sonores en fonction de la diversité des stimuli. L'analyse consiste à associer des RDMs avec des matrices de dissemblance de caractéristiques de stimuli (SDMs, *stimulus-feature dissimilarity matrices*). Avec leurs stimuli très hétérogènes, les auteurs ont pu tester 12 SDMs bas-niveaux basées sur le niveau sonore, le CGS, la hauteur, et le HNR, et 12 SDMs catégorielles basées sur le caractère vivant ou non, humain ou non, vocal ou non, des sons. La RSA de la sélectivité corticale a permis de mettre en évidence un encodage de certaines caractéristiques bas-niveaux et un encodage abstrait de certaines catégories sonores. L'utilisation de stimuli naturels riches en composantes spectro-temporelles permet ainsi de tester la sensibilité corticale à des différences acoustiques bas-niveaux, pouvant impliquer des recouvrements d'activations corticales (Giordano et al., 2013).

### 1.4.2 Construction du modèle de distance auditive entre catégories sonores

Dans notre étude (Isnard et al., 2016), nous nous sommes basés sur le calcul de distances auditives proposé par Agus et al. (2012) tout en considérant que des similarités auditives calculées entre des paires de représentations auditives des stimuli peuvent donner un indice sur l'appartenance de ces sons à une même catégorie ou non.

Tout d'abord, nous avons calculé une représentation auditive de tous les stimuli sous forme de STEPs (Moore, 2003). Le calcul des distances auditives a ensuite été réalisé à l'aide d'un algorithme de déformation temporelle dynamique (DTW, pour *dynamic time warping*). Cet algorithme consiste à évaluer le degré d'alignement entre des séries temporelles (ici, des spectrogrammes) en donnant une mesure de la déformation nécessaire de ces séries pour optimiser leur alignement, et estimer ainsi leur similarité. Cet algorithme a notamment été utilisé en reconnaissance automatique de la parole pour comparer des formes d'onde (Sakoe & Chiba, 1978). Dans notre cas où l'on compare des représentations temps-fréquence, le calcul d'alignement dépend toujours de la dimension temporelle, mais est fortement contraint par la dimension fréquentielle. Des exemples avec des matrices

particulières sont disponibles en Annexe B.

Une deuxième étape du calcul est plus spécifique à notre étude, elle porte sur la reconnaissance de catégories sonores. En effet, un certain nombre d'auteurs ont noté que les similarités acoustiques sont plus grandes entre les sons d'une même catégorie qu'entre ceux de catégories différentes (e.g. Murray et al., 2006 ; Staeren et al., 2009 ; Bigand et al., 2011). Cette remarque illustre les deux aspects dont nous avons voulu tenir compte dans notre modèle de distances auditives entre catégories sonores. Les distances d'un son par rapport aux sons d'une autre catégorie sont soustraites de celles par rapport aux sons de sa propre catégorie.

La forte corrélation que nous avons constatée entre les deux mesures de distances auditives et de performances de reconnaissance auditive (sensibilité $d'$) illustre la possibilité de faire correspondre des mesures obtenues sur différents types de représentations expérimentales (physiques, comportementales, etc. ; Kriegeskorte et al., 2008). Cette correspondance confirme que la discriminabilité perceptive entre des stimuli peut effectivement se concevoir comme une mesure de distance perceptive avec des propriétés géométriques (Macmillan & Creelman, 2005). Enfin, ces résultats semblent traduire des stratégies de reconnaissance auditive, et suggèrent que l'appartenance perceptive d'un son à une catégorie revienne à le comparer une référence moyenne interne à laquelle appartient le son, par rapport à une autre référence moyenne des catégories auxquelles ne peut pas appartenir le son. Ces références pourraient avoir été acquises par l'exposition continue à l'environnement sonore naturel.

# 2 Temps de traitement de sons courts

Ce travail a fait l'objet de communications orales :

- lors du 24$^{e}$ colloque de l'ED3C : "The timing of sound recognition in normal human adult : a behavioral study" (poster), le 21 mars 2016 ;
- lors du 13$^{e}$ Congrès Français d'Acoustique (CFA) : "Nouveau paradigme pour l'étude du temps de traitement auditif de sources sonores : Rapid Audio Sequential Presentation (RASP)", le 15 avril 2016.

Il est par ailleurs destiné à être publié dans différentes revues scientifiques. La première série d'expériences est présentée ici en anglais en vue d'une soumission prochaine.

## 2.1 Résumé

La reconnaissance auditive est robuste à de très fortes dégradations du signal acoustique, en particulier dans le domaine temporel. Des études ont montré que des sons extrêmement courts (de l'ordre de quelques millisecondes) peuvent être reconnus. Mais le traitement auditif met un certain temps pour traiter les stimuli auditifs, au départ pour former un percept statique d'un stimulus qui varie dans le temps (Patterson, 2000). Dans cette étude, nous utilisons le paradigme RASP (cf. paragraphe I.3.3.1) qui permet d'évaluer le temps mis par le système auditif pour traiter de courts extraits de sons.

Le corpus de sons est composé de sons de voix chantées et d'instruments de musique. Une première expérience vise à contrôler les performances des participants en fonction de la durée des sons, afin de vérifier qu'ils sont capables de reconnaître des sons courts présentés de façon isolée. Dans une deuxième expérience, des séquences de sons courts (16 et 32 ms) sont présentés aux participants selon le paradigme RASP (paragraphe II.2.2). Ces séquences sont composées d'une cible à reconnaître au milieu d'un flux de distracteurs (voix cible et instruments distracteurs, et vice-versa). Enfin, nous avons repris les deux premières expériences en les complétant par une mesure de temps de réaction (paragraphe II.2.3), afin de déterminer avec quelle rapidité les participants sont capables de reconnaître des sons présentés de façon isolée ou dans des séquences RASP. A chaque fois, les participants pouvaient reconnaître une cible dans une séquence de sons distracteurs présentée à un taux de présentation élevé, réduisant pourtant son temps de traitement à un intervalle très court. De plus, pour un taux de présentation donné, la reconnaissance d'une cible voix était meilleure que celle d'une cible instrument.

Des compléments d'analyses sont proposés pour l'analyse de la reconnaissance de sons courts isolés (paragraphe II.2.4) et de la reconnaissance d'une cible dans une séquence RASP (paragraphe II.2.5).

L'ensemble de ces résultats complètent ceux obtenus dans l'étude précédente, en explicitant l'efficacité de la reconnaissance auditive avec un nouvel éclairage apporté sur les différentes dynamiques temporelles impliquées dans ces traitements.

## 2.2 "Time course of auditory recognition using short natural sounds : the RASP paradigm"

### 2.2.1 Introduction

The recognition of natural objects in everyday life seems easy and instantaneous. Contrary to what we experience, however, the recognition process takes time. In a seminal paper, Massaro (1972a) proposed a new account to the understanding of processing time in auditory perception. Based on the argument that sound features cannot be processed as they reach the ear and the auditory system, because that would require instantaneous perception, he proposed that perceptual units of auditory perception are stored in a preperceptual store for further processing. A number of questions arise from this theoretical account of auditory processing, among which the minimal features necessary for recognition (see Gray, 1942 ; Robinson & Patterson, 1995b ; Suied et al., 2014 ; Isnard et al., 2016), and the processing time of these minimal features (Woods & Alain, 1993 ; Woods et al., 1993 ; Suied et al., 2013a). Using a new auditory paradigm, the RASP paradigm (Rapid Audio Sequential Presentation ; Suied et al., 2013a), the current study focus on the question of the processing time of natural sounds like voices and instruments.

The RASP paradigm is similar to the classic visual task termed Rapid Serial Visual Presentation (RSVP) (e.g. Chun & Potter, 1995 ; Subramaniam et al., 2000 ; Buffat et al., 2012). In the RSVP task, participants are presented with a rapid succession of flashed images, and have to detect or recognize a given target within this stream. RSVP is thus a masking paradigm, intensively used to study the limited processing time of the visual system to recognize a given target image. Physiological data have also been collected with RSVP ; the stimulus-onset asynchrony for which performance falls to chance have been shown to reflect neural time constants in the recognition process (Keysers et al., 2001).

A first auditory analogue was proposed by Woods & Alain (1993) to study the duration of feature processing with acoustical stimuli presented rapidly. They used short pure tones (10 ms) presented with ISIs ranging from 40 to 200 ms, and recorded event-related potentials to investigate the nature of the auditory processing (parallel or serial). Participants had to focus on combinations of basic

features (frequency, localization, and intensity) which defined the targets. Authors observed that if sounds in the sequences presented some or all the features corresponding to the target, then the evoked cortical waves had a longer duration than the ISI in the sequence. This overlap was therefore interpreted as a parallel processing of the sounds rather than a serial one.

With the RASP paradigm, Suied et al. (2013a) transposed the RSVP paradigm to the auditory world in a more systematic way. The authors assessed the processing time involved in the recognition of natural stimuli. They presented sequences of short distractor sounds with, in half of the trials, a short target sound at a random position in the sequence. They showed that performances decreased with an increased presentation rate, but recognition was still possible for presentations rates of up to 30 sounds per second, suggesting that the recognition of natural sounds was extremely efficient, both in terms of the acoustic duration and of the underlying time constants.

In the present study, we investigated the time course of natural sound recognition with the RASP paradigm, and extended the results of Suied et al. (2013a). Here, the sound corpus was balanced with the same number of voice and instrument sound sources, and presented a very large acoustical variability. Nevertheless, all the sounds were selected in the same pitch range, and were presented with the same duration and loudness, letting participants with only timbre cues to perform the tasks (e.g. Patil et al., 2012). In a control experiment, the participants were first tested in their ability to recognize short voice and instrument sounds presented in isolation. Then, 3 RASP experiments were run. In each of these, the recognition time of two different categories were compared : voice targets in a stream of musical instruments, and instrument targets in a stream of voice distractors. The 3 experiments were designed to study the processing time of auditory recognition, by varying the presentation rate (number of sounds per second). By comparing several experimental conditions through the experiments, we controlled that the processing we were studying was only due to the time necessary for the auditory system to process timbre cues, and not to other possible bias that could have played a role (frequency masking, quantity of information...).

### 2.2.2 Methods

**Participants.** Thirty-eight participants were recruited for this study. Eight of them did not take part in the experiments because they had more than 20 dB HL hearing loss at one or more of the audiometric frequencies between 0.125 and 4 kHz (audiograms performed with an Echodia Elios audiometer). The 30 remaining participants were first included in the control experiment. Based on the exclusion criteria defined in the Procedure, only 24 of them (14 women ; mean age = 24.0 ± 3.2 ; range = 18-29 years) could participate in the main experiments. Eight took part in the 'presentation rate' experiment, 7 in the 'number of sounds' experiment, and 9 in the 'pitch' experiment. The Institutional Review Board of the French Institute of Medical Research and Health ethically approved this specific study prior to the experiment (opinion n°15-211), and all participants provided written informed consent to participate. They were compensated for their participation.

**Stimuli.** Briefly, all stimuli were sequences of very short snippets of natural sounds. Sound samples were extracted from the RWC Music Database (Goto et al., 2003). Two sound categories were used : singing voices and musical instruments, with four original sound sources in each category. The voice sounds were two men singing the vowels /a/ and /i/, and two women singing the vowels /e/ and /o/. The instrument sounds were : a bassoon, a clarinet, a piano, and a saxophone. Twelve pitches were kept for each sound source, from A3 to G#4 (see Agus et al., 2012, for details).

For the control experiment, the short excerpts of sounds were presented individually in order to be sure that the participants could recognize the sound presented in isolation. Seven sound durations were tested : 2, 4, 8, 16, 32, 64, and 128 ms. The sounds were gated at these durations with a Hanning window. The starting point of the gating window was randomly chosen between 0 and 100 ms from the onset of the sound on each trial. Finally, stimulus intensities were normalized by their root-mean-square level and divided by the square root of their duration (see Suied et al., 2014, for details).

For the three main RASP experiments, sequences of these short gated sounds were created. How the sequences were constructed is described in the Procedure (see below). For each sequence, gated sounds were generated following the proce-

dure used for the gating control experiment (random beginning of the Hanning window, normalization...), and with the same sound corpus.

**Apparatus.** Participants were tested individually in a double-walled Industrial Acoustics (IAC) sound booth. Stimuli were presented through a RME Fireface digital-to-analog converter at 16-bits resolution and a 44.1 kHz sample-rate. They were presented diotically through a Sennheiser HD 650 headphone at a comfortable loudness level (~70 dB A). No time limit was imposed and a visual feedback (green for correct responses, red for incorrect) was provided once the participant had responded.

**Procedure.** First, a control experiment was run, in which all participants were tested in their ability to recognize very short sounds presented in isolation. On each trial, participants heard a short sound which could be either a voice or an instrument. They had to indicate whether the sound was a voice or an instrument (two-alternative forced-choice task). The two sound categories and the seven sound durations were presented in a randomized order, with equal probability. For each trial, the pitch and the beginning of the gating were chosen randomly. It contributed to generate a large acoustical variability in the sound corpus, since each short snippet of sound was different on each trial, and for each participant. The number of repetitions was reduced mid-experiment, but with no major impact on the results (see Results section below). The first ten participants performed 44 repetitions for each category and for each sound duration, whereas the twenty following participants performed 24 repetitions. Before the test, the first ten participants were familiarized with the task with a short training consisting in 28 trials, for which the seven sound durations were presented in a decreasing order, and with 4 trials per duration. The twenty following participants performed the same training with, first, the 8 original sounds with a 250-ms duration.

For the three main experiments ('presentation rate', 'number of sounds', and 'pitch'), sequences of short natural sounds were presented in rapid succession. Each short sound was gated with the same procedure as the one described for the control experiment ; the same sound database was used, with singing voices and instruments sounds. For all experiments, the task was similar : participants

heard a sequence of short sounds, and had to decide whether a target sound was present in the sequence or not (yes/no task). In 50% of the trials (the 'no' trials), sequences were composed of distractor sounds only ; in the other 50% (the 'yes' trials), one target sound was embedded in the sequence (except at the first and last positions of the sequence). Target and distractors were alternatively voice or instrument sounds, for each experiment. The presentation rate of the sequence (from slow to very fast sequences) was also varied through all three experiments.

The 'presentation rate' RASP experiment was performed to investigate the time limits of the sound recognition process, by limiting the potential factors to timbre factors : pitch was randomly varied and all individual durations were the same : 16-ms sounds. The 'number of sounds' and 'pitch' experiments were carried out to rule out two other important mechanisms that could have played a role as well. Firstly, the drop in performance with an increased presentation rate could be due to memory limitations rather than a reduced available time to analyze each sound in the sequence, because, with fixed-durations sequences, the number of sounds increased as the rate increased. This was tested in the 'number of sounds' experiment, by comparing fixed-duration sequences (500 ms) with fixed-number of sounds sequences (7 sounds). Secondly, forward masking could play a role as well, by distorting the spectrum of individual sound in the sequence, thus limiting its audibility. To test this, we compared, in the 'pitch' experiment, random-pitch sequences, where each individual sound in the sequence had a randomly selected pitch, with fixed-pitch sequences, where the pitches of every sound in the sequence were the same (while varying from sequence to sequence). An impact of forward-masking on performance would predict worse results for the fixed-pitch sequences, because of a larger frequency overlap between successive sounds. A summary of the different conditions used for each experiment is given in Table 5.

For the 'presentation rate' experiment, participants performed 44 repetitions for each target category (Voices, Instruments) and for each presentation rate. All participants performed both types of blocks (voices as a target and instruments as a target), but in a counterbalanced order between them. Presentation rates (from 5.3 Hz to 60 Hz ; see details in Table 5) were randomized within each block. A specific training with the adequate instruction (was there a voice or was there an instrument within the sequence) was done just before the corresponding blocks.

| Conditions | RASP experiments | | |
|---|---|---|---|
| | 'presentation rate' | 'number of sounds' | 'pitch' |
| Sound duration | 16 ms | 32 ms | 32 ms |
| Sequence type | Fixed duration | Fixed duration vs. fixed number of sounds | Fixed number of sounds |
| Pitch | Randomized | Randomized | Randomized vs. fixed |
| Presentation rate | 5.3, 7.5, 10.6, 15, 21.2, 30, and 60 Hz | 5.3, 7.5, 10.6, 15, 21.2, and 30 Hz | 5.3, 7.5, 10.6, 15, 21.2, and 30 Hz |

Tableau 5 – Tested conditions in the main experiment with sequences of short sounds presented rapidly.

For each target category block, the training started with sequences presented with an increasing presentation rate from 5.3 to 15 Hz ; firstly, individual sounds had a 64-ms sound duration, then 32-ms (64 trials) ; secondly, they performed 112 trials of the formal test still as a training.

For the 'number of sounds' and the 'pitch' experiments, participants performed 24 repetitions for each category, for each presentation rate, and for each condition : fixed duration/fixed number of sounds and randomized/fixed pitch respectively. As for the first 'presentation rate' experiment, presentation rates (from 5.3 Hz to 30 Hz ; see details in Table 5) were randomized within each block. For the 'number of sounds' experiment, four types of blocks were possible : target Voice with 'fixed duration' sequences, target Voice with 'fixed number of sounds' sequences, target Instruments with 'fixed duration', and target Instruments with 'fixed number of sounds'. These four conditions were counterbalanced between participants. A similar counterbalanced presentation of blocks was performed for the 'pitch' experiment (including the 'random-pitch' and the 'fixed-pitch' conditions). A similar training as for the first 'presentation rate' experiment was conducted before each experiment 'number of sounds' and 'pitch'.

All experiments lasted about 2 hours and half in total (including, for each participant, audiometric measures, gating, and one of the RASP experiment).

### 2.2.3 Results

For all experiments, d-prime scores were computed as a measure of performance (Macmillan & Creelman, 2005). Further analyses were all performed on these d-primes.

**Control experiment.** To ensure that participants could recognize a target sound within the RASP sequence, we performed a control experiment, where the recognition of short individual sounds was assessed. A strict exclusion criterion was thus fixed for the analyses of this gating experiment : mean – 1 standard deviation. Data from six participants were below this criterion, and were then discarded. All results presented below are from the 24 remaining participants.

First, we controlled that there was no effect of the different numbers of repetitions and trainings between participants, with a repeated-measures ANOVA on the d-prime scores with the sound duration as within-subject factor and the number of repetitions as categorical factor [repetitions : $F(1,22) = 0.012$, $p = 0.916$, $\eta_p^2 = 0.001$ ; two-way interaction between sound duration and repetitions : $F(6,132) = 0.305$, $p = 0.934$, $\eta_p^2 = 0.014$].

Then, a repeated-measures ANOVA was performed on the d-prime scores with the sound duration as a within-subjects factor. Results are shown in Figure 28. As expected from the literature, performances increased as the sound duration increased [$F(6, 138) = 348.296$, $p < 0.00001$, $\eta_p^2 = 0.938$]. All pairs of sound durations were significantly different between them (orthogonal contrasts) [$t(23) > 3$, $p < 0.002$], except between 2 and 4 ms [$t(23) = 0.785$, $p = 0.441$] and between 64 and 128 ms [$t(23) = 2.018$, $p = 0.055$]. Finally, participants recognized voice and instrument short sounds significantly above chance for durations equal or above 4 ms [at 2 ms : $t(23) = 1.007$, $p = 0.324$ ; at 4 ms : $t(23) = 2.191$, $p < 0.04$ ; at 8 ms : $t(23) = 6.585$, $p < 0.00001$].

All following RASP experiments were conducted with these 24 participants, who showed the ability to recognize short sounds presented individually.

**'Presentation rate' experiment.** Results are represented in Figure 29a. A repeated-measures ANOVA was performed on the d-prime scores with presentation rate and sound target as within-subjects factors. It revealed, surprisingly a

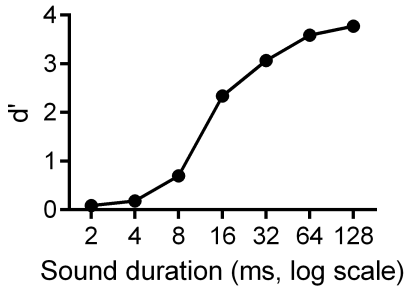


FIGURE 28 – Recognition of individual short sounds (control experiment, 24 participants). Error bars represent the standard errors of the means (too small to be visible in the graph). Performance, as measured by d-prime, increased as the sound duration increased.

better recognition for a voice target embedded in a sequence of instruments than the reverse [$F(1, 7) = 7.901$, $p < 0.03$, $\eta_p^2 = 0.530$]. However, the gain in d-prime for the voice target was, on average, relatively small : $\Delta$d-prime = 0.2. As expected, recognition performance decreased as presentation rate increased [$F(6, 42) = 29.590$, $p < 0.00001$, $\eta_p^2 = 0.809$]. Least-squares linear regressions were used to investigate whether the decreases in performance with presentation rate were linear. Separate regression lines were performed for each sound category. For both categories, the decrease was linear on a log-scale [instrument target : $R^2 = 0.6050$ ; voice target : $R^2 = 0.6524$ ; the slopes were significantly non-zeros : $p < 0.0001$].

To test the maximum rate which allow the recognition of a target within the sequence, performances for a given rate were compared to chance level (d-prime = 0). An instrument target could be recognized within a sequence of voice distractors up to sequences of 21.2 Hz [at 15 Hz : $t(7) > 5$, $p < 0.002$ ; at 21.2 Hz : $t(7) = 2.289$, $p = 0.056$ ; at 30 Hz and 60 Hz, respectively : $t(7) = 0.645$, $p = 0.539$, and $t(7) = -0.100$, $p = 0.924$]. For a voice target within a sequence of instruments distractors, recognition was possible up to 30 Hz [$t(7) > 3$, $p < 0.02$ ; at 60 Hz : $t(7) = 0.651$, $p = 0.536$]. These results confirm the better recognition for a voice target than an instrument one.

Finally, to investigate a potential difference between the slowest presentation rate and the sound presented in isolation, we compared d-prime obtained in the control experiment with d-prime obtained for the 5.3 Hz rate. There was no difference [$t(7) = 2.290$, $p = 0.056$], showing that at least for the slowest rate, no

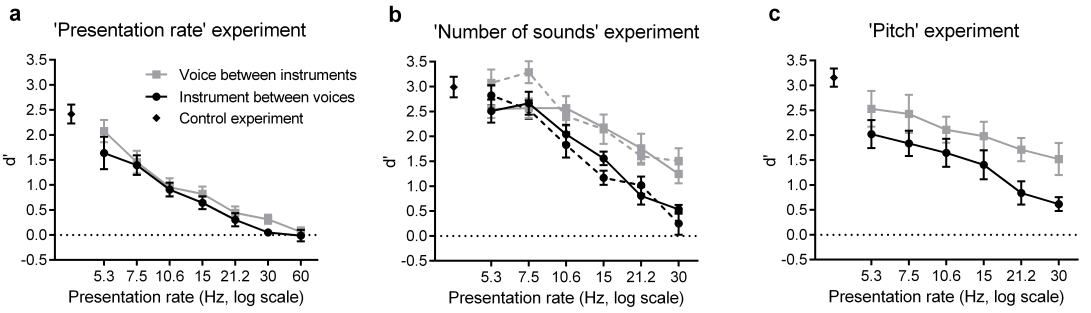


FIGURE 29 – RASP performances : recognition of a short target in a sequence of short distractors presented rapidly. Mean d-prime scores are plotted for each experiment condition as a function of presentation rates. The error bars represent the standard errors of the means. Results from the control experiment, when short sounds were presented in isolation, are represented by a diamond on the left of the curves. For all panels, performance linearly decreased (on a log scale) as the presentation rate increased, and voice were better recognized within instruments than the reverse. Panel a : performance for sequences composed of 16-ms sounds. Panel b : sequences of 32-ms sounds ; solid lines represent the 'fixed number of sounds' condition, dashed lines represent the 'fixed duration' condition. Panel c : sequences of 32-ms sounds were presented. Each line is an average of the 'random-pitch' and the 'control pitch' conditions, there was no difference between these two conditions for all presentation rates.

information was lost, compared to the individual sound recognition.

**'Number of sounds' experiment.** Results are represented on Figure 29b. The repeated-measures ANOVA with presentation rate, sound category, and the 'number of sounds' conditions as within-subjects factors revealed on average a better recognition for a voice target between instrument distractors than the reverse [$F(1, 6) = 10.023$, $p < 0.02$, $\eta_p^2 = 0.626$]. Similarly as in the first experiment, performances decreased with an increased presentation rate [$F(5, 30) = 71.433$, $p < 0.00001$, $\eta_p^2 = 0.923$]. There was no main effect of the sequence type [$F(1, 6) = 0.481$, $p = 0.514$, $\eta_p^2 = 0.074$]. However, the two-way interaction between sequence type and presentation rate was significant [$F(5, 30) = 3.097$, $p < 0.03$, $\eta_p^2 = 0.340$]. Performances seemed to plateau at 7.5 Hz in the case of sequences with a fixed duration, and at 10.6 Hz in the case of sequences with a fixed number of sounds. Indeed, in the case of sequences with a fixed duration, the differences were not significant between the presentation rates 5.3 and 7.5 Hz [$t(6) = 0.314$, $p = 0.765$], whereas in the case of sequences with a fixed number of sounds, the

differences were not significant between the presentation rates : 5.3 and 7.5 Hz, and 10.6 Hz, [respectively : t(6) = -0.765, p = 0.474 ; t(6) = 1.568, p = 0.168].

The two-way interaction between sound category and presentation rate was also significant [F(5, 30) = 3.523, p < 0.02, $\eta_p^2$ = 0.370]. This interaction was probably due to the equivalent performances between voice and instrument targets only for the 5.3 Hz and 7.5 Hz presentation rates [respectively : t(6) = -0.465, p = 0.658 ; t(6) = -1.289, p = 0.245], compared to the other presentation rates [t(6) < -2, p < 0.04]. The two-way interaction between sequence type and sound category was not significant, nor the three-way interaction [respectively : F(1, 6) = 2.655, p = 0.154, $\eta_p^2$ = 0.307 ; F(5, 30) = 2.073, p = 0.097, $\eta_p^2$ = 0.257]. Due to these significant interactions, and to the plateau between the first two presentation rates, the linear regressions were conducted only from the 7.5 Hz to the 30 Hz presentation rates. As in the first experiment, for both categories, performance linearly decreased on a log scale as a function of presentation rate [instrument target : $R^2$ = 0.8071 ; voice target : $R^2$ = 0.5737 ; the slopes were significantly non-zeros : p < 0.0001].

In this experiment, the target was detected above chance for all presentation rates [t(6) > 4, p < 0.005], except in the case of an instrument target in sequences with a fixed duration, at 30 Hz [t(6) = 1.099, p = 0.314].

Finally, there was no difference in performances between the 5.3-Hz condition and the sound presented in isolation [t(6) = 1.44, p = 0.2].

**'Pitch' experiment.** Results for the pitch experiment are plotted on Figure 29c. The repeated-measures ANOVA conducted with presentation rates, sound category, and pitch conditions as within-subjects factors again confirmed the better recognition for a voice target than an instrument [F(1, 8) = 9.118, p < 0.02, $\eta_p^2$ = 0.533], and the decrease in performances with an increased presentation rate [F(5, 40) = 26.378, p < 0.00001, $\eta_p^2$ = 0.767]. Moreover, there was no effect of the pitch condition, suggesting no impact of frequency masking on the RASP performance : the main effect of the pitch condition was not significant, nor the two-way interactions with sound category and with presentation rate [respectively : F(1, 8) = 0.551, p = 0.479, $\eta_p^2$ = 0.065 ; F(1, 8) = 0.002, p = 0.971, $\eta_p^2$ = 0.000 ; F(5, 40) = 1.995, p = 0.100, $\eta_p^2$ = 0.200]. The two-way interaction between sound category

and presentation rate, and the three-way interaction were not significant either [respectively : $F(5, 40) = 1.327$, $p = 0.273$, $\eta_p^2 = 0.142$ ; $F(5, 40) = 0.198$, $p = 0.962$, $\eta_p^2 = 0.024$].

In this experiment as well, the performance decreased linearly on a log scale as presentation increased [instrument target : $R^2 = 0.3554$ ; voice target : $R^2 = 0.1471$ ; the slopes were significantly non-zeros, respectively : $p < 0.0001$ and $p < 0.005$].

The target was recognized above chance for all presentation rates in all conditions [$t(8) > 3$, $p < 0.02$].

Finally, the comparison between the 5.3 Hz condition and the individual sound condition was in this experiment significant [$t(8) = 3.658$, $p < 0.007$], probably linked to a general downwards offset on the RASP performance for this experiment, maybe due to the participant's fatigue.

**Effect of the sound duration.** We assessed the effect of the sound duration on the RASP performance by comparing the performances of the participants of the 'presentation rate' and the 'number of sounds' experiments, for which sounds had respectively a 16-ms and a 32-ms duration. For the 'number of sounds' experiment, we analyzed the data only for the fixed duration condition, for comparison purposes with the 'presentation rate' experiment. For similar reasons, as the 60-Hz presentation rate could not be tested with a 32-ms sound duration, we excluded from the analysis the corresponding data with a 16-ms sound duration in the 'presentation rate' experiment.

We performed a repeated-measures ANOVA on the d-prime scores with the sound categories and the 6 presentation rates as within-subject factors, and the experiment (i.e. the sound duration) as between-subjects factors. Again, a voice target between instrument distractors was better recognized than the reverse [$F(1, 13) = 16.202$, $p < 0.002$, $\eta_p^2 = 0.555$], and recognition decreased with an increased presentation rate [$F(5, 65) = 83.267$, $p < 0.00001$, $\eta_p^2 = 0.865$].

As expected, the recognition was better for 32-ms sounds than for 16-ms sounds when presented in sequences [$F(1, 13) = 37.842$, $p < 0.00004$, $\eta_p^2 = 0.744$]. The two-way interaction between sound category and sound duration was also significant [$F(1, 13) = 5.692$, $p < 0.04$, $\eta_p^2 = 0.305$] (see Figure 30). This effect was due to a

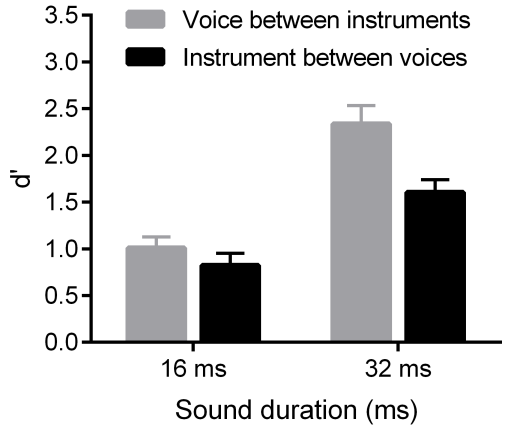


FIGURE 30 – Voices are better recognized in a sequence of instruments than the reverse. The effect is more pronounced for the easiest condition, with 32-ms sounds. Mean d-prime scores (average over presentation rates) are plotted as a function of individual sound duration, for each target type (voice and instrument). The error bars represent the standard errors of the means.

larger voice effect for the 32-ms sequences than for the 16-ms sequences [16-ms : $t(7) = 1.200$, $p = 0.252$; 32-ms : $t(6) = 4.389$, $p < 0.0008$].

The two-way interaction between presentation rate and sound duration was also significant [$F(5, 65) = 2.813$, $p < 0.03$, $\eta_p^2 = 0.178$]. There is no need here to describe this interaction, as the performances have already been described in the previous analyses. The two-way interaction between sound category and presentation rate was not significant, nor the three-way interaction [respectively : $F(5, 65) = 1.677$, $p = 0.153$, $\eta_p^2 = 0.114$ ; $F(5, 65) = 2.283$, $p = 0.057$, $\eta_p^2 = 0.149$].

**Effect of the target position in the sequence.** To investigate potential memory effects, we analyzed an a posteriori effect of the target position in the sequence. We performed a pooled analysis on the blocks in which sequences had a fixed number of sounds ('number of sounds' and 'pitch' experiments). There was no enough data for each target position for the fixed-duration sequences.

Results are represented in Figure 31. A repeated-measures ANOVA was performed on the d-prime scores with the sound categories and the five target positions as within-subject factors, and the experiment as a categorical factor. Performances were equivalent in both experiments [$F(1, 14) = 0.433$, $p = 0.521$, $\eta_p^2 = 0.030$]. As assessed previously, voice targets were better recognized than instrument targets

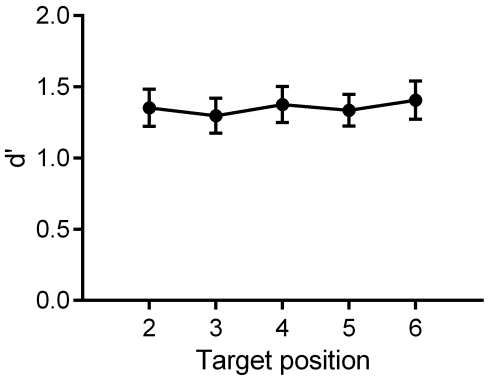


FIGURE 31 – There was no effect of the target position on its recognition in a sequence of distractors. Mean d-prime scores are plotted as a function of target position in a sequence of seven sounds (average over presentation rates and target type). The error bars represent the standard errors of the means.

[$F(1, 14) = 12.139$, $p < 0.004$, $\eta_p^2 = 0.464$]. More importantly here, the effect of the target position was not significant as a main effect, nor in the two-way interactions with experiment and with sound category [respectively : $F(4, 56) = 0.710$, $p = 0.589$, $\eta_p^2 = 0.048$ ; $F(4, 56) = 1.090$, $p = 0.370$, $\eta_p^2 = 0.072$ ; $F(4, 56) = 1.128$, $p = 0.353$, $\eta_p^2 = 0.075$]. The two-way interactions between experiment and sound category, and the three-way interaction were not significant either [respectively : $F(1, 14) = 0.079$, $p = 0.783$, $\eta_p^2 = 0.006$ ; $F(4, 56) = 1.894$, $p = 0.124$, $\eta_p^2 = 0.119$].

### 2.2.4 Discussion

We provide here two new results to the general study of the time course of recognition, and more specifically to the study of recognition of natural sounds like voices and musical instruments. Firstly, we found that auditory recognition was extremely fast, as evidenced by the highest presentation rate up to which participants could still recognized a target sound embedded in a rapid sequence (30 Hz). Secondly, we found a robust voice effect : voice targets in an instrument sequence were better recognized at all rates than instrument targets in a voice sequence.

For all experimental conditions, short sounds could be recognized reasonably well when presented individually (d-prime around 2.5 and 3 for short 16-ms and 32-ms sounds, respectively). When presented in rapid sequences with one target

embedded in a series of distractors, the recognition performance for these same short sounds linearly decreased on a log-scale as the presentation rate increased. Recognition could still be above chance for the fastest rate possible, 30 Hz, for voice sounds, and above or just at chance at 30 Hz for instrument sounds. As shown extensively in vision research with the RSVP paradigm (e.g. Potter, 1976 ; Subramaniam et al., 2000), this new way of measuring performance allows to study the time course of processing in the central nervous system. Perceiving and attending to the stimulus N+1 interferes with the processing of the stimulus N, thus suggesting a time window for the processing of the stimulus N. From our data, we can establish two limits of this time window : an upper limit which defines optimal recognition with no loss of information, and a lower limit, for which the recognition was possible, above chance, which helps to define the shortest time for a first 'read-out' of the information by the auditory system. Here, the upper limit was found to be around 200 ms, corresponding to the slowest presentation rate, 5.3 Hz. This is the necessary time window for which performance was optimal, with no difference in the performance for sounds presented in sequences and for sounds presented individually (control experiment). This upper limit is in accordance with the time course of recognition identified with other paradigms (Massaro, 1972a). Results reported in Massaro's study were derived from stimuli with considerably less acoustical variability (tones and very few examples of syllables) than in the present study. He predicted much longer processing time for natural stimuli ; interestingly, we found a similar time window as in his report, but with a much larger, diverse, but controlled set of natural stimuli. The lower limit was found at 33 ms, and corresponds to the 30-Hz presentation rate. This short time window is also in accordance with the literature, from different paradigms and research fields. Massaro (1972a) proposed the concept of a pre-perceptual store, from which a 'read-out' could be performed, in order to recognize a stimulus. He showed that the read-out started as the stimulus was still ongoing, as early as 30 ms, with similar levels of performance achieved (d-prime around 1). With the view that our natural sounds were decomposed and analyzed in terms of features (see Isnard et al., 2016), this result argue in favor of a parallel processing of features (see also Woods & Alain, 1993 ; Duncan et al., 1997). Interestingly, this processing time is almost as short as the time took by the information to

flow from the periphery to the brain, which is around 15 – 30 ms (see Helmholtz, 1895 ; Posner, 1978 ; Liegeois-Chauvel et al., 1994). It can also be compared, within the same time range, to the 70 ms found with an ERP paradigm, for the first responses to brain discrimination of sounds of objects (Murray et al., 2006). Finally, it is also worth noting that these 30 ms corresponds also to the lowest limit of pitch perception (Pressnitzer et al., 2001).

The new RASP paradigm was proposed, as its visual analogue, to study the time course of auditory recognition, and helps, by bringing strong constraints on the task, to reveal the important auditory features used to recognize a diverse set of natural sounds. The robustness of the results throughout all the experimental conditions permits to rule out all the other plausible alternative explanations for the fast recognition of natural sounds. Firstly, forward frequency masking could have played a role, by limiting the detectability of each sound in a sequence (Moore, 2012). This hypothesis was ruled out with the 'pitch' experiment, as a frequency masking hypothesis would have predicted better performance for the random-pitch condition, with the harmonics of successive sounds better separated in frequency than for the constant-pitch condition. Secondly, the decrease in performance as the presentation rate increase could have been due to the increase of information : with a fixed-duration sequence, as the rate increase, so does the number of sounds. The 'number of sounds' experiment confirmed that this was not the case, and that here, by increasing the presentation rate, we decreased effectively the available time for the auditory system to analyze the information (see Massaro, 1972a). Finally, the absence of effect of the target position on the sequence on the performance shows that the RASP effect was not even partly due to memory effect, with better recall of first and last items in a sequence (e.g. Murdock, 1962).

The robust results obtained in the present study confirm what was obtained in the previous RASP experiments (Suied et al., 2013a), and extend it with a more balanced set of sounds (same number of voices and instruments), different targets, and a large group of participants, with controlled audiometry. A new, solid, and unexpected result also emerged from our current study : a voice advantage. There was no trivial acoustical explanation for this voice advantage. The voice sounds used as a target in the voice blocks were the voice distractors in the instruments

blocks, and vice versa. The acoustical variability of the sound set was large for both categories, and listeners were left only with timbre cues to perform the task (similar loudness cues, and random pitches within an octave range, with no possibility to use the pitch as a cue). This effect, however, is not a surprise in the realm of a very large number of neuroimaging studies, providing evidence for a specific voice treatment in the auditory cortex (e.g. Belin et al., 2000, 2002, 2004 ; Moerel et al., 2012). There is, however, few behavioral studies studying voice per se (and not speech). The few behavioral evidences collected show a similar trend as in this study : a behavioral voice advantage, which might reflect its specific neural treatment. Comparing voice and instrument stimuli, Agus et al. (2012) showed faster response times for voice stimuli. The faster recognition of voice could be explained by a smaller set of timbre features necessary to recognize voices compared to instruments, or at least more rapidly available. With the same corpus of sounds, but this time presenting to the participants only short snippets of sounds (like in the present control experiment), Suied et al. (2014) showed that the minimal duration necessary to recognize a voice sound was lower (4 ms) than for an instrument sound (8 ms). It is possible that the higher information rate present in the short voice sounds contributes to the asymmetry observed in our study. The dependence of time processing with the duration of the sounds in the sequences reflects this accumulation of timbre features in temporal perceptual windows.

However, the apparent better underlying auditory coding of voice sounds could have facilitated the recognition of a target instrument, by a simple auditory contrast with the voice distractors. This was clearly not the case. Interestingly, Cusack & Carlyon (2003) obtained a similar asymmetry in the recognition of basic features in auditory sequences with, in particular, a better recognition of frequency-modulated targets between pure tone distractors than the reverse. They suggested that frequency modulation is coded as an extra feature which consequently drives the recognition of the target. Here, it could be argued that the voice pops out of the instrument distractors, because voice is coded as a specific feature. The fact that the voice effect did not depend on the number of distractors (comparison, for a given presentation rate, of the 'fixed number of sounds' condition with the 'fixed sequence duration' condition) is an additional argument

in favor of a pop-out effect. The voice features are probably located in conjoint temporal and spectral modulations, which can characterize natural sounds like vocalizations or environmental sounds (Singh & Theunissen, 2003). Using natural sounds, our results revealed an outstanding temporal processing for the voice timbre, and demonstrate that, whatever these voice features are, they are detected in a very efficient way by the auditory system.

These results were obtained by means of a new paradigm in audition, RASP, which, together with complementary behavioral (e.g. reaction times) or electrophysiological techniques (e.g. auditory evoked potential), should be a powerful technique to study mental chronometry of the mind (see Posner, 1978), and more specifically here the time course of auditory recognition.

## 2.3 Réponses rapides à des voix dans des séquences RASP

### 2.3.1 Introduction

Dans l'étude précédente, le paradigme RASP a permis de révéler un avantage fort et significatif pour la reconnaissance d'une cible voix dans une séquence de distracteurs instruments présentée sériellement et rapidement, en comparaison à une cible instrument dans une séquence de distracteurs voix. La présente étude a pour objectif de déterminer si cet avantage pour la voix peut être corroboré à l'aide de temps de réaction (TRs), afin d'évaluer la confiance dans la réalisation de cette tâche en fonction des catégories sonores (e.g. Emmerich et al., 1972).

Il s'agit donc de mesurer avec quelle rapidité des participants répondent lorsqu'ils ont reconnu une cible isolée ou dans une séquence de distracteurs, en fonction de la catégorie des sons. Dans une expérience contrôle, nous avons contrôlé la rapidité de reconnaissance de sons courts présentés isolément. Dans l'expérience principale, nous avons évalué la rapidité de reconnaissance d'une cible sonore courte dans une séquence de courts distracteurs présentés séquentiellement, suivant le paradigme RASP.

### 2.3.2 Matériel et méthode

**Participants.** Quatorze participants, qui avaient également participé à l'étude RASP précédente (sans en avoir été exclu, d'après les critères d'exclusion mentionnés précédemment) dans les 7 mois avant cette nouvelle étude, ont été recrutés (8 femmes ; âge moyen = $24.3 \pm 3.1$). L'audition de tous les participants a été contrôlée en effectuant un audiogramme avec un audiomètre Echodia Elios, de façon à vérifier que les pertes auditives ne dépassaient pas 20 dB HL entre 0.125 et 4 kHz. Le comité d'évaluation éthique de l'INSERM (*institutional review board*) a préalablement approuvé cette étude spécifique (avis n°15-211), et tous les participants ont signé un consentement de participation. Ils ont été compensés financièrement pour leur participation à cette étude, qui durait environ 1 heure.

**Stimuli et équipement.** Comme dans l'étude précédente, 2 catégories sonores ont été utilisées : voix et instruments, avec 4 sources sonores originales dans chaque catégorie. Les sources sonores de voix sont 2 voix chantées d'hommes sur les

voyelles /a/ et /i/, et deux voix chantées de femmes sur les voyelles /e/ et /o/. Les sources sonores d'instruments sont : un basson, une clarinette, un piano, et un saxophone. La hauteur de chaque son (présenté seul ou dans une séquence, voir ci-dessous) a été choisie aléatoirement dans l'octave de A3 à G#4. Les sons originaux ont été tronqués sur une durée de 32 ms avec une fenêtre de Hanning, et avec, pour chaque extrait, le début de la troncation choisi aléatoirement entre 0 et 100 ms à partir du début du son.

Dans l'expérience contrôle, les stimuli sont des sons individuels. Pour l'expérience principale, nous avons créé des séquences de sons courts présentés rapidement. Les séquences ont un nombre fixe de 7 sons par séquence. La durée des séquences varie en fonction de 6 taux de présentation : 5.3, 7.5, 10.6, 15, 21.2, et 30 Hz. Toutes les séquences sonores incluent des distracteurs de l'une des catégories sonores (voix ou instruments), et peuvent inclure une cible de l'autre catégorie placée à une position aléatoire dans la séquence, excepté aux première et dernière positions.

Les participants effectuent le test individuellement dans une cabine acoustique *Industrial Acoustics* (IAC). Les stimuli sont présentés à l'aide d'un convertisseur numérique-analogique TDT RM1 Mobile Processor (Tucker-Davis Technologies) à une résolution de 16 bits et une fréquence d'échantillonnage de 48828 Hz, après une conversion de 44100 à 48828 Hz dans Matlab. Ils sont présentés à travers un casque Sennheiser HD 650 à 70 dBA.

**Procédure.** Dans les 2 expériences, les participants effectuent une tâche *Go/No-Go*. La cible est présentée dans 80% des essais. A chaque essai, les participants doivent garder l'index de leur main dominante sur le bouton réponse, écoutent un stimulus et doivent appuyer sur le bouton réponse le plus rapidement et précisément possible s'ils reconnaissent la cible (même si la séquence n'est pas terminée dans le cas de l'expérience principale). Sinon, ils doivent attendre le prochain stimulus. Le délai entre les stimuli est de 1.5 s pour les sons individuels (expérience contrôle), et de 2.2 s pour les séquences sonores (expérience principale). Un retour visuel indique si la réponse donnée était correcte ou fausse.

Dans l'expérience contrôle, pour chaque catégorie cible (voix ou instruments), les participants sont d'abord familiarisés avec la tâche avec un court entraînement

de 40 essais. La première moitié des essais comporte les sons originaux d'une durée de 250 ms, les essais suivants sont des sons tronqués de 32 ms. Puis, les participants effectuent 60 essais aléatoires (48 répétitions par catégorie).

Similairement, dans l'expérience principale, pour chaque catégorie cible successivement (voix ou instruments), les participants effectuent d'abord un court entraînement de 40 essais. La première moitié des essais est présentée avec un taux de présentation de 5.3 Hz, les essais suivants avec un taux de présentation de 15 Hz. Puis, les participants effectuent 3 blocs de 120 essais aléatoires (48 répétitions par catégorie et par taux de présentation). Les participants effectuent les 2 expériences avec l'ordre de passage de chaque cible contrebalancé entre les participants.

### 2.3.3 Résultats

**Expérience contrôle : reconnaissance de sons courts individuels.** Nous avons calculé les scores $d'$ comme mesure de la performance (Macmillan & Creelman, 2005). La Figure 32 représente les résultats de l'expérience contrôle en termes de scores $d'$ et de TRs des détections correctes. Les réponses égales ou inférieures à 100 ms étaient considérées comme des erreurs. Pour analyser les TRs, nous avons vérifié la log-normalité des distributions pour chaque participant et pour chaque catégorie sonore avec des tests de Kolmogorov-Smirnov. La différence avec une distribution log-normale théorique n'était pas significative dans 26 des 28 distributions ($p > 0.05$). Puis, la moyenne de chaque distribution a été calculée, après les avoir transformées en distributions normales avec la fonction logarithme. Finalement, la moyenne a été reconvertie en millisecondes avec la fonction exponentielle.

Les performances pour reconnaître rapidement un son court de voix ou d'instrument présenté isolément sont équivalentes (test t pour échantillons appariés : $t(13) = -1.856$, $p = 0.086$), bien qu'une tendance suggère une meilleure reconnaissance des voix en termes de scores $d'$ moyens ($\Delta d' = 0.41$). En termes de TRs cependant, les réponses sont significativement plus rapides pour des cibles voix que pour des cibles instruments (test t pour échantillons appariés : $t(13) = 3.366$, $p < 0.006$).

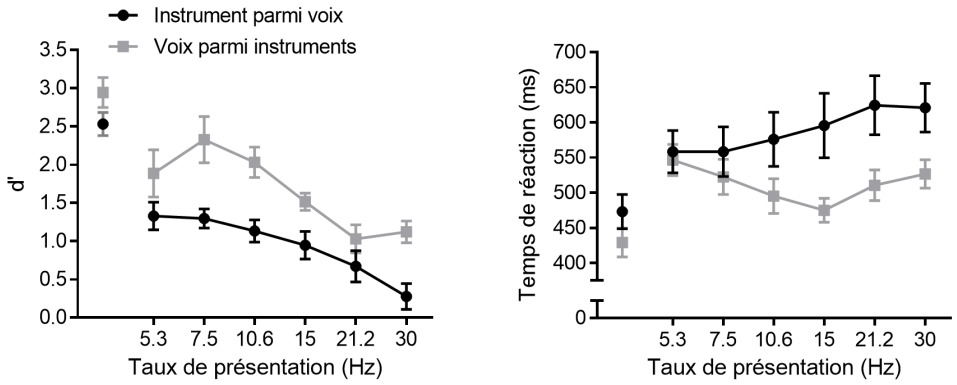


FIGURE 32 – Reconnaissance d'une cible sonore présentée isolément, ou dans une séquence de distracteurs en fonction du taux de présentation. Scores d' moyens (à gauche) et TRs (en ms ; à droite) pour des cibles de voix et d'instruments. Les résultats pour des cibles isolées sont représentés par des points isolés à gauche des courbes. Les barres d'erreurs représentent les erreurs types.

**Expérience principale : reconnaissance d'une cible dans une séquence sonore.**

**En fonction du taux de présentation.** La Figure 32 représente les résultats de l'expérience principale en termes de scores $d'$ et de TRs des détections correctes, en fonction du taux de présentation. Un participant a été exclu des analyses à cause de ses scores $d'$ qui étaient proches ou inférieurs à 0, indiquant qu'il effectuait la tâche au hasard (Macmillan & Creelman, 2005). Comme dans l'expérience contrôle, les réponses égales ou inférieures à 100 ms à partir du début de la cible dans la séquence sont considérées comme des erreurs. Nous avons aussi vérifié la log-normalité des distributions pour chaque participant et pour chaque condition (catégorie et taux de présentation) avec des tests de Kolmogorov-Smirnov. La différence avec une distribution log-normale théorique n'était pas significative dans 153 des 156 distributions ($p > 0.05$). Ensuite, les réponses dépassant 2 s ont été supprimées (0.6% des détections correctes). Enfin, les TRs moyens de chaque distribution ont été calculés comme dans l'expérience contrôle.

Deux ANOVAs ont été effectuées, avec pour variable dépendante respectivement le score $d'$ et le TR, et avec pour facteurs inter-participants : catégorie sonore (voix ou instruments) et taux de présentation.

Comme dans l'étude précédente, l'analyse de la performance a révélé une meilleure reconnaissance d'une cible voix parmi des distracteurs instruments que l'inverse ($F(1, 12) = 14.314$, $p < 0.003$, $\eta_p^2 = 0.544$). Par ailleurs, les performances diminuent significativement avec le taux de présentation croissant ($F(5, 60) = 21.666$, $p < 0.00001$, $\eta_p^2 = 0.644$). Cet effet a été analysé avec un test post-hoc de Tukey-HSD. Les performances ne sont pas différentes entre 5.3 et 7.5, 10.6, 15 Hz (respectivement : $p = 0.666$, $p = 1.000$, $p = 0.076$), ni entre 7.5 et 10.6 Hz, 10.6 et 15 Hz, 15 et 21.2 Hz, 21.2 et 30 Hz (respectivement : $p = 0.548$, $p = 0.115$, $p = 0.073$, $p = 0.879$). Les différences sont significatives dans tous les autres cas ($p < 0.004$). L'interaction double entre catégorie et taux de présentation n'est pas significative ($F(5, 60) = 1.036$, $p = 0.405$, $\eta_p^2 = 0.080$).

Finalement, nous avons effectué des tests un-échantillon sur les scores $d'$ comparés à 0 (niveau de chance), pour tous les taux de présentation et catégories sonores. La cible est détectée au-dessus de la chance pour tous les taux de présentation, qu'il s'agisse d'une voix ou d'un instrument ($t(12) > 3$, $p < 0.007$), excepté dans le cas d'une cible instrument parmi des distracteurs voix à 30 Hz ($t(12) = 1.635$, $p = 0.128$).

En termes de TRs, similairement aux résultats obtenus pour des sons isolés, les réponses pour une cible voix sont en moyenne plus rapides que pour une cible instrument ($F(1, 12) = 8.064$, $p < 0.02$, $\eta_p^2 = 0.402$). Il n'y a pas d'effet principal du taux de présentation ($F(5, 60) = 1.444$, $p = 0.222$, $\eta_p^2 = 0.107$). Cependant, l'interaction double est significative ($F(5, 60) = 2.396$, $p < 0.05$, $\eta_p^2 = 0.166$), et a été analysée plus en détail avec un test post-hoc de Tukey-HSD. Ce test révèle des TRs plus rapides pour une cible voix comparée à une cible instrument à : 7.5 vs. 21.2 et 30 Hz (respectivement : $p < 0.03$, $p < 0.04$), 10.6 vs. 15, 21.2, et 30 Hz (respectivement : $p < 0.04$, $p < 0.002$, $p < 0.003$), 15 vs. 10.6, 15, 21.2, et 30 Hz (respectivement : $p < 0.03$, $p < 0.004$, $p < 0.0003$, $p < 0.0003$), 21.2 vs. 21.2 et 30 Hz (respectivement : $p < 0.008$, $p < 0.02$), 30 vs. 21.2 Hz ($p < 0.05$). Les différences ne sont pas significatives dans tous les autres cas ($p > 0.05$).

**En fonction de la position de la cible.** La Figure 33 présente les résultats de l'expérience principale en termes de scores $d'$, cette fois en fonction de la position de la cible dans la séquence. Une ANOVA a été effectuée avec pour

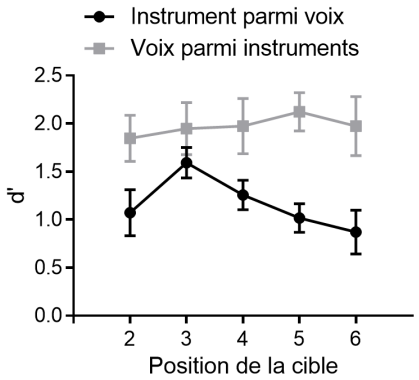


FIGURE 33 – Reconnaissance d'une cible dans une séquence de distracteurs en fonction de la position de la cible dans la séquence. Scores d'moyens pour des cibles de voix et d'instruments. Les barres d'erreurs représentent les erreurs types.

variable dépendante le score $d'$, et avec pour facteurs inter-participants : catégorie sonore (voix ou instruments) et position de la cible. Comme dans les analyses précédentes des performances, la reconnaissance d'une cible voix parmi des distracteurs instruments est meilleure que l'inverse ($F(1, 12) = 14.586$, $p < 0.003$, $\eta_p^2 = 0.549$). L'effet principal de la position de la cible n'est pas significatif, ni l'interaction double avec la catégorie (respectivement : $F(4, 48) = 0.986$, $p = 0.424$, $\eta_p^2 = 0.076$ ; $F(4, 48) = 1.624$, $p = 0.184$, $\eta_p^2 = 0.119$).

### 2.3.4 Discussion

Les performances de reconnaissance d'un son court de voix ou d'instrument sont équivalentes, avec cependant une tendance vers une meilleure reconnaissance des voix. En revanche, les voix sont reconnues significativement plus rapidement que les instruments, ce qui avait déjà été observé avec différentes tâches de reconnaissance de voix et d'instruments impliquant la mesure de TRs (Agus et al., 2012). Cette efficacité dans la reconnaissance des voix semblent refléter des traitements issus de mécanismes cérébraux spécifiques (e.g. Belin et al., 2000).

Dans le cas de séquences RASP, les résultats de cette étude en termes de performances confirment ceux de notre étude précédente : une cible voix parmi des distracteurs instruments est mieux reconnue que si les deux catégories sonores sont inversées, et à des taux de présentation plus élevés (30 vs. 21.2 Hz). Cu-

sack & Carlyon (2003) ont aussi observé des asymétries dans la reconnaissance d'une cible parmi des distracteurs, qui semblaient s'expliquer par la présence de caractéristiques uniques dans la cible comparée aux distracteurs (e.g. modulation fréquentielle, durée). Selon les auteurs, ces caractéristiques doivent générer une plus grande activation des structures cérébrales lorsqu'elles sont présentes comparée à lorsqu'elles ne sont pas présentes. A nouveau, un mécanisme similaire pourrait entrer en jeu dans le cas de sons de voix (Belin et al., 2000). En outre, cette meilleure reconnaissance pour les voix est aussi plus rapide pour des taux de présentation élevés (15 et 21.2 Hz), ce qui semble induire une plus grande confiance dans la prise de décision relative à la reconnaissance d'une cible voix (Emmerich et al., 1972).

Enfin, les performances de reconnaissance ne dépendent pas de la position de la cible dans la séquence, pour les deux catégories sonores, confirmant les résultats obtenus dans notre étude précédente. L'influence de processus liés à la mémoire (e.g. Murdock, 1962) sur la reconnaissance d'une cible dans une séquence RASP serait donc limitée.

## 2.4 Compléments d'analyses (1/2) : quels indices permettent de reconnaître des sons individuels très courts ?

### 2.4.1 Contrôle du CGS et du HNR

L'expérience contrôle de l'étude du temps de traitement auditif visait à établir la durée minimale de présentation nécessaire pour la reconnaissance de stimuli auditifs courts et présentés individuellement. Comme détaillé précédemment, chaque catégorie (voix ou instruments) était composée de quatre sources sonores originales.

Pour la reconnaissance des stimuli auditifs courts, les participants ne pouvaient se servir que du timbre des sons. En effet, la hauteur était choisie aléatoirement sur une octave de A3 à G#4, l'intensité sonore était égalisée en niveau RMS, et les sons étaient présentés aléatoirement avec une durée de 2 à 128 ms. Cependant, en plus de ces égalisations en hauteur, intensité sonore, et durée, nous avons aussi estimé le CGS et le HNR de tous les sons originaux, à l'aide du logiciel Praat (Boersma & Weenink, 2015). Le CGS est un corrélat acoustique trouvé de façon concordante dans plusieurs études sur le timbre (e.g. Grey, 1977 ; Krimphoff et al., 1994 ; McAdams et al., 1995 ; Caclin et al., 2005 ; Elliott et al., 2013), tandis que l'influence des composantes périodiques et apériodiques du signal quantifiée par le HNR a notamment été mis en évidence dans des études d'imagerie sur l'encodage de sons naturels par le cortex auditif (Staeren et al., 2009 ; Giordano et al., 2013).

La moyenne (± écart-type) des valeurs de CGS pour les voix était de 785.8 ± 244.2 Hz, et de 754.1 ± 337.7 Hz pour les instruments. La moyenne du HNR pour les voix était de 17.1 ± 4.4 dB et de 20.0 ± 3.9 dB pour les instruments. Nous avons effectué un test t pour deux échantillons indépendants sur les valeurs de CGS et de HNR, entre les deux catégories sonores. La différence n'était pas significative entre les voix et les instruments en termes de CGS ($t(94) = -0.527$, $p = 0.599$). La différence était significative en termes de HNR ($t(94) = 3.418$, $p < 0.001$). Cependant, cette différence entre les deux moyennes était relativement faible ($\Delta$HNR = 2.9 dB) en comparaison à d'autres catégories de sons naturels (cf. Isnard et al., 2016).

### 2.4.2 Modèle de distance auditive

Comme expliqué dans les paragraphes précédents, les stimuli étaient égalisés dans un grand nombre de dimensions perceptives : hauteur, intensité sonore, durée, CGS, tandis que les valeurs de HNR étaient proches entre les deux catégories. Pour tenter de comprendre comment il pouvait encore être possible de reconnaître des sons de quelques millisecondes, nous avons utilisé le modèle de distance auditive entre catégories présenté dans notre précédente étude (Isnard et al., 2016), que nous avons proposé pour expliquer des résultats perceptifs dans une tâche de catégorisation avec des sons naturels qui étaient simplifiés dans le domaine spectro-temporel. Ce modèle pourrait nous aider à détecter quels indices spectro-temporels sont encore présents dans les sons courts utilisés dans la présente étude, et dans quelle mesure ils peuvent nous permettre d'évaluer l'appartenance de ces sons à leur catégorie sonore.

Tout d'abord, pour chaque durée sonore, nous avons créé un ensemble de sons de simulation comprenant 96 sons courts (2 catégories $\times$ 4 sources sonores $\times$ 12 hauteurs), avec la même méthode que celle employée avec les participants. Puis, nous avons calculé les distances entre les STEPs (Moore, 2003) de chaque son grâce à un algorithme DTW, avant de calculer les distances auditives entre les catégories voix et instruments pour chaque ensemble de sons de simulation (i.e. pour chaque durée sonore).

Les résultats ont révélé que les distances auditives moyennes entre les deux catégories sont très corrélées avec la durée sonore (corrélation linéaire de Pearson : $r = 0.99$, $p < 0.00001$). Il n'est donc pas surprenant d'observer ensuite que les performances sont aussi très corrélées avec les distances auditives, avec cette fois une corrélation non-paramétrique de Spearman ($rs = 0.96$, $p < 0.003$ ; Figure 34).

Cette corrélation indique la quasi-monotonie (excepté pour les durées 8 et 16 ms avec une très faible baisse de distances auditives autour de 0.001) de l'augmentation des performances avec l'augmentation des distances auditives, ce qui était déjà observé en fonction de la durée des sons. Les distances auditives entre catégories sonores ont permis de rendre compte de la possibilité de capturer des indices spectro-temporels lorsqu'ils sont toujours présents dans les sons courts (indépendamment pour chaque durée). Donc la corrélation observée sur la Figure 34 n'est pas une simple transposition des performances affichées en fonction de la

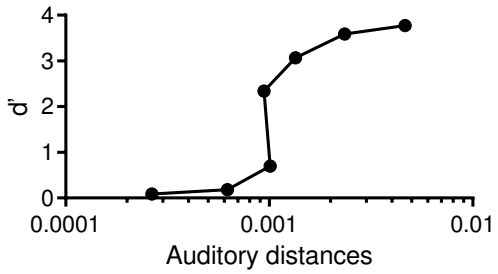


FIGURE 34 – Résultats de l'expérience contrôle exprimés en fonction des distances auditives entre les catégories voix et instruments. Scores d' moyens en fonction des distances auditives, avec les barres d'erreur représentation les écart-types des moyennes (trop petites pour être visibles sur la figure). L'utilisation d'une échelle logarithmique permet de compenser la décroissance en distances auditives avec la durée sonore, laquelle suivait une échelle logarithmique.

durée sonore, mais bien un réel effet perceptif dû à la présence ou à l'absence d'indices spectro-temporels – ou spectraux dans le cas des durées très courtes (Suied et al., 2014). La nature de ces indices reste cependant encore à déterminer pour expliquer plus précisément leur influence sur les performances de reconnaissance auditive.

### 2.4.3 Bilan sur les indices permettant la reconnaissance de sons courts isolés

Dans cette expérience contrôle, nous avons vérifié que des participants peuvent reconnaître des sons très courts de voix et d'instruments présentés seuls, en accord avec des résultats de la littérature (Gray, 1942 ; Robinson & Patterson, 1995b,a ; Suied et al., 2013a, 2014). Tous les sons étaient égalisés en hauteur, niveau sonore, et durée, donc leur reconnaissance se faisait sur la base de leur timbre uniquement. De plus, les valeurs de CGS étaient aussi égalisées et celles de HNR étaient proches entre les deux catégories. Malgré le fait que tous ces indices aient été égalisés, la reconnaissance de sons courts de voix et d'instruments était encore possible au-dessus de la chance pour des durées de présentation jusqu'à 4 ms, et avec de très bonnes performances dès 16 ms.

La corrélation entre les performances et la durée sonore indique seulement des limites perceptives dans le domaine temporel. En revanche, les distances auditives

entre catégories sonores peuvent expliquer les performances en termes d'indices spectraux et spectro-temporels, et pourraient être associées à une dimension de timbre comparable à d'autres corrélats spectro-temporels de dimensions perceptives trouvés dans des études sur le timbre (e.g. flux spectral ; Krumhansl, 1989 ; McAdams et al., 1995 ; Kendall et al., 1999).

## 2.5 Compléments d'analyses (2/2) : variabilité des distances auditives entre les sons des séquences RASP

Les distances auditives obtenues pour l'expérience contrôle avec des sons courts isolés sont détaillées ici pour les voix et pour les instruments séparément. Il s'avère que les distances auditives pour les voix ne sont pas significativement différentes de celles pour les instruments, pour la durée de 16 ms : 0.001 $\pm$ 0.0009 pour les voix et 0.0009 $\pm$ 0.0009 pour les instruments (test t pour 2 échantillons indépendants : $t(94) = -0.240$, $p = 0.811$). Mais les distances auditives pour les voix sont significativement plus grandes que celles pour les instruments pour la durée sonore de 32 ms : 0.002 $\pm$ 0.001 pour les voix et 0.0009 $\pm$ 0.001 pour les instruments (test t pour 2 échantillons indépendants : $t(94) = -3.626$, $p < 0.0005$). En d'autres termes, pour la durée sonore de 32 ms, les voix sont à la fois plus loin des instruments et plus proches entre elles, que les instruments comparés aux voix.

Les distances auditives indiquent une plus grande discrimination des voix comparées aux instruments pour les sons de 32 ms, tandis que les instruments sont plus variables entre eux. On peut donc supposer que des instruments peuvent être davantage confondus avec des voix, que des voix ne peuvent l'être avec des instruments. Pour les sons de 16 ms, l'effet n'est pas significatif en termes de distance auditive, mais on constate aussi que la taille d'effet de la catégorie observée sur les résultats perceptifs est aussi réduite pour cette durée sonore, comparée à la durée sonore de 32 ms. Notre modèle de distance auditive donne donc certains éléments pour comprendre quels indices spectro-temporels restent dans le cas de sons courts et très courts, et comment leur variabilité entre catégories permet de distinguer les cibles d'une catégorie des distracteurs d'une autre catégorie dans une séquence de sons courts présentés rapidement.

# Discussion générale

Deux aspects de l'efficacité de la reconnaissance auditive ont été traités dans cette thèse : la quantité d'information et la rapidité du traitement de reconnaissance. Pour évaluer l'efficacité auditive suivant chacun de ces deux aspects, nous avons mené des expériences visant à déterminer les limites perceptives en termes de quantité minimale d'information et de temps de traitement minimal nécessaires pour reconnaître un son.

### *Reconnaissance d'indices auditifs parcimonieux*

Tout d'abord, nous nous sommes inscrits dans un cadre d'étude qui prend à contre-pied la notion d'espace de timbre multidimensionnel (Pressnitzer et al., 2013). Ce cadre d'étude fait écho aux résultats sur le codage parcimonieux des représentations cérébrales auditives, mis en évidence par des expériences en neurophysiologie (Smith & Lewicki, 2006 ; Hromadka & Zador, 2009). A travers des expériences comportementales, nous avons montré qu'en sélectionnant parcimonieusement l'information, seuls quelques pics d'énergie sélectionnés sur un spectrogramme auditif suffisent à reconnaître la catégorie sonore du son original. Autrement dit, la reconnaissance des sons semble s'appuyer sur des combinaisons d'indices acoustiques contenus dans les sons naturels, transformés et codés par les différents traitements auditifs successifs. Dans nos expériences avec des sons très simplifiés, nous sommes parvenus à sélectionner les indices qui semblent prendre part à ce codage parcimonieux, étant donné qu'une extrêmement faible quantité d'information auditive suffit à leur reconnaissance (Suied et al., 2013b ; Isnard et al., 2016).

### *Rapidité de reconnaissance auditive*

Dans une deuxième série d'expériences, nous avons d'abord montré que la reconnaissance auditive est possible même pour des extraits sonores très courts. Les représentations auditives issues d'un très petit extrait sonore sont nécessairement très succinctes mais encore suffisantes pour sa reconnaissance (Occelli et al., 2015).

A la suite de ce résultat préliminaire, le paradigme RASP proposé par Suied et al. (2013a) nous a permis de tester le seuil de rapidité de la reconnaissance auditive de sons naturels courts. Selon nos estimations obtenues sur des données expérimentales, des sons courts semblent être analysés dans des fenêtres temporelles de l'ordre de 25-50 ms, au-delà de la durée du signal sonore jusqu'à pouvoir reconnaître effectivement le son. Cette rapidité de reconnaissance est de l'ordre de celle trouvée par Suied et al. (2013a), et est aussi plus rapide que les estimations d'études comportementales utilisant des paradigmes similaires de masquage de reconnaissance (cf. Massaro, 1972a) ou obtenues à partir de données électrophysiologiques (e.g. Murray et al., 2006 ; Charest et al., 2009).

***Troubles du traitement auditif rapide***

Un aspect complémentaire de ces résultats serait à approfondir, il s'agit du cas où le traitement n'est pas rapide mais au contraire anormalement lent. Suied et al. (2013a) ont exclu deux participants de leurs analyses de leur expérience RASP à cause de scores trop faibles, bien que ces mêmes participants avaient effectué correctement la tâche de reconnaissance de sons courts présentés isolément. Dans notre étude, nous n'avons pas exclu de participants car ils avaient déjà été sélectionnés avec des critères stricts, d'abord d'après leur audiométrie clinique, ensuite en contrôlant strictement leurs performances pour la reconnaissance de sons courts présentés isolément. Toutefois, nous avons observé une certaine variabilité dans nos résultats RASP. Il serait intéressant de tester un plus grand nombre de participants avec le paradigme RASP pour étudier spécifiquement cette variabilité, qui pourrait suggérer des pertes auditives cachées qui ne seraient pas révélées par l'audiométrie clinique.

Ce type de déficit du traitement auditif rapide a déjà pu être observé chez des participants avec des troubles dys (e.g. dyslexie, dysphasie ; Hari & Kiesila, 1996 ; Benasich et al., 2002 ; Tallal, 2004). Un trouble dys est diagnostiqué lorsqu'un participant, avec un développement intellectuel normal par ailleurs (i.e. sans retard mental, ni trouble de l'audition ou de la vision, ni schizophrénie, ni autisme), échoue à réaliser correctement des tâches liées au langage (Benasich et al., 2002 ; Tallal, 2004). Un traitement temporel efficace des propriétés acoustiques basiques

serait nécessaire pour une acquisition normale du langage, étant donné que la parole présente un taux élevé d'évènement acoustiques rapides changeant rapidement et que sa compréhension passe par la reconnaissance de courtes transitions formantiques (environ 40 ms ; Schwartz & Tallal, 1980 ; Johnsrude et al., 1997 ; Tallal, 2004). Ainsi, les difficultés de traitement de stimuli auditifs brefs et présentés rapidement en succession pourraient permettre de prédire significativement le développement ultérieur du langage (Benasich et al., 2002).

Un grand nombre d'études – la majorité conduites par Tallal (e.g. Tallal, 1975, 1980, 1984, 2004 ; Tallal & Piercy, 1973 ; Schwartz & Tallal, 1980 ; Jernigan et al., 1991 ; Frenkel et al., 2000 ; Benasich et al., 2002) – ont eu pour objectif de mettre en évidence un lien de causalité entre un déficit dans le traitement de stimuli acoustiques présentés rapidement et la déficience phonologique chez les participants dys, plutôt que l'implication de facteurs linguistiques plus complexes [17]. Par exemple, Tallal & Piercy (1973) ont observé que les performances de reconnaissance de deux tons complexes présentés successivement commencent à chuter dès un ISI inférieur à 305 ms, ce qui n'est pas le cas avec des participants normaux. Avec des séquences de 3 ou 4 de ces tons, très peu de participants sont parvenus à réaliser correctement la tâche. Des déficits similaires ont été observés chez des souris sur lesquelles on avait pratiqué des ectopies du cortex frontal ou pariétal (Frenkel et al., 2000), donnant des indications sur les mécanismes cellulaires possiblement impliqués dans les troubles dys.

L'utilisation du paradigme RASP dans le contexte de la dyslexie pourrait systématiser l'étude des troubles de traitement auditif rapide en général, avec la possibilité de tester un grand nombre de paramètres acoustiques et montrer peut-être aussi l'existence d'asymétries de traitement auditif rapide similaires ou non à celles qu'on a pu en observer dans notre étude. En effet, le paradigme RASP a l'avantage d'être polyvalent. En particulier, il serait intéressant de le tester avec des sons simplifiés telles que des esquisses auditives (Isnard et al., 2016), afin

17. Malgré le nombre de ces études concordantes, il est important de mentionner que l'hypothèse d'un lien entre traitement auditif rapide et troubles dys a été remise en cause ces dernières années par différents auteurs présentant des explications alternatives aux troubles dys, avec des facteurs d'ordre sensoriel, attentionnel, ou phonologique (e.g. Ramus, 2001 ; Ramus et al., 2006 ; Hari & Renvall, 2001 ; Goswami, 2015). Ces explications alternatives ne remettent toutefois pas en cause les résultats expérimentaux présentés ici mais seulement le modèle global dans lequel ils s'insèrent.

de comparer l'effet du taux d'information sur le traitement temporel chez des participants normaux et chez des participants présentant des troubles dys.

### *L'efficacité de la reconnaissance de la voix*

Dans nos deux études, il est apparu que la voix se distingue des autres catégories sonores testées (instruments, oiseaux, et véhicules dans la première étude ; instruments dans la deuxième étude). Dans la première étude, la reconnaissance des voix était étonnamment basse, alors que d'autres études comportementales montrent une très bonne reconnaissance de la voix en comparaison à d'autres catégories sonores (Agus et al., 2012 ; Suied et al., 2014). La voix présente un profil acoustique particulier qui semble traité de façon spécifique par le système auditif (Belin, 2006). Notre technique de simplification parcimonieuse des sons ne semble pas s'appliquer aux voix, ce qui ne veut pas dire que les sons de voix ne puissent pas se simplifier efficacement en conservant quelques composantes acoustiques, bien au contraire (e.g. Remez et al., 1981). La représentation auditive des voix semble se distinguer de celle d'autres catégories sonores à cause d'indices spectraux particuliers (Formisano et al., 2008).

Dans la deuxième étude sur la rapidité de reconnaissance de sons naturels, nous avons observé comportementalement une asymétrie en faveur des voix. La reconnaissance des voix est non seulement plus rapide que celle d'instruments de musique, mais les réponses sont aussi données avec plus d'assurance, d'après la mesure de TRs également plus rapides. Ces résultats complètent d'autres résultats récents qui montrent que les TRs pour des voix présentées isolément sont plus courts que pour d'autres catégories sonores (Agus et al., 2012), et que des sons courts sont mieux reconnus, pour une durée donnée, s'il s'agit de voix plutôt que d'autres catégories sonores (Suied et al., 2014).

### *Outils d'analyse en fonction des catégories sonores*

Dans le but de saisir les disparités observées dans la reconnaissance des différentes catégories sonores, nous avons proposé un modèle de distance auditive entre catégories sonores (Isnard et al., 2016). Ce modèle s'appuie sur un modèle

de la périphérie auditive et tient compte des distances entre des sons à l'intérieur d'une catégorie mais également des distances avec les sons d'autres catégories. En combinant ce modèle de distance auditive à une méthode de recherche d'indices auditifs de type *Bubbles* (e.g. Venezia et al., 2016), il devrait théoriquement être possible de chercher automatiquement les indices importants pour la reconnaissance auditive pour des catégories sonores comme la voix où la méthode d'esquisse auditive était moins efficace. Cela constitue un axe de recherche envisagé dans la continuité de cette thèse.

Nous avons également proposé une extension d'un modèle de TDS pour analyser la sensibilité des participants par classe de stimuli dans une tâche m-AFC (Isnard et al., 2016). La modélisation de la sensibilité pourrait constituer un travail de thèse en soi, et nous n'avons fait qu'aborder cette thématique. Nous avons néanmoins pu relever qu'à l'heure actuelle, il existe des moyens computationnels pour calculer la sensibilité en tenant compte du biais, ainsi que d'évaluer leur écart-type.

***Propositions d'applications de ces recherches***

En plus de la contribution théorique et des outils d'analyse de stimuli et de performances de reconnaissance entre catégories sonores, nos résultats pourraient aussi concourir à l'élaboration d'applications concrètes dans différents domaines. Les esquisses auditives sont une proposition qui pourrait servir de base à la sélection d'indices acoustiques ou auditifs pour élaborer des sons de synthèse complexes. L'utilisation de sons naturels apporte d'emblée une complexité spectro-temporelle qui est difficile à atteindre avec de la synthèse sonore qui utilise des composantes plus basiques (e.g. en synthèse additive). De plus, l'utilisation d'un modèle auditif pour sélectionner des composantes discrètes permet aussi de garantir une plus grande fiabilité que dans le cas d'une sélection plus arbitraire.

Un champ d'application des esquisses auditives pourrait être celui des alarmes sonores. En effet, dans un environnement saturé d'informations auditives et visuelles (e.g. un cockpit d'avion), il est impératif que ces informations soient hiérarchisées en fonction de la priorité de traitement qu'elles requièrent, notamment lorsque le temps de réponse requis pour réagir est très limité (e.g. une panne mo-

teur). Or, les informations transmises via la modalité auditive le sont ou bien via des communications vocales, qui nécessitent un niveau d'intelligibilité suffisant, qui sont nécessairement étalées dans la durée, et qui sont difficilement traitées simultanément (Brungart & Simpson, 2005) ; ou bien à partir de sons de synthèse basiques (tons purs modulés) sous la forme d'alarmes sonores dont le lien entre efficacité (e.g. en termes de TRs) et complexité spectro-temporelle n'est que partiellement établi (e.g. Suied et al., 2010). C'est pourquoi des esquisses sonores telles que celles que nous avons proposées pourraient permettre de transmettre une information sémantique sur l'objet auditif de façon efficace, c'est-à-dire avec une bonne reconnaissance et une complexité spectro-temporelle restreinte.

# Annexes

## Annexe A : Calcul de la sensibilité

### A.1 Programme Matlab pour reproduire le calcul de Hacker & Ratcliff (1979)

```
function dprime = hk1979_vincent(hitrate,M)
eq1 = @(x) abs(hitrate - simps(x,M));
dprime = fminsearch(eq1,0);

function hitrate_tmp = simps(dprime_tmp,M)
eq2 = @(y) pdf('norm',y - dprime_tmp,0,1)*cdf('norm',y,0,1)^(M-1);
hitrate_tmp = tsimpsonsrule(eq2,-4,10,175);
```

### A.2 Comparaisons de différentes méthodes de calcul de la sensibilité tenant compte du biais

#### A.2.1 Programmes OpenBUGS

**Modèle 4-AFC avec biais (DeCarlo, 2012).**

```
#4AFC with positional bias parameters
model 4AFC
{
#priors for parameters d, b1, b2, b3
d ~ dnorm(0,.1)
b1 ~ dnorm(0,.1)
b2 ~ dnorm(0,.1)
b3 ~ dnorm(0,.1)

for (i in 1:N) {
   eps1[i] ~ dnorm(0,1.0)
   eps2[i] ~ dnorm(0,1.0)
   eps3[i] ~ dnorm(0,1.0)
   z[i] <- x1[i]-x2[i]
   za[i] <- x1[i]-x3[i]
   zb[i] <- x2[i]-x3[i]
   z1[i] <- 1-2*x1[i]-x2[i]-x3[i]
   z2[i] <- 1-x1[i]-2*x2[i]-x3[i]
   z3[i] <- 1-x1[i]-x2[i]-2*x3[i]
   p1[i,1] <- phi(b1-b2+d*z[i]+eps1[i])*phi(b1-b3+d*za[i]+eps1[i])*phi(b1-d*z1[i]+eps1[i])
   p1[i,2] <- 1-p1[i,1]
   p2[i,1] <- phi(-b1+b2-d*z[i]+eps2[i])*phi(b2-b3+d*zb[i]+eps2[i])*phi(b2-d*z2[i]+eps2[i])
   p2[i,2] <- 1-p2[i,1]
   p3[i,1] <- phi(-b1+b3-d*za[i]+eps3[i])*phi(-b2+b3-d*zb[i]+eps3[i])*phi(b3-d*z3[i]+eps3[i])
   p3[i,2] <- 1-p3[i,1]
   y1[i] ~ dcat(p1[i,1:2])
   y2[i] ~ dcat(p2[i,1:2])
   y3[i] ~ dcat(p3[i,1:2])
    }
}

#data
list(N=117)

#priors
list(d=1,b1=0,b2=0,b3=0)
```

**Modèle 4-AFC par catégorie avec biais (Isnard et al., 2016).**

```
#4AFC with positional bias parameters
model 4AFC
{
#priors for parameters d, b1, b2, b3
d1 ~ dnorm(0,.1)
d2 ~ dnorm(0,.1)
d3 ~ dnorm(0,.1)
d4 ~ dnorm(0,.1)
b1 <- 0.01145
b2 <- 0.004945
b3 <- 0.006606

for (i in 1:N) {
   eps1[i] ~ dnorm(0,1.0)
   eps2[i] ~ dnorm(0,1.0)
   eps3[i] ~ dnorm(0,1.0)
   x4[i] <- 1-x1[i]-x2[i]-x3[i]
   p1[i,1] <- phi(d1*x1[i]-d2*x2[i]+b1-b2+eps1[i])*phi(d1*x1[i]-d3*x3[i]+b1-
b3+eps1[i])*phi(d1*x1[i]-d4*x4[i]+b1+eps1[i])
   p1[i,2] <- 1-p1[i,1]
   p2[i,1] <- phi(d2*x2[i]-d1*x1[i]+b2-b1+eps2[i])*phi(d2*x2[i]-d3*x3[i]+b2-
b3+eps2[i])*phi(d2*x2[i]-d4*x4[i]+b2+eps2[i])
   p2[i,2] <- 1-p2[i,1]
   p3[i,1] <- phi(d3*x3[i]-d1*x1[i]+b3-b1+eps3[i])*phi(d3*x3[i]-d2*x2[i]+b3-
b2+eps3[i])*phi(d3*x3[i]-d4*x4[i]+b3+eps3[i])
   p3[i,2] <- 1-p3[i,1]
   y1[i] ~ dcat(p1[i,1:2])
   y2[i] ~ dcat(p2[i,1:2])
   y3[i] ~ dcat(p3[i,1:2])
   }
}

#data
list(N=117)

#priors
list(d1=1,d2=1,d3=1,d4=1)
```

| Condition | Position cible | "1" | "2" | "3" | Total |
|---|---|---|---|---|---|
| WW | 1 | 54 | 5 | 1 | 60 |
| | 2 | 0 | 60 | 0 | 60 |
| | 3 | 5 | 3 | 52 | 60 |
| SS | 1 | 40 | 12 | 8 | 60 |
| | 2 | 6 | 49 | 4 | 59 |
| | 3 | 6 | 5 | 49 | 60 |

Tableau 6 – Fréquences des réponses pour deux conditions d'une expérience 3-AFC (Ennis & O'Mahony, 1995). "1", "2", et "3" indiquent la fréquence avec laquelle la position était choisie comme étant le signal.

### A.2.2 Tests avec des données expérimentales

**Données expérimentales pour une tâche 3-AFC.**

**Méthodes de calculs.**

1. Ramener la tâche à du 2-AFC en considérant la tâche de reconnaissance comme des 2-AFC multiples : on compare une catégorie par rapport à toutes les autres catégories regroupées.
2. Hacker & Ratcliff (1979) : la table proposée par les auteurs donne les valeurs de $d'$ pour des valeurs de proportion correcte de 0 à 1, et pour différents nombres d'alternatives. La proportion correcte correspond dans ce cas à la somme des valeurs de la diagonale du tableau où sont répertoriées les fréquences de réponses.
3. Hacker & Ratcliff (1979) adapté par catégorie : pour chaque catégorie, on calcule une proportion correcte qui correspond au taux de détections et rejets corrects. Puis on reporte cette valeur dans le tableau de Hacker & Ratcliff (1979) à la colonne 2-AFC.
4. DeCarlo (2012).
5. DeCarlo (2012) sans biais.
6. Isnard et al. (2016) : le biais est calculé d'après la méthode de DeCarlo (2012), puis réinjecté pour faire le calcul de sensibilité par catégorie.
7. Isnard et al. (2016) sans biais.

| Méthodes | Sensibilités | | | | Biais | | |
|---|---|---|---|---|---|---|---|
| | d1 | d2 | d3 | d | b1 | b2 | b3 |
| Equivalent 2-AFC | 3.01 | 3.90 | 3.51 | 3.47 | 0.23 | -0.45 | 0.64 |
| Hacker & Ratcliff (1979) | – | – | – | 2.39 | – | – | – |
| Hacker & Ratcliff (1979) adapté par catégorie | 2.20 | 2.48 | 2.33 | 2.34 | – | – | – |
| DeCarlo (2012) | – | – | – | 2.73 (0.23) | 0.37 (0.30) | 1.00 (0.36) | – |
| DeCarlo (2012) sans biais | – | – | – | 2.40 (0.16) | – | – | – |
| Isnard et al. (2016) | 2.56 (0.22) | 4.61 (1.48) | 2.73 (0.27) | 3.30 | 0.37 | 1.00 | – |
| Isnard et al. (2016) sans biais | 2.09 (0.23) | 5.19 (1.40) | 2.00 (0.28) | 3.09 | – | – | – |

Tableau 7 – Résultats des sensibilités et des biais avec différentes méthodes de calcul pour une expérience 3-AFC avec des données présentant un biais important. La valeur d correspond à la moyenne de d1, d2, et d3 lorsque le calcul est effectué par catégorie. Les valeurs entre parenthèses indiquent les écart-types.

**Résultats des sensibilités et des biais.**

**Données WW : biais important.** Ces données présentent un biais non-négligeable pour la réponse 2, d'après son estimation avec la méthode de DeCarlo (2012) (cf. Tableau 7). L'augmentation du biais augmente la valeur du $d'$ par rapport à celle obtenue avec la méthode de Hacker & Ratcliff (1979), avec laquelle le biais est négligé. Si l'on néglige le biais dans le calcul avec la méthode de DeCarlo, on retrouve bien la valeur de $d'$ calculée avec la méthode de Hacker & Ratcliff (1979). Par ailleurs, le $d'$ pour la deuxième catégorie est très élevé car les résultats expérimentaux sont parfaits pour cette catégorie (maximum de hits, aucune fausse alarme). Même si les résultats des calculs de sensibilité et de biais convergent significativement (cf. DeCarlo (2012)), les écart-types sont plus importants. Qualitativement, on observe les mêmes tendances dans les $d'$ des différentes catégories, avec les différentes méthodes de calcul.

| Méthodes | Sensibilités | | | | Biais | | |
|---|---|---|---|---|---|---|---|
| | d1 | d2 | d3 | d | b1 | b2 | b3 |
| Equivalent 2-AFC | 1.71 | 2.03 | 2.18 | 1.97 | 0.42 | 0.06 | 0.19 |
| Hacker & Ratcliff (1979) | – | – | – | 1.52 | – | – | – |
| Hacker & Ratcliff (1979) adapté par catégorie | 1.29 | 1.47 | 1.59 | 1.45 | – | – | – |
| DeCarlo (2012) | – | – | – | 1.48 (0.12) | -0.28 (0.20) | 0.05 (0.21) | – |
| DeCarlo (2012) sans biais | – | – | – | 1.45 (0.12) | – | – | – |
| Isnard et al. (2016) | 1.35 (0.18) | 1.52 (0.21) | 1.62 (0.26) | 1.50 | -0.28 | 0.05 | – |
| Isnard et al. (2016) sans biais | 1.05 (0.18) | 1.75 (0.21) | 1.72 (0.26) | 1.51 | – | – | – |

Tableau 8 – Résultats des sensibilités et des biais avec différentes méthodes de calcul pour une expérience 3-AFC avec des données présentant un biais faible. La valeur d correspond à la moyenne de d1, d2, et d3 lorsque le calcul est effectué par catégorie. Les valeurs entre parenthèses indiquent les écart-types.

**Données SS : biais faible.** Cette fois, les données présentent des biais faibles, d'après leur estimation avec la méthode de DeCarlo (2012) (cf. Tableau 8). Les sensibilités sont aussi plus faibles et divergent moins. L'arrondi au dixième les rend globalement équivalentes.

# Annexe B : Calcul de distance auditive

La mesure de distance auditive élaborée par Agus et al. (2012), et reprise entre des catégories sonores par Isnard et al. (2016), consiste à calculer des distances entre des représentations temps-fréquence auditives (STEPs ; Moore, 2003) à l'aide d'un algorithme de DTW. La Figure 35 donne quelques exemples de distances mesurées entre des matrices particulières illustrant des représentations temps-fréquence. L'algorithme permet de faire une correspondance temporelle entre des évènements temporels placés aléatoirement s'ils sont en nombre équivalent, tandis que l'augmentation et la diversité des composantes fréquentielles accroît significativement le résultat de la mesure de distance.

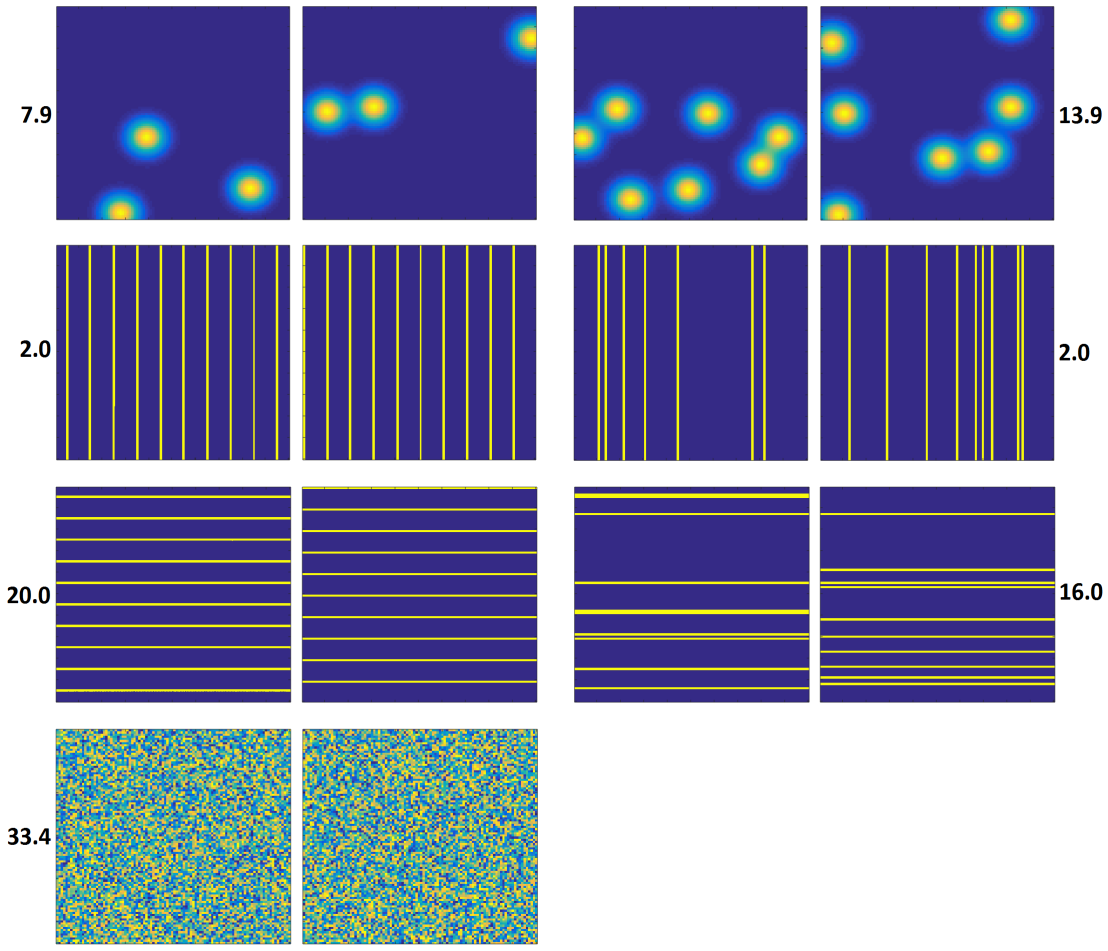


FIGURE 35 – Exemples de calculs de distances auditives pour des matrices caractéristiques. La similarité est évaluée entre des paires de matrices à l'aide d'un algorithme de *dynamic time warping*, illustrant des représentations auditives temps-fréquence, et indiquée de chaque côté des paires de matrices. Chaque ligne illustre un cas particulier : (1) des bulles gaussiennes réparties aléatoirement (3 et 7 bulles respectivement), (2) des clics temporels réguliers ou irréguliers (10 évènements), (3) une répartition de 10 partiels stationnaires harmoniques ou non, (4) du bruit. La mesure de similarité est fortement dépendante de la dimension fréquentielle, tant que le nombre d'évènements temporels est équivalent entre deux matrices.

# Références

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature*, 433(7021) :68–72.

Agus, T. R., Suied, C., Thorpe, S. J., & Pressnitzer, D. (2012). Fast recognition of musical sounds based on timbre. *J Acoust Soc Am*, 131(5) :4124–33.

Ahumada, A. J. & Lovell, J. (1971). Stimulus features in signal detection. *The Journal of the Acoustical Society of America*, 49(6B) :1751–1756.

Ahumada, A. J., Marken, R., & Sandusky, A. (1975). Time and frequency analyses of auditory signal detection. *The Journal of the Acoustical Society of America*, 57(2) :385–390.

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., & Grady, C. L. (2001). "What" and "where" in the human auditory system. *Proceedings of the National Academy of Sciences*, 98(21) :12301–12306.

ANSI (1973). S3. 20. *New York, NY : American National Standards Institute.*

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological review*, 61(3) :183.

Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of experimental psychology : human perception and performance*, 19(2) :250.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. *Sensory Communications, Rosenblith W.A. (ed), MIT press*, pages 217–234.

Bathellier, B., Ushakova, L., & Rumpel, S. (2012). Discrete neocortical dynamics predict behavioral categorization of sounds. *Neuron*, 76(2) :435–449.

Belin, P. (2006). Voice processing in human and non-human primates. *Philos Trans R Soc Lond B Biol Sci*, 361(1476) :2091–107.

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice : neural correlates of voice perception. *Trends Cogn Sci*, 8(3) :129–35.

Belin, P. & Grosbras, M.-H. (2010). Before speech : cerebral voice processing in infants. *Neuron*, 65(6) :733–735.

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research*, 13(1) :17–26.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403(6767) :309–312.

Benasich, A. A., Thomas, J. J., Choudhury, N., & Leppanen, P. H. (2002). The importance of rapid auditory processing abilities to early language development : evidence from converging methodologies. *Developmental psychobiology*, 40(3) :278–292.

Bestelmeyer, P. E., Maurage, P., Rouger, J., Latinus, M., & Belin, P. (2014). Adaptation to vocal expressions reveals multistep perception of auditory emotion. *The Journal of Neuroscience*, 34(24) :8098–8105.

Bestelmeyer, P. E., Rouger, J., DeBruine, L. M., & Belin, P. (2010). Auditory adaptation in vocal affect perception. *Cognition*, 117(2) :217–223.

Bigand, E., Delbe, C., Gerard, Y., & Tillmann, B. (2011). Categorization of extremely brief auditory stimuli : Domain-specific or domain-general processes ? *PloS one*, 6(10) :e27024.

Bleeck, S., Ives, T., & Patterson, R. D. (2004). Aim-mat : the auditory image model in matlab. *Acta Acustica United with Acustica*, 90(4) :781–787.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences*, volume 17, pages 97–110. Amsterdam.

Boersma, P. & Weenink, D. (2015). Praat : doing phonetics by computer [computer program]. version 5.4.12, retrieved 10 july 2015 from http ://www.praat.org/.

Boulez, P. (1987). Timbre and composition - timbre and language. *Contemporary Music Review*, 2(1) :161–171.

Boulez, P. & Archimbaud, M. (2016). *Entretiens avec Michel Archimbaud*. Editions Gallimard.

Brungart, D. S. & Simpson, B. D. (2005). Improving multitalker speech communication with advanced audio displays. Report, DTIC Document.

Buffat, S., Plantier, J., Roumes, C., & Lorenceau, J. (2012). Repetition blindness for natural images of objects with viewpoint changes. *Front Psychol*, 3 :622.

Buffat, S., Plantier, J., Roumes, C., & Lorenceau, J. (2013). Repetition blindness for natural images of objects with viewpoint changes. *Frontiers in psychology*, 3.

Caclin, A. (2004). *Interactions et independances entre dimensions du timbre des sons complexes : approche psychophysique et electrophysiologique chez l'Humain*. Thesis.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions : A confirmatory study using synthetic tones. *The Journal of the Acoustical Society of America*, 118(1) :471.

Carroll, J. D. & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, 35(3) :283–319.

Carron, M. (2016). *Methodes et outils pour definir et vehiculer une identite sonore : application au design sonore identitaire de la marque SNCF*. Thesis.

Cavanagh, P. (2005). The artist as neuroscientist. *Nature*, 434(7031) :301–307.

Charest, I., Pernet, C. R., Rousselet, G. A., Quinones, I., Latinus, M., Fillion-Bilodeau, S., Chartrand, J.-P., & Belin, P. (2009). Electrophysiological evidence for an early processing of human voices. *Bmc Neuroscience*, 10(1) :127.

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2) :887.

Chon, S. & McAdams, S. (2012). Exploring blending as a function of timbre saliency. In *Proceedings of the 12th International Conference of Music Perception and Cognition*.

Chon, S. H., Schwartzbach, K., Smith, B., & McAdams, S. (2013). Effect of timbre on melody recognition in three-voice counterpoint music. In *Proceedings of the Sound and Music Computing Conference 2013*.

Chun, M. M. & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental psychology : Human perception and performance*, 21(1) :109.

Clark, M., Luce, D., Abrams, R., Schlossberg, H., & Rome, J. (1963). Preliminary experiments on the aural significance of parts of tones of orchestral instruments and on choral tones. *Journal of the Audio Engineering Society*, 11(1) :45–54.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3) :1562–1573.

Craven, B. (1992). A table of d' for m-alternative odd-man-out forced-choice procedures. *Perception & psychophysics*, 51(4) :379–385.

Creel, W., Boomsliter, P. C., & Powers, S. R. (1970). Sensations of tone as perceptual forms. *Psychological Review*, 77(6) :534.

Cusack, R. & Carlyon, R. P. (2003). Perceptual asymetries in audition. *Journal of Experimental Psychology : Human Perception and Performance*, 29(3) :713–725.

Cutzu, F. & Edelman, S. (1996). Faithful representation of similarities among three-dimensional shapes in human vision. *Proceedings of the National Academy of Sciences*, 93(21) :12046–12050.

Dai, H. & Micheyl, C. (2010). Psychophysical reverse correlation with multiple response alternatives. *J Exp Psychol Hum Percept Perform*, 36(4) :976–93.

De Lucia, M., Clarke, S., & Murray, M. M. (2010). A temporal hierarchy for conspecific vocalization discrimination in humans. *The Journal of Neuroscience*, 30(33) :11210–11221.

DeCarlo, L. T. (2012). On a signal detection approach to -alternative forced choice with bias, with maximum likelihood and bayesian approaches to estimation. *Journal of Mathematical Psychology*, 56(3) :196–207.

deCharms, C. R., Blake, D. T., & Merzenich, M. M. (1998). Optimizing sound features for cortical neurons. *Science*, 280(5368) :1439–1444.

Dirks, D. D. & Bower, D. (1970). Effect of forward and backward masking on speech intelligibility. *The Journal of the Acoustical Society of America*, 47(4B) :1003–1008.

Donders, F. C. (1969). On the speed of mental processes. *Acta psychologica*, 30 :412–431.

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *The Journal of the Acoustical Society of America*, 102(4) :2403–2411.

Duncan, J., Martens, S., & Ward, R. (1997). Within but not between sensory modalities. *Nature*, 387 :809.

Efron, R. (1970a). The minimum duration of a perception. *Neuropsychologia*, 8(1) :57–63.

Efron, R. (1970b). The relationship between the duration of a stimulus and the duration of a perception. *Neuropsychologia*, 8(1) :37–55.

Ehresman, D. & Wessel, D. L. (1978). Perception of timbral analogies. *Rapport IRCAM No13*.

Elad, M. (2012). Sparse and redundant representation modeling - what next ? *IEEE Signal Processing Letters*, 19(12) :922–928.

Elliot, P. B. (1964). Tables of d'. *In Swets, JA (Eds), Signal detection and recognition by human observers. New York : Wiley.*

Elliott, C. A. (1975). Attacks and releases as factors in instrument identification. *Journal of Research in Music Education*, 23(1) :35–40.

Elliott, L. L. (1967). Development of auditory narrow-band frequency contours. *The Journal of the Acoustical Society of America*, 42(1) :143–153.

Elliott, T. M., Hamilton, L. S., & Theunissen, F. E. (2013). Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. *J Acoust Soc Am*, 133(1) :389–404.

Emmerich, D. S., Gray, J. L., Watson, C. S., & Tanis, D. C. (1972). Response latency, confidence, and rocs in auditory signal detection. *Perception & Psychophysics*, 11(1) :65–72.

Ennis, D. M. & O'Mahony, M. (1995). Probabilistic models for sequential taste effects in triadic choice. *Journal of Experimental Psychology : Human Perception and Performance*, 21(5) :1088.

Fecteau, S., Armony, J. L., Joanette, Y., & Belin, P. (2004). Is voice processing species-specific in human auditory cortex ? an fmri study. *Neuroimage*, 23(3) :840–8.

Felsen, G. & Dan, Y. (2005). A natural approach to studying vision. *Nat Neurosci*, 8(12) :1643–6.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A*, 4(12) :2379–2394.

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "who" is saying "what" ? brain-based decoding of human voice and speech. *Science*, 322(5903) :970–3.

Franco, A., Eberlen, J., Destrebecqz, A., Cleeremans, A., & Bertels, J. (2015). Rapid serial auditory presentation : A new measure of statistical learning in speech segmentation. *Experimental psychology.*

Frenkel, M., Sherman, G. F., Bashan, K. A., Galaburda, A. M., & LoTurco, J. J. (2000). Neocortical ectopias are associated with attenuated neurophysiological responses to rapidly changing auditory stimuli. *Neuroreport*, 11(3) :575–579.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience*, 25(20) :5004–5012.

Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., & Belin, P. (2013). Abstract encoding of auditory objects in cortical activity patterns. *Cereb Cortex*, 23(9) :2025–37.

Giordano, B. L., McDonnell, J., & McAdams, S. (2010). Hearing living symbols and nonliving icons : category specificities in the cognitive processing of environmental sounds. *Brain Cogn*, 73(1) :7–19.

Glasberg, B. R. & Moore, B. C. (2002). A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society*, 50(5) :331–342.

Gosselin, F. & Schyns, P. G. (2001). Bubbles : a technique to reveal the use of information in recognition tasks. *Vision research*, 41(17) :2261–2271.

Gosselin, F. & Schyns, P. G. (2002). Rap : A new framework for visual categorization. *Trends in cognitive sciences*, 6(2) :70–77.

Gosselin, F. & Schyns, P. G. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological Science*, 14(5) :505–509.

Gosselin, F. & Schyns, P. G. (2004). No troubles with bubbles : a reply to murray and gold. *Vision Research*, 44(5) :471–477.

Gosselin, F. & Schyns, P. G. (2005). Bubbles : A user's guide. *Building object categories in developmental time*, pages 91–106.

Goswami, U. (2015). Sensory theories of developmental dyslexia : three challenges for research. *Nat Rev Neurosci*, 16(1) :43–54.

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2003). Rwc music database : Music genre database and musical instrument sound database. In *ISMIR*, volume 3, pages 229–230.

Gray, G. W. (1942). Phonemic microtomy : The minimum duration of perceptible speech sounds. *Communications Monographs*, 9(1) :75–90.

Green, D. & Swets, J. (1966). Signal detection theory and psychophysics. *New York*, 888 :889.

Green, D. M. (1964). Consistency of auditory detection judgments. *Psychological review*, 71(5) :392.

Grey, J. M. (1975). *An exploration of musical timbre*. Dept. of Music, Stanford University.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, 61(5) :1270–1277.

Grey, J. M. & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America*, 63(5) :1493–1500.

Grey, J. M. & Moorer, J. A. (1977). Perceptual evaluations of synthesized musical instrument tones. *The Journal of the Acoustical Society of America*, 62(2) :454–462.

Griffiths, T. D. & Warren, J. D. (2004). What is an auditory object ? *Nature Reviews Neuroscience*, 5(11) :887–892.

Gygi, B., Kidd, G. R., & Watson, C. S. (2004). Spectral-temporal factors in the identification of environmental sounds. *The Journal of the Acoustical Society of America*, 115(3) :1252.

Gygi, B., Kidd, G. R., & Watson, C. S. (2007). Similarity and categorization of environmental sounds. *Perception & psychophysics*, 69(6) :839–855.

Hacker, M. J. & Ratcliff, R. (1979). A revised table of d' for m-alternative forced choice. *Attention, Perception, & Psychophysics*, 26(2) :168–170.

Halpern, A. R., Zatorre, R. J., Bouffard, M., & Johnson, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, 42(9) :1281–92.

Handel, S. & Erickson, M. L. (2001). A rule of thumb : The bandwidth for timbre invariance is one octave. *Music Perception : An Interdisciplinary Journal*, 19(1) :121–126.

Harding, S., Cooke, M., & Konig, P. (2007). Auditory gist perception : an alternative to attentional selection of auditory streams ? In *International Workshop on Attention in Cognitive Systems*, pages 399–416. Springer.

Hari, R. (1995). Illusory directional hearing in humans. *Neuroscience Letters*, 189(1) :29–30.

Hari, R. & Kiesila, P. (1996). Deficit of temporal auditory processing in dyslexic adults. *Neuroscience letters*, 205(2) :138–140.

Hari, R. & Renvall, H. (2001). Impaired processing of rapid stimulus sequences in dyslexia. *Trends in cognitive sciences*, 5(12) :525–532.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539) :2425–2430.

Helmholtz, H. (1895). On the sensation of tone as a physiological basis for the theory of music (AJ Ellis, Trans.). *New York : Longman, Green, and Co. (Original work published 1877).*

Hromadka, T., DeWeese, M. R., & Zador, A. M. (2008). Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biol*, 6(1) :e16.

Hromadka, T. & Zador, A. M. (2009). Representations in auditory cortex. *Current opinion in neurobiology*, 19(4) :430–433.

Isnard, V., Taffou, M., Viaud-Delmon, I., & Suied, C. (2016). Auditory sketches : Very sparse representations of sounds are still recognizable. *PloS one*, 11(3) :e0150313.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11) :1254–1259.

Iverson, P. & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. *The Journal of the Acoustical Society of America*, 94(5) :2595–2603.

Jernigan, T. L., Hesselink, J. R., Sowell, E., & Tallal, P. A. (1991). Cerebral structure on magnetic resonance imaging in language-and learning-impaired children. *Archives of neurology*, 48(5) :539–545.

Johnsrude, I. S., Zatorre, R. J., Milner, B. A., & Evans, A. C. (1997). Left-hemisphere specialization for the processing of acoustic transients. *NeuroReport*, 8(7) :1761–1765.

Joos, M. (1948). Acoustic phonetics. *Language*, 24(2) :5–136.

Kaas, J. H. & Hackett, T. A. (1999). 'What' and 'where' processing in auditory cortex. *Nature neuroscience*, 2(12).

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area : a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience*, 17(11) :4302–4311.

Kaplan, J. T. & Iacoboni, M. (2005). Listen to my actions ! *Behavioral and Brain Sciences*, 28(02) :135–136.

Kaplan, J. T. & Iacoboni, M. (2007). Multimodal action representation in human left ventral premotor cortex. *Cognitive Processing*, 8(2) :103–113.

Kapoor, A. & Allen, J. B. (2012). Perceptual effects of plosive feature modification. *The Journal of the Acoustical Society of America*, 131(1) :478–491.

Karadogan, S. G., Larsen, J., SyskindPedersen, M., & Boldt, J. B. (2010). Robust isolated speech recognition using binary masks. In *Signal Processing Conference, 2010 18th European*, pages 1988–1992. IEEE.

Kawahara, H. & Matsui, H. (2003). Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In *Acoustics, Speech, and Signal Processing (ICASSP'03)*, volume 1, pages 256–259. IEEE.

Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention : an auditory saliency map. *Curr Biol*, 15(21) :1943–7.

Kendall, R. A., Carterette, E. C., & Hajda, J. M. (1999). Perceptual and acoustical features of natural and synthetic orchestral instrument tones. *Music Perception : An Interdisciplinary Journal*, 16(3) :327–363.

Keysers, C., Xiao, D., Foldiak, P., & Perrett, D. I. (2001). The speed of sight. *Cognitive Neuroscience, Journal of*, 13(1) :90–101.

King, A. J. & Nelken, I. (2009). Unraveling the principles of auditory cortical processing : can we learn from the visual system ? *Nat Neurosci*, 12(6) :698–701.

Klein, M. D. & Stolz, J. A. (2015). Looking and listening : A comparison of intertrial repetition effects in visual and auditory search tasks. *Atten Percept Psychophys*, 77(6) :1986–97.

Kording, K. P., Konig, P., & Klein, D. J. (2002). Learning of sparse auditory receptive fields. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, volume 2, pages 1103–1108.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2 :4.

Kriegstein, K. V. & Giraud, A.-L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage*, 22(2) :948–955.

Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caracterisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique. *Le Journal de Physique IV*, 04(C5) :C5–625–C5–628.

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand. *Structure and perception of electroacoustic sound and music*, 9 :43–53.

Lakatos, P., Pincze, Z., Fu, K.-M. G., Javitt, D. C., Karmos, G., & Schroeder, C. E. (2005a). Timing of pure tone and noise-evoked responses in macaque auditory cortex. *Neuroreport*, 16(9) :933–937.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005b). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of neurophysiology*, 94(3) :1904–1911.

Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, 23(12) :1075–1080.

Leaver, A. M. & Rauschecker, J. P. (2010). Cortical representation of natural complex sounds : effects of acoustic features and auditory object category. *J Neurosci*, 30(22) :7604–12.

Lemaitre, G., Susini, P., Winsberg, S., & McAdams, S. (2003). Perceptively based design of new car horn sounds.

Lemaitre, G., Susini, P., Winsberg, S., McAdams, S., & Letinturier, B. (2007). The sound quality of car horns : a psychoacoustical study of timbre. *Acta acustica united with Acustica*, 93(3) :457–468.

Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, 442(7102) :572–575.

Levy, D. A., Granot, R., & Bentin, S. (2001). Processing specificity for human voice stimuli : electrophysiological evidence. *Neuroreport*, 12(12) :2653–2657.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nat Neurosci*, 5(4) :356–63.

Lewis, J. W. (2006). Cortical networks related to human use of tools. *The Neuroscientist*, 12(3) :211–231.

Lewis, J. W., Brefczynski, J. A., Phinney, R. E., Janik, J. J., & DeYoe, E. A. (2005). Distinct cortical pathways for processing tool versus animal sounds. *J Neurosci*, 25(21) :5148–58.

Lewis, J. W., Talkington, W. J., Walker, N. A., Spirou, G. A., Jajosky, A., Frum, C., & Brefczynski-Lewis, J. A. (2009). Human cortical organization for processing vocalizations indicates representation of harmonic structure as a signal attribute. *J Neurosci*, 29(7) :2283–96.

Li, F., Menon, A., & Allen, J. B. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *J Acoust Soc Am*, 127(4) :2599–610.

Liberman, A. M. & Mattingly, I. G. (1989). A specialization for speech perception. *Science*, 243(4890) :489–494.

Liegeois-Chauvel, C., Musolino, A., Badier, J., Marquis, P., & Chauvel, P. (1994). Evoked potentials recorded from the auditory cortex in man : evaluation and topography of the middle latency components. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 92(3) :204–214.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *The Journal of the acoustical society of America*, 96(4) :2076–2087.

Livingstone, M. & Hubel, D. (1988). Segregation of form, color, movement, and depth : anatomy, physiology, and perception.

Lomber, S. G. & Malhotra, S. (2008). Double dissociation of 'what' and 'where' processing in auditory cortex. *Nature neuroscience*, 11(5) :609–616.

Luce, R. D. (1963). *Detection and recognition*, pages 103–189. New York : Wiley.

Lyon, R. (1982). A computational model of filtering, detection, and compression in the cochlea. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, volume 7, pages 1282–1285. IEEE.

Lyon, R. (1984). Computational models of neural auditory processing. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, volume 9, pages 41–44. IEEE.

Lyon, R. & Shamma, S. (1996). *Auditory representations of timbre and pitch*, pages 221–270. Springer.

Lyon, R. F., Katsiamis, A. G., & Drakakis, E. M. (2010). History and future of auditory filter models. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 3809–3812. IEEE.

Lyon, R. F. & Mead, C. (1988). An analog electronic cochlea. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7) :1119–1134.

Macherey, O. & Delpierre, A. (2013). Perception of musical timbre by cochlear implant listeners : a multidimensional scaling study. *Ear and hearing*, 34(4) :426–436.

Macmillan, N. & Creelman, C. (2005). Detection theory : A user's guide lawrence erlbaum associates. *New York*.

Maeder, P. P., Meuli, R. A., Adriani, M., Bellmann, A., Fornari, E., Thiran, J.-P., Pittet, A., & Clarke, S. (2001). Distinct pathways involved in sound recognition and localization : a human fMRI study. *Neuroimage*, 14(4) :802–816.

Mandel, M. (2013). Learning an intelligibility map of individual utterances. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, pages 1–4. IEEE.

Mandel, M. I., Yoho, S. E., & Healy, E. W. (2014). Generalizing time-frequency importance functions across noises, talkers, and phonemes. In *Fifteenth Annual Conference of the International Speech Communication Association.*

Mangini, M. & Biederman, I. (2004). Making the ineffable explicit : estimating the information employed for face classifications. *Cognitive Science*, 28(2) :209–226.

Marozeau, J., deCheveigne, A., McAdams, S., & Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *The Journal of the Acoustical Society of America*, 114(5) :2946–2957.

Massaro, D. W. (1970). Preperceptual auditory images. *Journal of Experimental Psychology*, 85(3) :411.

Massaro, D. W. (1971). Effect of masking tone duration on preperceptual auditory images. *Journal of Experimental Psychology*, 87(1) :146.

Massaro, D. W. (1972a). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, 79(2) :124.

Massaro, D. W. (1972b). Stimulus information vs processing time in auditory pattern recognition. *Perception & Psychophysics*, 12(1) :50–56.

Massaro, D. W. (1975). Backward recognition masking. *The Journal of the Acoustical Society of America*, 58(5) :1059–1065.

Massaro, D. W. (1977). Rate of perceptual processing. *Psychological research*, 39(3) :277–283.

McAdams, S. (1994). Reconnaissance de sources et d'evenements sonores. *Penser les sons, Psychologie cognitive de l'audition*, pages 157–214.

McAdams, S. & Cunible, J.-C. (1992). Perception of timbral analogies. *Philosophical Transactions of the Royal Society of London B : Biological Sciences*, 336(1278) :383–389.

McAdams, S., Susini, P., Misdariis, N., & Winsberg, S. (1998). Multidimensional characterisation of perceptual and preference judgements of vehicle and environmental noises. In *Euro-Noise 98*.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes. *Psychological research*, 58(3) :177–192.

McDermott, H. J. (2004). Music perception with cochlear implants : a review. *Trends in amplification*, 8(2) :49–82.

McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nat Neurosci*, 16(4) :493–8.

McDermott, J. H. & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery : evidence from sound synthesis. *Neuron*, 71(5) :926–40.

Meddis, R. (1986). Simulation of mechanical to neural transduction in the auditory receptor. *The Journal of the Acoustical Society of America*, 79(3) :702–711.

Meddis, R. (1988). Simulation of auditory-neural transduction : Further studies. *The Journal of the Acoustical Society of America*, 83(3) :1056–1063.

Meddis, R. & Hewitt, M. J. (1991a). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i : Pitch identification. *The Journal of the Acoustical Society of America*, 89(6) :2866–2882.

Meddis, R. & Hewitt, M. J. (1991b). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. ii : Phase sensitivity. *The Journal of the Acoustical Society of America*, 89(6) :2883–2894.

Meddis, R., Hewitt, M. J., & Shackleton, T. M. (1990). Implementation details of a computation model of the inner hair-cell auditory-nerve synapse. *The Journal of the Acoustical Society of America*, 87(4) :1813–1816.

Miller, J. R. & Carterette, E. C. (1975). Perceptual space for musical structures. *The Journal of the Acoustical Society of America*, 58(3) :711–720.

Minard, A., Susini, P., Misdariis, N., Lemaitre, G., Mcadams, S., & Parizet, E. (2008). Environmental sound description : a meta-analysis of timbre perception. In *Workshop SID, CHI, Florence*.

Misdariis, N., Minard, A., Susini, P., Lemaitre, G., McAdams, S., & Parizet, E. (2010). Environmental sound perception : Metadescription and modeling based

on independent primary studies. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010 :1–26.

Misdariis, N., Smith, B. K., Pressnitzer, D., Susini, P., & McAdams, S. (1998). Validation of a multidimensional distance model for perceptual dissimilarities among musical timbres. In *Proc. of Joint meeting of the 16th congress on ICA, 135th meeting of ASA*.

Moerel, M., De Martino, F., & Formisano, E. (2012). Processing of natural sounds in human auditory cortex : tonotopy, spectral tuning, and relation to voice sensitivity. *The Journal of Neuroscience*, 32(41) :14205–14216.

Moore, B. C. (2012). *An introduction to the psychology of hearing*. Brill.

Moore, B. C. J. (2003). Temporal integration and context effects in hearing. *Journal of Phonetics*, 31(3-4) :563–574.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of experimental psychology*, 64(5) :482.

Murray, M. M., Camen, C., Andino, S. L. G., Bovet, P., & Clarke, S. (2006). Rapid brain discrimination of sounds of objects. *The Journal of Neuroscience*, 26(4) :1293–1302.

Murray, R. F. (2011). Classification images : A review. *J Vis*, 11(5).

Murray, R. F. & Gold, J. M. (2004). Troubles with bubbles. *Vision Research*, 44(5) :461–470.

Narayanan, A. & Wang, D. (2010). Robust speech recognition from binary masks. *J Acoust Soc Am*, 128(5) :EL217–22.

Navon, D. (1977). Forest before trees : The precedence of global features in visual perception. *Cognitive psychology*, 9(3) :353–383.

Nelken, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex. *Curr Opin Neurobiol*, 14(4) :474–80.

Nelken, I. & de Cheveigne, A. (2013). An ear for statistics. *Nat Neurosci*, 16(4) :381–2.

Nelken, I., Rotman, Y., & Yosef, O. B. (1999). Responses of auditory-cortex neurons to structural features of natural sounds. *Nature*, 397(6715) :154–157.

Neri, P. & Heeger, D. J. (2002). Spatiotemporal mechanisms for detecting and identifying image features in human vision. *Nat Neurosci*, 5(8) :812–6.

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, 88(6) :1281–1296.

Occelli, F., Suied, C., Pressnitzer, D., Edeline, J. M., & Gourevitch, B. (2015). A neural substrate for rapid timbre recognition ? neural and behavioral discrimination of very brief acoustic vowels. *Cereb Cortex*.

O'Connor, K. N., Petkov, C. I., & Sutter, M. L. (2005). Adaptive stimulus optimization for auditory cortical neurons. *Journal of Neurophysiology*, 94(6) :4051–4067.

Oliva, A. & Schyns, P. G. (1997). Coarse blobs or fine edges ? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive psychology*, 34 :72–107.

Oliva, A. & Torralba, A. (2006). Building the gist of a scene : The role of global image features in recognition. *Progress in brain research*, 155 :23–36.

Olshausen, B. A. & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *NATURE*, 381 :13.

Olshausen, B. A. & Field, D. J. (1997). Sparse coding with an overcomplete basis set : A strategy employed by v1 ? *Vision research*, 37(23) :3311–3325.

Olshausen, B. A. & Field, D. J. (2004). Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4) :481–487.

Olshausen, B. A. & O'Connor, K. N. (2002). A new window on sound. *Nature neuroscience*, 5(4) :292–294.

Overath, T., Kumar, S., Stewart, L., von Kriegstein, K., Cusack, R., Rees, A., & Griffiths, T. D. (2010). Cortical mechanisms for the segregation and representation of acoustic textures. *J Neurosci*, 30(6) :2070–6.

Parizet, E., Guyader, E., & Nosulenko, V. (2008). Analysis of car door closing sound quality. *Applied Acoustics*, 69(1) :12–22.

Patil, K., Pressnitzer, D., Shamma, S., & Elhilali, M. (2012). Music in our ears : the biological bases of musical timbre perception. *PLoS Comput Biol*, 8(11) :e1002759.

Patterson, R. D. (2000). Auditory images. how complex sounds are represented in the auditory system. *Journal of the Acoustical Society of Japan (E)*, 21(4) :183–190.

Patterson, R. D., Allerhand, M. H., & Giguere, C. (1995). Time-domain modeling of peripheral auditory processing : A modular architecture and a software platform. *The Journal of the Acoustical Society of America*, 98(4) :1890–1894.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., & Allerhand, M. (1992). Complex sounds and auditory images. *Auditory physiology and perception*, 83 :429–446.

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project.

Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E., Watson, R. H., Fleming, D., Crabbe, F., & Valdes-Sosa, M. (2015). The human voice areas : Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage*, 119 :164–174.

Peterson, G. E. (1939). The significance of various portions of the wave length in the minimum duration necessary for the recognition of vowel sounds. *Unpublished doctoral dissertation, Department of Speech, Louisiana State University.*

Pizzamiglio, L., Aprile, T., Spitoni, G., Pitzalis, S., Bates, E., D'Amico, S., & Di Russo, F. (2005). Separate neural systems for processing action-or non-action-related sounds. *Neuroimage*, 24(3) :852–861.

Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. *Frequency analysis and periodicity detection in hearing*, pages 397–414.

Plomp, R. (1975). Auditory analysis and timbre perception. *Auditory analysis and perception of speech*, pages 7–22.

Plomp, R. (1979). Fysikaliska motsvarigheter till klanfarg hos stationara ljud. *Var horsel och musiken. Stockholm : Klung. Musikaliska Akademien.*

Plomp, R. & Steeneken, H. (1969). Effect of phase on the timbre of complex tones. *The Journal of the Acoustical Society of America*, 46(2B) :409–421.

Plumbley, M. D., Blumensath, T., Daudet, L., Gribonval, R., & Davies, M. E. (2010). Sparse representations in audio and music : from coding to source separation. In *Proceedings of the IEEE*, volume 98, pages 995–1005.

Ponsot, E., Susini, P., & Meunier, S. (2015). A robust asymmetry in loudness between rising- and falling-intensity tones. *Atten Percept Psychophys*, 77(3) :907–20.

Posner, M. I. (1978). *Chronometric explorations of mind.* Lawrence Erlbaum.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of experimental psychology : human learning and memory*, 2(5) :509.

Powell, R. L. & Tosi, O. (1970). Vowel recognition threshold as a function of temporal segmentations. *Journal of Speech, Language, and Hearing Research*, 13(4) :715–724.

Pressnitzer, D., Agus, T., & Suied, C. (2013). *Acoustic Timbre Recognition*, pages 1–6. Springer New York, New York, NY.

Pressnitzer, D., Patterson, R. D., & Krumbholz, K. (2001). The lower limit of melodic pitch. *The Journal of the Acoustical Society of America*, 109(5) :2074–2084.

Qi, Y. & Hillman, R. E. (1997). Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals. *The Journal of the Acoustical Society of America*, 102(1) :537–543.

Rabinowitz, N. C., Willmore, B. D., King, A. J., & Schnupp, J. W. (2013). Constructing noise-invariant representations of sound in the auditory pathway. *PLoS Biol*, 11(11) :e1001710.

Ramus, F. (2001). Dyslexia : Talk of two theories. *Nature*, 412(6845) :393–395.

Ramus, F., White, S., & Frith, U. (2006). Weighing the evidence between competing theories of dyslexia. *Developmental Science*, 9(3) :265–269.

Rasch, R. & Plomp, R. (1982). The perception of musical tones. *The psychology of music*, 2 :89–112.

Rauschecker, J. P. (1998). Cortical processing of complex sounds. *Current opinion in neurobiology*, 8(4) :516–521.

Rauschecker, J. P. & Scott, S. K. (2009). Maps and streams in the auditory cortex : nonhuman primates illuminate human speech processing. *Nature neuroscience*, 12(6) :718–724.

Rauschecker, J. P. & Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences*, 97(22) :11800–11806.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an rsvp task : An attentional blink ? *Journal of experimental psychology : Human perception and performance*, 18(3) :849.

Redies, C. (2007). A universal model of esthetic perception based on the sensory coding of natural stimuli. *Spatial vision*, 21(1) :97–117.

Remez, R. E., Pardo, J. S., Piorkowski, R. L., & Rubin, P. E. (2001). On the bistability of sine wave analogues of speech. *Psychological Science*, 12(1) :24–29.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, 212(4497) :947–949.

Remez, R. E. & Thomas, E. F. (2013). Early recognition of speech. *Wiley Interdisciplinary Reviews : Cognitive Science*, 4(2) :213–223.

Riede, T., Herzel, H., Hammerschmidt, K., Brunnberg, L., & Tembrock, G. (2001). The harmonic-to-noise ratio applied to dog barks. *The Journal of the Acoustical Society of America*, 110(4) :2191–2197.

Robinson, K. & Patterson, R. D. (1995a). The duration required to identify the instrument, the octave, or the pitch chroma of a musical note. *Music Perception*, pages 1–15.

Robinson, K. & Patterson, R. D. (1995b). The stimulus duration required to identify vowels, their octave, and their pitch chroma. *The Journal of the Acoustical Society of America*, 98(4) :1858–1865.

Rocchesso, D., Mauro, D. A., & Monache, S. D. (2016). mimic : The microphone as a pencil. In *Proceedings of the TEI'16 : Tenth International Conference on Tangible, Embedded, and Embodied Interaction*, pages 357–364. ACM.

Rolls, E. T. & Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology*, 73(2) :713–726.

Rolls, E. T., Tovee, M. J., & Panzeri, S. (1999). The neurophysiology of backward visual masking : information analysis. *Journal of Cognitive Neuroscience*, 11(3) :300–311.

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., & Rauschecker, J. P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nature neuroscience*, 2(12) :1131–1136.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8 :382–439.

Sakoe, H. & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1) :43–49.

Samson, S., Zatorre, R. J., & Ramsay, J. O. (1997). Multidimensional scaling of synthetic musical timbre : Perception of spectral and temporal characteristics. *Canadian Journal of Experimental Psychology*, 51(4) :307.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput Biol*, 10(1) :e1003412.

Schwartz, J. & Tallal, P. (1980). Rate of acoustic change may underlie hemispheric specialization for speech perception. *Science*, 207(4437) :1380–1381.

Schwartz, O. & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature neuroscience*, 4(8) :819–825.

Schyns, P. G., Bonnar, L., & Gosselin, F. (2002). Show me the features ! understanding recognition from the use of visual information. *Psychological Science*, 13(5) :402–409.

Scurto, H., Lemaitre, G., Francoise, J., Voisin, F., Bevilacqua, F., & Susini, P. (2015). Combining gestures and vocalizations to imitate sounds. *The Journal of the Acoustical Society of America*, 138(3) :1780–1780.

Seneff, S. (1988). A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, 16 :55–76.

Seymour, K. J., Scott McDonald, J., & Clifford, C. W. (2009). Failure of colour and contrast polarity identification at threshold for detection of motion and global form. *Vision Res*, 49(12) :1592–8.

Shamma, S. (2001). On the role of space and time in auditory processing. *Trends in cognitive sciences*, 5(8) :340–348.

Shamma, S. A. (1985a). Speech processing in the auditory system i : The representation of speech sounds in the responses of the auditory nerve. *The Journal of the Acoustical Society of America*, 78(5) :1612–1621.

Shamma, S. A. (1985b). Speech processing in the auditory system ii : Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *The Journal of the Acoustical Society of America*, 78(5) :1622–1632.

Shamma, S. A., Chadwick, R. S., Wilbur, W. J., Morrish, K. A., & Rinzel, J. (1986). A biophysical model of cochlear processing : Intensity dependence of pure tone responses. *The Journal of the Acoustical Society of America*, 80(1) :133–145.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234) :303–304.

Simoncelli, E. P. & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1) :1193–1216.

Simpson, A. J., Terrell, M. J., & Reiss, J. D. (2013). A practical step-by-step guide to the time-varying loudness model of Moore, Glasberg, and Baer (1997; 2002). In *Audio Engineering Society Convention 134*. Audio Engineering Society.

Singh, N. C. & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, 114(6) :3394–3411.

Sivaram, G. S., Nemala, S. K., Elhilali, M., Tran, T. D., & Hermansky, H. (2010). Sparse coding for speech recognition. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4346–4349. IEEE.

Slaney, M. (1998). Auditory toolbox. *Interval Research Corporation, Tech. Rep*, 10 :1998.

Slaney, M. & Lyon, R. F. (1991). Apple hearing demo reel. *Cupertino, CA : Apple Computer, Inc.*

Smith, E. C. & Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, 439(7079) :978–82.

Smith, J. E. K. (1982). Simple algorithms for m-alternative forced-choice calculations. *Attention, Perception, & Psychophysics*, 31(1) :95–96.

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychol Sci*, 16(3) :184–9.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416(6876) :87–90.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., & Formisano, E. (2009). Sound categories are represented as distributed patterns in the human auditory cortex. *Curr Biol*, 19(6) :498–502.

Stanislaw, H. & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1) :137–149.

Subramaniam, S., Biederman, I., & Madigan, S. (2000). Accurate identification but no priming and chance recognition memory for pictures in rsvp sequences. *Visual Cognition*, 7(4) :511–535.

Suen, C. Y. & Beddoes, M. P. (1972). Discrimination of vowel sounds of very short duration. *Perception & Psychophysics*, 11(6) :417–419.

Suga, N. (1992). Philosophy and stimulus design for neuroethology of complex-sound processing. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 336(1278) :423–428.

Suga, N., Zhang, Y., & Yan, J. (1997). Sharpening of frequency tuning by inhibition in the thalamic auditory nucleus of the mustached bat. *Journal of Neurophysiology*, 77(4) :2098–2114.

Suied, C., Agus, T. R., Thorpe, S. J., Mesgarani, N., & Pressnitzer, D. (2014). Auditory gist : recognition of very short sounds from timbre cues. *J Acoust Soc Am*, 135(3) :1380–91.

Suied, C., Agus, T. R., Thorpe, S. J., & Pressnitzer, D. (2013a). *Processing of short auditory stimuli : the rapid audio sequential presentation paradigm (RASP)*, pages 443–451. Springer.

Suied, C., Dremeau, A., Pressnitzer, D., & Daudet, L. (2013b). *Auditory sketches : Sparse representations of sounds based on perceptual models*, pages 154–170. Springer.

Suied, C., Susini, P., McAdams, S., & Patterson, R. D. (2010). Why are natural sounds detected faster than pips ? *J Acoust Soc Am*, 127(3) :EL105–10.

Susini, P., McAdams, S., Misdariis, N., Lemaitre, G., & Winsberg, S. (2005). Timbre des sons environnementaux. In *Proceedings of the Conference on Interdisciplinary Musicology (CIM05)*.

Susini, P., McAdams, S., & Winsberg, S. (1997). Caracterisation perceptive des bruits de vehicules. In *CFA : Congres Francais d'Acoustique*, pages 543–546.

Susini, P., McAdams, S., & Winsberg, S. (1999). A multidimensional technique for sound quality assessment. *Acta acustica united with Acustica*, 85(5) :650–656.

Susini, P., McAdams, S., Winsberg, S., Perry, I., Vieillard, S., & Rodet, X. (2004). Characterizing the sound quality of air-conditioning noise. *Applied Acoustics*, 65(8) :763–790.

Susini, P., Misdariis, N., Winsberg, S., & McAdams, S. (1998). Caracterisation perceptive de bruits. *Acoustique et Techniques*, 13(4) :11–15.

Susini, P., Perry, I., Vieillard, S., Winsberg, S., McAdams, S., & Rodet, X. (2001). Sensory evaluation of air-conditioning noise : Sound design and psychoacoutic evaluation. In *Proc. of the International Congress on Acoustics*.

Tallal, P. (1975). A different view of "auditory processing factors in language disorders". *Journal of Speech and Hearing Disorders*, 40(3) :413–414.

Tallal, P. (1980). Auditory temporal perception, phonics, and reading disabilities in children. *Brain and language*, 9(2) :182–198.

Tallal, P. (1984). Temporal or phonetic processing deficit in dyslexia ? That is the question. *Applied Psycholinguistics*, 5(02) :167–169.

Tallal, P. (2004). Improving language and literacy is a matter of time. *Nature Reviews Neuroscience*, 5(9) :721–728.

Tallal, P. & Piercy, M. (1973). Defects of non-verbal auditory perception in children with developmental aphasia.

Tallal, P. & Piercy, M. (1974). Developmental aphasia : Rate of auditory processing and selective impairment of consonant perception. *Neuropsychologia*, 12(1) :83–93.

Theunissen, F. E. & Elie, J. E. (2014). Neural processing of natural sounds. *Nat Rev Neurosci*, 15(6) :355–66.

Theunissen, F. E., Sen, K., & Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *The Journal of Neuroscience*, 20(6) :2315–2331.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *nature*, 381(6582) :520–522.

Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science*, 292(5515) :290–293.

Tremblay, S., Vachon, F., & Jones, D. M. (2005). Attentional and perceptual sources of the auditory attentional blink. *Perception & Psychophysics*, 67(2) :195–208.

Tsuchida, T. & Cottrell, G. W. (2012). Auditory saliency using natural statistics. In *Proc. Annual Meeting of the Cognitive Science (CogSci)*, pages 1048–1053.

Turmukhambetov, D., Campbell, N. D., Goldman, D. B., & Kautz, J. (2015). Interactive sketch-driven image synthesis. In *Computer Graphics Forum*, volume 34, pages 130–142. Wiley Online Library.

Vachon, F., Tremblay, S., Hughes, R. W., & Jones, D. M. (2010). Capturing and unmasking the mask in the auditory attentional blink. *Exp Psychol*, 57(5) :346–53.

Varnet, L., Knoblauch, K., Meunier, F., & Hoen, M. (2013). Using auditory classification images for the identification of fine acoustic cues used in speech perception. *Front Hum Neurosci*, 7 :865.

Venezia, J. H., Hickok, G., & Richards, V. M. (2016). Auditory "bubbles" : Efficient classification of the spectrotemporal modulations essential for speech intelligibility. *The Journal of the Acoustical Society of America*, 140(2) :1072–1088.

Vinje, W. E. & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456) :1273–1276.

Vitevitch, M. S. (2003). Change deafness : the inability to detect changes between two voices. *Journal of Experimental Psychology : Human Perception and Performance*, 29(2) :333.

Voss, R. F. & Clarke, J. (1975). 1/f noise in music and speech. *Nature*, 258 :317–318.

Voss, R. F. & Clarke, J. (1978). "1/f noise" in music : Music from 1/f noise. *The Journal of the Acoustical Society of America*, 63(1) :258–263.

Walker, K. M., Bizley, J. K., King, A. J., & Schnupp, J. W. (2011). Multiplexed and robust representations of sound features in auditory cortex. *The Journal of Neuroscience*, 31(41) :14565–14576.

Wang, D. (2005). *On ideal binary mask as the computational goal of auditory scene analysis*, pages 181–197. Springer.

Wang, D., Kjems, U., Pedersen, M. S., Boldt, J. B., & Lunner, T. (2008). Speech perception of noise with binary gains. *The Journal of the Acoustical Society of America*, 124(4) :2303–2307.

Wang, K. & Shamma, S. (1994). Self-normalization and noise-robustness in early auditory representations. *IEEE transactions on speech and audio processing*, 2(3) :421–435.

Warren, J. D. & Griffiths, T. D. (2003). Distinct mechanisms for processing spatial sequences and pitch sequences in the human auditory brain. *The journal of neuroscience*, 23(13) :5799–5804.

Warren, J. D., Scott, S. K., Price, C. J., & Griffiths, T. D. (2006). Human brain mechanisms for the early analysis of voices. *Neuroimage*, 31(3) :1389–97.

Wedin, L. & Goude, G. (1972). Dimension analysis of the perception of instrumental timbre. *Scandinavian Journal of Psychology*, 13(1) :228–240.

Wessel, D. L. (1973). Psychoacoustics and music : A report from michigan state university. *PACE : Bulletin of the Computer Arts Society*, 30 :1–2.

Wessel, D. L. (1979). Timbre space as a musical control structure. *Computer music journal*, pages 45–52.

Winsberg, S. & Carroll, J. D. (1989). A quasi-nonmetric method for multidimensional scaling via an extended euclidean model. *Psychometrika*, 54(2) :217–229.

Winsberg, S. & De Soete, G. (1993). A latent class approach to fitting the weighted euclidean model, clascal. *Psychometrika*, 58(2) :315–330.

Woods, D. L. & Alain, C. (1993). Feature processing during high-rate auditory selective attention. *Perception & psychophysics*, 53(4) :391–402.

Woods, D. L., Alain, C., Covarrubias, D., & Zaidel, O. (1993). Frequency-related differences in the speed of human auditory processing. *Hearing research*, 66(1) :46–52.

Woolley, S. M., Fremouw, T. E., Hsu, A., & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nat Neurosci*, 8(10) :1371–9.

Yang, X., Wang, K., & Shamma, S. A. (1992). Auditory representations of acoustic signals. *Information Theory, IEEE Transactions on*, 38(2) :824–839.

Yost, W. A. (2015). Psychoacoustics : A brief historical overview. *Acoustics Today : A Publication of the Acoustical Society of America*, 11(3) :46–53.

Young, M. P. & Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science*, 256(5061) :1327–1331.

Yund, E. W., Uno, A., & Woods, D. L. (1999). Preattentive control of serial auditory processing in dichotic listening. *Brain and language*, 66(3) :358–376.

Zatorre, R. J., Bouffard, M., & Belin, P. (2004). Sensitivity to auditory object features in human temporal neocortex. *The journal of neuroscience*, 24(14) :3637–3642.

Zhang, Y., Suga, N., & Yan, J. (1997). Corticofugal modulation of frequency processing in bat auditory system. *Nature*, 387(6636) :900–903.

Zwicker, E. & Scharf, B. (1965). A model of loudness summation. *Psychological review*, 72(1) :3.