


Caractérisation du code acoustique de l'attribut « chaud » pour des sons vocaux et non-vocaux par corrélation inverse

Master Acoustique, Traitement du signal, Informatique Appliqués à la Musique

Rapport de stage

28 juillet 2018

Auteur :
Pierre Rampon

Lieu du stage :
STMS Ircam-CNRS-SU, équipes Perception et Design Sonores & Analyse et Synthèse des sons

Encadrants :
Nicolas Misdariis
Nicolas Obin
Emmanuel Ponsot *(ENS-Ulm Laboratoire des Systèmes Perceptifs)*

Résumé

A la suite d'un travail lexical sur les attributs sonores en design sonore et d'études sur la reconnaissance spécifique des sons vocaux, ce stage rapporte une expérience de type 2-AFC utilisant la méthode de corrélation inverse, destinée à caractériser physiquement la perception de « chaleur » dans les sons vocaux et non-vocaux, chez les professionnels du son. Les liens de l'attribut avec l'enveloppe spectro-temporelle et le rapport harmoniques-sur-bruit (HNR) des sons sont testés. Les résultats montrent un lien entre le son « chaud » et une concentration d'énergie dans les basses-fréquences (60-400 Hz) pour les sons vocaux et non-vocaux. De plus, cette concentration d'énergie semble être localisée temporellement, dans la première moitié du son. Le « chaud » des sons vocaux comporterait une accentuation du HNR en basses fréquences (60-300 Hz) alors que celui des sons non-vocaux un HNR moins élevé en hautes fréquences (4000-5000 Hz). Cette première modélisation du son « chaud » devra être validée expérimentalement dans une étude ultérieure.

Mots-clefs : Design sonore ; voix ; corrélation inverse ; descripteurs acoustiques, enveloppe spectro-temporelle, rapport harmonique/bruit.

Abstract

In the wake of a lexical work on sound attributes in acoustical features and studies of specific recognition of vocal sounds, this report is about a 2-AFC experiment using the reverse correlation method. The experiment is designed to physically characterize the perception of a "warm" sound in vocal and non-vocal sounds among sound experts. The relationships between sound "warmth" and two dimensions are tested, precisely with the spectro-temporal envelope and the harmonic-to-noise ratio (HNR). It appears that a "warm" sound has a low-frequency energy concentration (60-400 Hz). This concentration seems to be temporally localized in the first half of the sound. Furthermore, it seems that there is a relationship with HNR that would be accentuated in low-frequency area for non-vocal sounds (60-300 Hz) but lowered in high frequencies for vocal sounds (4000-5000 Hz). This model of "warmth" should be confirmed by further studies.

Mots-clefs : Sound design ; voice ; reverse correlation ; acoustical features ; spectro-temporal envelope ; harmonic-to-noise ratio envelope.

Remerciements

Tout d'abord, je dois le bon déroulement de ce stage à mes encadrants : Nicolas Misdariis, Nicolas Obin et Emmanuel Ponsot. Je les remercie donc vivement pour tous leurs bons conseils et le temps qu'ils m'ont réservé. Ensuite, je remercie également toute l'équipe PDS pour leur bon accueil.

Je tiens également à remercier vivement les participants qui ont courageusement passé l'expérience de corrélation inverse, sans oublier les stagiaires de l'IRCAM qui se sont portés volontaires pour l'expérience préliminaire.

Table des matières

Chapitre 1

Introduction

Dans le domaine de l'audio, les spécialistes - ingénieurs du son, compositeurs, designers sonores etc. - utilisent un vocabulaire particulier pour décrire les sensations sonores. Le premier, Schaeffer [1] étudie et distingue les modes d'écoute des sons : on peut *ouïr*, *écouter*, *entendre* ou *comprendre* un son. Le premier mode d'écoute fait référence à une écoute purement passive, mais les trois derniers permettent de percevoir différents éléments présents dans le son. Carron [2], reprenant ses travaux, définit trois types de description d'un son, en parallèle aux trois derniers types d'écoute : On peut décrire un son par des concepts simples et de bas-niveau, ce qui fait référence au *discours réduit*. Par exemple, « j'entends un son long, aigu et fort ». C'est la description liée à l'action d'*écouter*. Ensuite, on peut décrire le son par sa source physique : « c'est un son de cloche frappée par le battant ». La description porte davantage sur le type de la source et sur les modes d'interaction (ici, le battant tient le rôle de l'excitateur). C'est le *discours causal*, lié à l'action d'*entendre*. Enfin, on peut associer la perception auditive à une émotion, à un concept plus qu'à un ressenti purement physique (long, fort) ou qu'à une cause physique (son de cloche). Pour suivre l'exemple, de ce même son aigu et fort qui est d'ailleurs un son de cloche, je peux dire : « c'est un son alarmant ». On pourrait aussi le qualifier de son désagréable, ou rond. Cette description relève du discours *sémantique*, liée à l'action de *comprendre*. Dans ce cas, on s'applique davantage à expliquer et comprendre en profondeur le phénomène sonore. On peut dire que l'écoute porte plus sur la forme que sur la matière de l'objet sonore ; on qualifie cette dernière description comme étant de haut-niveau.

1.1 Descripteurs haut-niveau

Lorsqu'on qualifie un son à un haut niveau sémantique (c'est-à-dire suivant la description sémantique ou l'action de *comprendre*), on peut reprocher un manque de clarté et de concision dans les termes employés. En effet, les représentations physiques de ces

mots ne sont pas immédiates, au contraire du discours réduit par exemple (qui est un discours bas-niveau), où les termes peuvent être reliés à des propriétés physiques assez simples : l'intensité du son se mesure par rapport à l'énergie du signal, la hauteur d'un son harmonique est liée aux harmoniques du spectre qui se calcule avec l'analyse de Fourier. Pour les qualificatifs haut-niveau, une question intéressante est de savoir s'il existe un rapport entre la sémantique et la physique. Autrement dit, si la notion à haut niveau sémantique est physiquement quantifiable, mesurable. On voit une autre difficulté émerger : dès lors qu'elle est perçue par des sujets, la connaissance de la notion est personnelle mais comme tout mot du langage, elle désigne un concept qui tend à être commun à plusieurs individus. On peut alors se demander si la notion haut-niveau est réellement commune entre les sujets.

M. Carron a élaboré pendant son travail de thèse [3] un vocabulaire d'attributs sonores les plus utilisés par les professionnels du son (cf figure 1.1). La liste exhaustive se trouve en annexe (cf figure A.1). Elle a été construite de manière inductive : l'auteur a recoupé 52 travaux de description verbale de sons. Il a ensuite sélectionné 46 mots les plus récurrents. Puis il a questionné des praticiens du son sur cette liste pour arriver à cette liste de 35 mots, correspondant aux attributs sonores les plus usités.

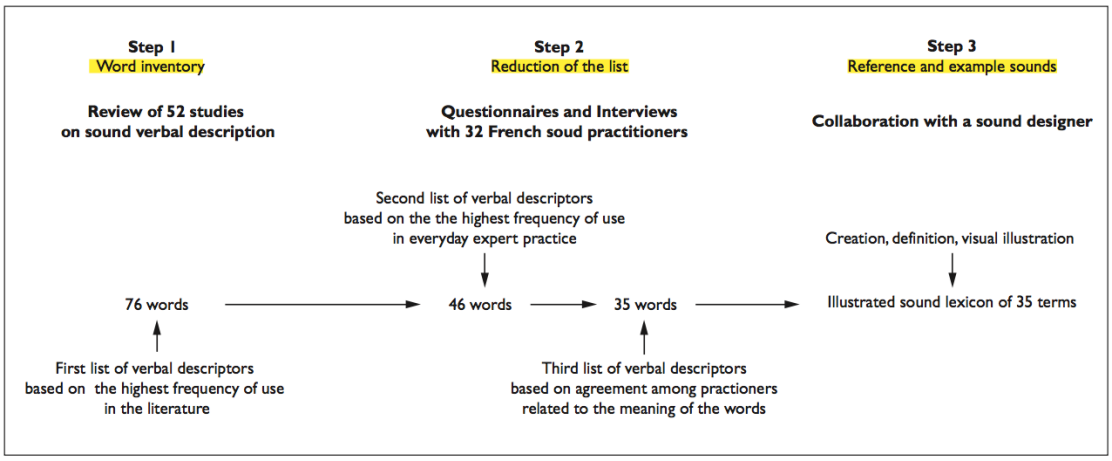


FIGURE 1.1 – Méthodologie pour la génération du lexique sonore de Maxime Carron [3]

Parmi ces qualificatifs, le terme « chaud » se place en tête de la liste triée par ordre d'occurence décroissant. Il a peu été étudié pour l'instant. Sabin a réalisé une expérience [4] destinée à caractériser physiquement les attributs sonores « chaud », « noir », « métallique » et « brillant ». Pour cela, il a construit 75 courbes d'égalisation générées aléatoirement. Il s'est assuré, par une construction statistique, qu'elles étaient suffisamment dissemblables entre elles. Ces courbes ont ensuite servi de filtres à deux fichiers audio (un son de voix féminine et un son de percussion). Ensuite, il a fait écouter les sons obtenus à huit participants qui ont dû juger de la correspondance du *stimulus* écouté avec les

Concept	Définition	Exemple
Mat/Résonant	Un son mat est un son qui s'arrête sèchement sans qu'aucun phénomène de résonance ne vienne le prolonger dans le temps. A l'inverse, un son résonant verra une partie de son énergie s'étaler dans le temps.	Exemples de sons environnementaux mats et résonants (porte, verre etc.).
Chaud	Un son chaud est un son plutôt grave qui donne une impression de largeur, de déploiement et d'une certaine richesse qui le rend plutôt agréable.	Exemples de voix d'homme, d'un bruit de moteur.

TABLE 1.1 – Exemple de définition du couple d'attributs « mat/résonant » et de l'attribut « chaud ». A ces définitions l'auteur joint des exemples sonores. [3]

cinq adjectifs, *via* un slider à 5 positions (allant de « très opposé à « très ressemblant »). C'est à partir des réponses aux 75 essais différents qu'il a tracé les courbes d'égalisation du « chaud » sonore pour les deux sources audio (le son de voix et celui de percussion, cf figure 1.2).

1.2 Sens de l'attribut sonore « chaud » pour les sons vocaux et de design sonore

D'autre part, on peut se demander s'il n'y a pas deux concepts compris dans l'idée de « son chaud ». En effet, lorsque l'on écoute une voix sans pour autant voir le locuteur, on peut percevoir certains traits psychologiques comme sociaux que l'on associe non pas au son que l'on entend mais à la personne qui parle : la personnalité (dont l'âge, le sexe), les émotions (la joie, la tristesse, la peur, mais aussi la confiance, la dominance) que le locuteur laisse transparaître dans sa voix [5, 6, 7]. Ainsi, lors de l'écoute d'un son de voix, la qualification du son par l'attribut « chaud » peut désigner : 1) la chaleur de la personne, c'est-à-dire le fait qu'elle est chaleureuse ; 2) la chaleur du son lui-même, comme l'entendent les experts travaillant dans le design sonore, sans regarder la personnalité que l'on infère à partir de la voix.

Par ailleurs, plusieurs travaux [8, 9] montrent que le cerveau humain ne traite pas de la même manière les sons humains des autres *stimuli* auditifs. Tout d'abord, Belin a identifié une zone cérébrale sélective aux sons vocaux à partir d'images par résonance magnétique (IRM). A la suite de cette publication, Agus a mené plusieurs expériences où il a comparé la performance des participants à reconnaître et analyser des sons vocaux

et instrumentaux. Il trouve de meilleurs résultats lorsque le son est vocal. Enfin, Isnard a conclu de son travail de thèse que le système auditif humain reconnaît plus efficacement les sons naturels et, parmi les sons naturels, les sons vocaux sont parmi ceux les mieux reconnus.

Enfin, les résultats de l'expérience de Sabin [4] montrent que le terme "warm" n'est pas représenté identiquement pour tous les participants, selon que le terme est caractérisé pour un son vocal ou instrumental (de percussion, cf figure 1.2).

Ces résultats mènent à penser que l'attribut sonore « chaud » peut être représenté différemment selon que le son est vocal ou non-vocal.

1.3 Problématique

De cette discussion, on voit que des attributs sonores haut-niveau sont couramment employés par les personnes travaillant dans le domaine du son. Elles-mêmes remarquent que chacun construit son propre vocabulaire par l'expérience et que pourtant elles arrivent à s'entendre sur ces attributs sonores abstraits. Le qualificatif « chaud » rentre dans cette catégorie. Parmi les chercheurs qui ont commencé à travailler sur cette notion, Sabin a tenté de caractériser cet attribut par une expérience de psychoacoustique [4]. Il a réalisé cette expérience dans le but de construire un outil simple d'utilisation, permettant de manipuler des sons suivant des attributs comme "warm" (chaud) ou "tinny" (métallique). Il obtient des égaliseurs donnant pour chaque participant de l'expérience leur représentation mentale des descripteurs audio (cf figure 1.2).

Le lien perceptif/physique pour la « chaleur » sonore n'est donc pas encore bien défini. L'expérience va se poser les questions suivantes : 1) est-ce que le qualificatif « chaud » peut être mesurable physiquement, comme la brillance d'un son se caractérise par le centre de gravité spectral du son ? [10, 11] 2) Dans ce cas, est-ce que la caractéristique physique d'un son « chaud » est commune entre les personnes utilisant ce vocabulaire ? 3) Est-ce que le sens du « chaud » sonore est différent selon que le son est vocal ou non ?

Ce sont ces trois questions qui vont porter tout ce travail expérimental. On peut les reformuler de la manière suivante :

1) La représentation mentale de « chaleur sonore » est-elle physiquement caractérisable par la méthode de corrélation inverse ?
2) Existe-t-il une représentation commune entre différentes personnes ?
3) Existe-t-il une représentation commune entre les sons vocaux et les sons non-vocaux ?

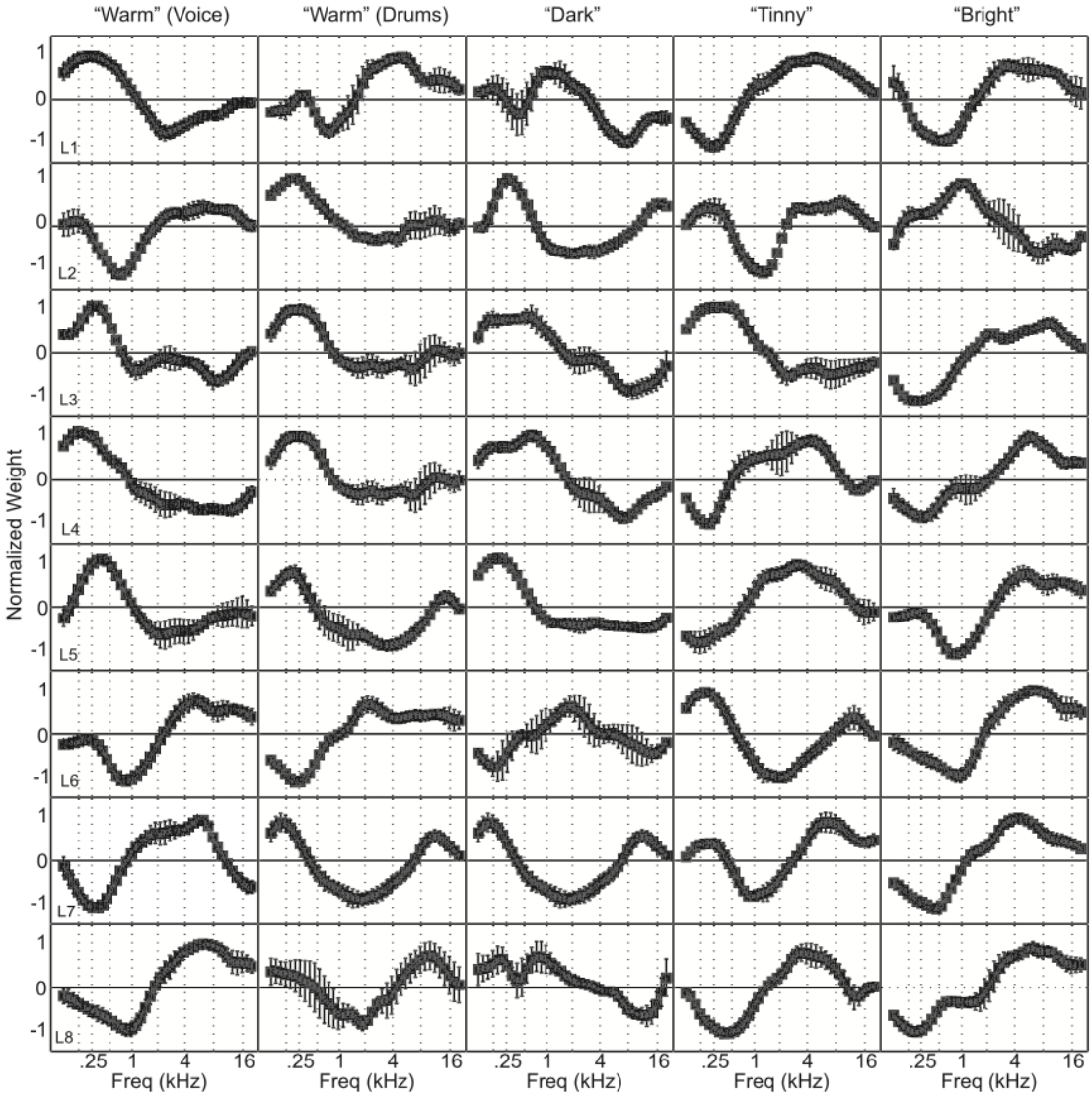


FIGURE 1.2 – Egaliseurs à bandes fréquentielles par participant pour les différents descripteurs obtenus par Sabin [4]. Les deux premières colonnes correspondent au descripteur « chaud » dans un son vocal et de percussion. Les barres sont les erreurs type de la moyenne.

1.4 Hypothèses sur les sons « chauds »

Pour répondre à ces questions, la première chose est de définir les hypothèses sur les dépendances d'un son « chaud » vis-à-vis de ses propriétés physiques : est-ce qu'un son « chaud » est un son inharmonique ? un son riche en hautes-fréquences ? ou bien

est-ce que la réverbération joue un rôle ? Le travail de Carron fournit potentiellement des éléments de réponse. Plus qu'un simple listage de descripteurs audio de haut-niveau, il aboutit à l'énoncé d'une définition pour chacun des termes (deux exemples sont reportés dans la table 1.1). La définition se veut compréhensible par tous, on ne trouve pas de terme technique dans ces définitions (même si parfois on les devine sans trop de difficultés).

son plutôt **grave**	→	basses fréquences
qui donne une impression de **largeur**		
et d'une **certaine richesse**	→	présence de bruit
qui le rend plutôt agréable.		

TABLE 1.2 – Transposition de la définition de « chaud » en termes de traitement du signal.

Dans la continuité du travail de thèse de M. Carron, l'équipe PDS a organisé des séances de travail (*workshops*) entre les chercheurs de l'équipe, un ingénieur du son et un compositeur en octobre 2017. A partir des définitions données par M. Carron, ils ont formulé des hypothèses précises sur ce lien verbal/signal. Pour l'attribut « chaud », le côté « grave » est traduit par l'hypothèse d'un spectre riche en basses-fréquences (BF). La « richesse » serait liée à un aspect bruité dans le son (cf table 1.2).

1.5 Dimensions manipulées

L'objectif est d'arriver à faire varier la « chaleur » d'un son en manipulant certaines dimensions physiques, puis ensuite, de faire ressortir le lien entre les dimensions manipulées et l'attribut « chaud ». Avec les hypothèses formulées plus haut, quelle(s) dimension(s) choisir ? Pour vérifier l'hypothèse portant sur le côté *grave*, la première dimension est logiquement le spectre. On manipulera cette dimension par un égaliseur (EQ, *equalizer*). Comme le remarque Sabin, l'EQ a l'avantage d'être un outil extrêmement répandu ; de plus il permet de changer le timbre. Là où Sabin a choisi un EQ statique, ici on prendra un EQ dynamique, c'est-à-dire variant aussi dans le temps.

D'autre part, le rapport harmoniques-sur-bruit (HNR, *Harmonic-to-Noise Ratio*) est choisi pour qualifier l'aspect *bruité* des signaux [12]. C'est en effet un descripteur du bruit pour les sons vocaux ou harmoniques. Ce descripteur se rapproche de certains descripteurs que l'on trouve dans les listes de descripteurs audio usuels en traitement du signal [13] et a l'avantage de mesurer la quantité de bruit dans plusieurs bandes de temps et de fréquences. On peut donc contrôler le bruit de manière plus fine. Cette dimension est une nouveauté par rapport à l'expérience menée par Sabin [4] mais s'appuie sur les récents *workshops* de l'équipe PDS.

Les hypothèses puis les dimensions utilisées maintenant détaillées, il reste à choisir la méthode expérimentale.

1.6 Méthode de corrélation inverse

On cherche une méthode expérimentale qui présente à des participants des *stimuli* audio qu'ils jugent selon le critère demandé et qui, en sortie, donne une image représentative de ce que les participants disent être un son « chaud » dans les dimensions choisies. Finalement, le problème est similaire à celui d'E. Ponsot lorsqu'il a voulu décrire comment on se représente une voix interrogative [14] en terme de variation de hauteur, ou bien une voix souriante en terme de déplacement fréquentiel des formants [15]. Il a utilisé une méthode dite de corrélation inverse (ou *reverse correlation*) qui lui a donné pour l'exemple de la voix interrogative la variation moyenne de hauteur d'une voix interrogative au cours du temps (cf figure 1.5), soit en quelque sorte le filtre moyen en hauteur d'une voix interrogative.

C'est cette même méthode expérimentale qui sera employée.

> « L'idée générale de la corrélation inverse est de présenter à un système (ici, un observateur humain) un *stimulus* légèrement perturbé sur beaucoup d'essais. [...] Les *stimuli* perturbés vont, sur différents essais, mener à des réponses différentes de la part du système, et les outils de corrélation inverse seront utilisés pour inférer les propriétés fonctionnelles du système à partir du schéma du bruit ajouté au *stimulus* et des réponses associées. » [15]

Le principe de corrélation inverse a d'abord été utilisé en psychophysique pour caractériser le processus sensitif humain (par exemple, la détection de tons dans du bruit [16]) puis en vision :

> « En vision, la corrélation inverse a été appliquée pour extraire la représentation mentale des observateurs qui fait, par exemple, un visage joyeux (Mangini and Biederman, 2004), comment les expressions faciales diffèrent selon les cultures (Jack et al., 2004) ou même ce qui fait que Mona Lisa semble sourire (Kontsevich and Tyler, 2004). » [15]

Egalement entre sciences sociales et traitement de l'image, une autre expérience (cf figures 1.3 et 1.4) vise à étudier les traits proprement féminins d'un visage. Enfin, cette méthode statistique a été introduite en audio :

> « Quelques récentes études ont commencé à utiliser cette approche pour des tâches auditives comme l'intelligibilité de la parole (Varnet et al., 2016 ; Venezia et al., 2016) ou comme la reconnaissance d'instruments musicaux (Thoret et al., 2016). En particulier, Brimijoin et al. (2013) a utilisé la corrélation inverse pour découvrir la représentation interne d'une voyelle chuchotée en présentant à des observateurs humains des voyelles bruitées selon une variation aléatoire du spectre de bruit. » [15]

Plus récemment encore, deux expériences en psychoacoustique ont utilisé cet outil [15, 14]. La première a extrait le code spectral d'une son de voyelle « souriant » en mo-

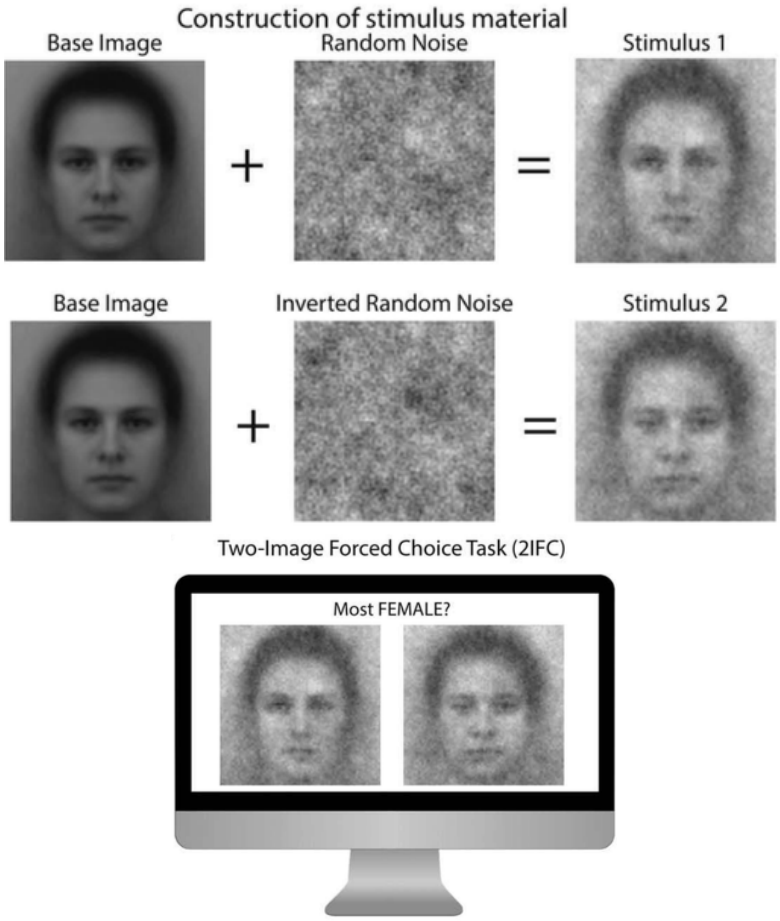


FIGURE 1.3 – Exemple d'expérience de *reverse correlation* en vision : les visages sont floutés alétoirement ; la tâche est de désigner quel visage parmi la paire présentée est le plus féminin. On présente un grand nombre de paires pour obtenir un résultat moyen en fonction des réponses[17].

difiant le contenu spectral d'un enregistrement de voyelle par le biais d'un égaliseur à bandes. La seconde a étudié les attributs « digne de confiance » et « dominant » d'un enregistrement de voix (prononçant le mot "hello") en faisant varier aléatoirement dans cet enregistrement la hauteur de voix pour mimer différentes prosodies.

Cette technique nécessite de bruiter un *stimulus* (un son pour le cas présent) dans une ou plusieurs dimensions, puis de le soumettre au jugement de personnes. On répète l'opération un certain nombre de fois. L'ajout d'un bruit tiré aléatoirement permet d'explorer la dimension précisée dans toute son étendue, les répétitions de mettre en évidence une tendance générale statistique (cf figure 1.5).

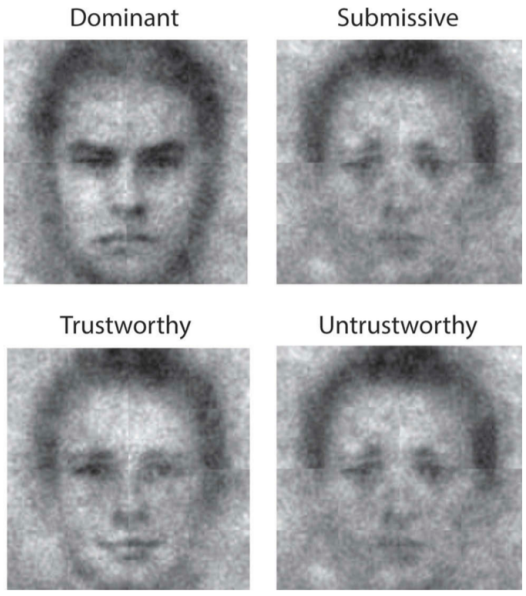


FIGURE 1.4 – Résultat de l'expérience en vision : on obtient des filtres-moyens d'un visage dominant ou soumis (en haut), digne de confiance ou peu fiable (en bas) [17].

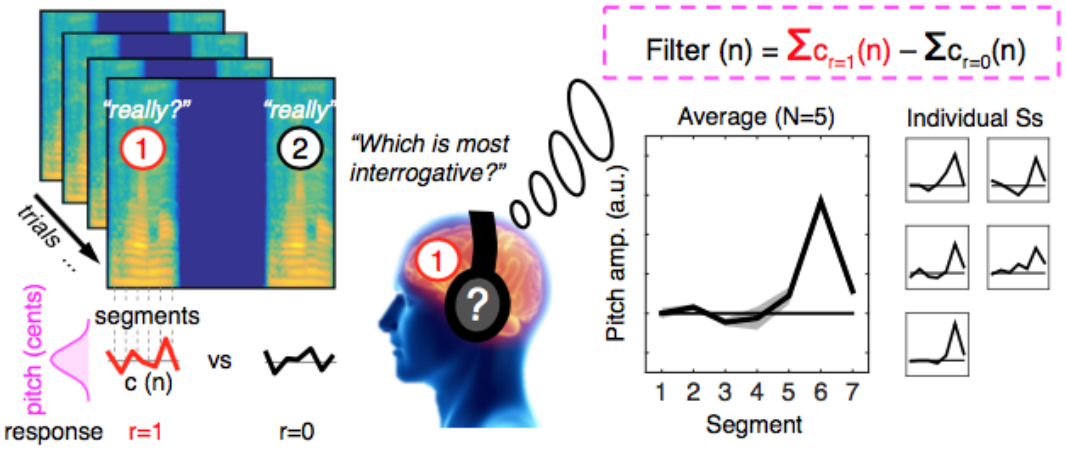


FIGURE 1.5 – Expérience de corrélation inverse en psychoacoustique [14] : on présente des *stimuli* audio d'une voix disant "really ?" sous forme de paires. On corrèle les réponses du participant (à la question "which is most interrogative ?") avec les modifications de hauteur ("pitch") effectuées aléatoirement sur toutes les paires. Le filtre-moyen de contour de hauteur donnant une voix interrogative est tracé sur le graphe de droite.

Chapitre 2

Présentation de l'expérience

La problématique générale porte sur la notion de « chaleur sonore » : est-elle liée à des propriétés physiques, comme l'interrogativité de la voix est liée à une augmentation de hauteur en fin de phrase ? Est-elle commune aux personnes utilisant cet attribut sonore ? Est-elle différente pour les sons vocaux et non-vocaux ?

Pour tenter de répondre à ces questions, on met en place une expérience de 2-AFC (de choix forcé à 2 alternatives, *two-Alternative Forced Choice*), utilisant la méthode de corrélation inverse. Pour une telle expérience, on construit des *stimuli* que l'on présente ensuite aux participants. Ces *stimuli* sont bruités de manière aléatoire, dans des dimensions précises, qui dépendent des hypothèses de départ. Ici, on va tester la dépendance de l'attribut « chaud » dans deux dimensions : le contenu spectral (bruité par un égaliseur à bandes fréquentielles et temporelles, suite à l'hypothèse qu'un son « chaud » est *grave*) et le rapport harmoniques sur bruit (suite à la supposition de l'aspect *bruité*). Ce chapitre va détailler d'abord la construction des *stimuli* puis la réalisation de l'expérience.

2.1 Génération du corpus de sons

Pour répondre à la question sur la différence de concept entre les sons vocaux et non-vocaux, le corpus initial est composé de sons vocaux monosyllabiques, sons vocaux bisyllabiques et sons non-vocaux (cf figure 2.1). L'hypothèse est que pour les sons de parole suffisamment courts (sons monosyllabiques), il n'y a pas d'inférence sur les qualités de la personne qui parle alors que pour les sons plus longs (sons bisyllabiques), on commence à inférer sur la personnalité (les termes « suffisamment courts » et « plus longs » seront précisés par la suite).

Premier sous-corpus : sons vocaux monosyllabiques Le son vocal et court le plus élémentaire est un son de voyelle (un phonème), d'où ce choix de sons vocaux monosyllabiques (le terme de son monosyllabique désigne donc plus précisément un phonème).

De plus, ce seront uniquement des sons issus des voyelles [a], [i] et [ou]. Le choix de ces trois voyelles se justifie par la modélisation de C. F. Hellwag [18]. Ce dernier a introduit la notion de triangle vocalique. Il s'agit d'une représentation qui place les trois voyelles [a], [i] et [ou] sur les sommets d'un triangle. En considérant la formation des sons, Hellwag classe toutes les voyelles à l'intérieur de ce triangle. Cela permet de les décomposer dans cette « base des voyelles » selon la fréquence des deux premiers formants. En prenant les voyelles qui constituent cette base, on travaille sur les éléments qui permettent à eux seuls de reconstruire n'importe quelle syllabe.

Deuxième sous-corpus : sons vocaux bisyllabiques Pour tester l'hypothèse d'inférence sur le trait psychologique sous-tendu au son vocal, le deuxième sous-corpus de sons est composé de sons vocaux mais cette fois ressemblant davantage à de la parole. Ces mots sont constitués de deux voyelles, chacune précédée par une consonne (modèle le plus simple de bisyllabe). On prend un mot qui n'a pas de signification pour les observateurs (autrement, le jugement pourrait être influencé par l'interprétation de ce mot par l'observateur, ce qui introduirait un biais indésirable). La consonne est choisie selon un critère d'ordre fréquenciel : on souhaite pouvoir distinguer facilement la consonne de la voyelle. D'où le choix de la consonne occlusive (ou explosive) [b] qui a les propriétés d'une impulsion (concentration temporelle et son spectre contient de l'énergie dans une plage continue de fréquences, contrairement aux voyelles qui sont des signaux harmoniques). On introduit donc les sons bisyllabiques : [ba-ba], [bi-bi], [bou-bou], en parallèle aux [a], [i] et [ou].

Troisième sous-corpus : sons non-vocaux Il reste à traiter des sons non-vocaux. Leur analyse permettra d'obtenir le code acoustique de l'attribut pris dans son sens sonore et non psychologique. Toujours dans l'optique de garder une certaine homogénéité entre ces trois sous-corpus, les sons sont générés par un modèle physique source-filtre. Les sons vocaux monosyllabiques forment les excitateurs et les modèles physiques d'instruments, les résonateurs (plaques, membranes, cordes par exemple). L'effet de cette synthèse par modèle physique est d'avoir des sons reconnus comme non-vocaux, mais dont les propriétés harmoniques sont proches de celles des enregistrements de voix originaux, donc des autres sous-corpus.

Ces trois sous-corpus de sons ont été construits en plusieurs étapes :

- Enregistrement des voix
- Modification des timbres de voix et passage dans un modèle de résonateur physique avec *Modalys* [19]
- Sélection à partir des résultats d'une expérience de catégorisation
- Normalisation du corpus en HNR, en contenu spectro-temporel et en volume sonore.

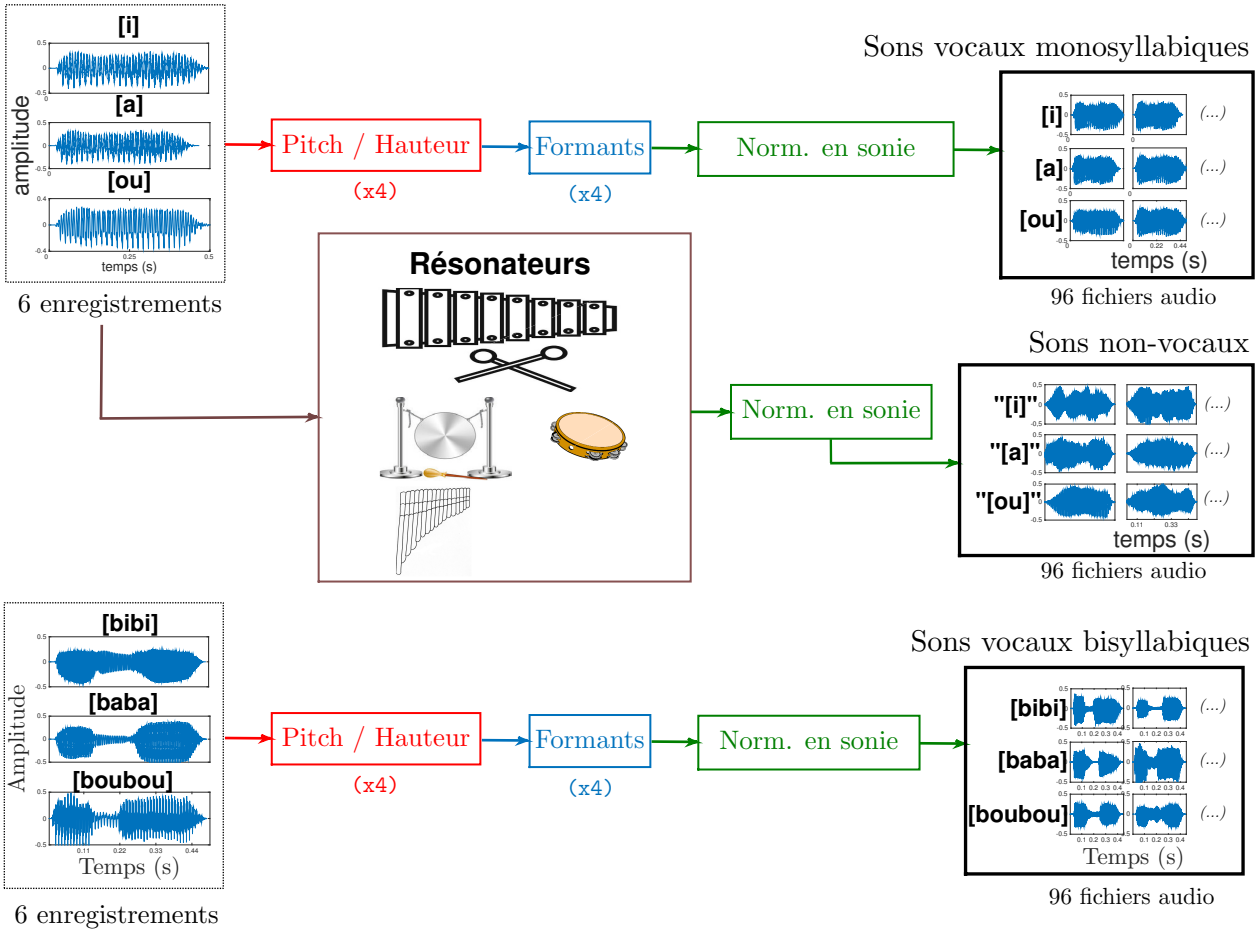


FIGURE 2.1 – Schéma décrivant les deux premières étapes de la construction du corpus : après enregistrement, sélection, « nettoyage », formatage (normalisations en durée, intensité, ajouts de fondus et de silences), chaque fichier audio subit un certain nombre de transformations pour obtenir un corpus de 3x96=288 enregistrements différents.

2.1.1 Enregistrement des sons vocaux

Cinq voix d'hommes et deux de femmes prononçant successivement et par trois fois « a », « i », « ou », « ba-ba », « bi-bi », « bou-bou » sont enregistrées. La consigne précisait de prendre un ton neutre et de faire durer chacun des sons environ 500ms. Pour aider les participants à respecter les consignes, ils ont préalablement écouté un enregistrement servant de modèle. Les voix ont été enregistrées dans les cabines de l'équipe PDS [1]. La prise de son s'est faite à partir de deux micros [2]. Sur ces prises de son, deux enregistrements par voyelle (par ex. [a]) et mot (par ex. [bi-bi]) sont retenus, en prenant une voix masculine et une féminine pour chaque, indépendamment du micro d'enregistrement. Puis on applique un fondu en ouverture de 10 ms et en sortie de 30 ms, ainsi qu'un silence de 5 ms au début et à la fin de l'extrait. Enfin, les sons sont normalisés par la moyenne quadratique du signal.

A ce stade, on a 12 enregistrements, soit 6 pour chacun des sous-corpus de sons monosyllabiques et bisyllabiques.

2.1.2 Agrandissement du corpus

Lors de l'expérience de Sabin (cf section 1), les *stimuli* sont issus de plusieurs filtres aléatoires mais que d'une seule source sonore. On ne sait donc pas si les résultats qu'il trouve (cf figure 1.2) s'appliquent aux sons vocaux en général ou uniquement à cet enregistrement de voix féminine. Dans un objectif de généralisation des futurs résultats, il est donc souhaitable d'avoir une estimation qui ne dépende pas d'un son en particulier. C'est pourquoi une centaine de sons composera chaque sous-corpus (cf figure 2.1). Ainsi, les spécificités spectrales de chaque corpus seront gommées du fait qu'en moyenne, les corpus de sons auront les mêmes propriétés.

Constitution des sous-corpus de sons vocaux mono- et bisyllabiques

A partir des 12 enregistrements audio, on génère 192 sons (soit 96 sons monosyllabiques et autant de bisyllabiques). Pour ce faire, on transpose la hauteur de la voix aléatoirement (tirage uniforme dans l'intervalle $[-130; -30[\cup]30; 130]$ cents) et modifie les huit premiers formants grâce à l'outil `creamDBgenerator` [20], afin de changer le timbre de la voix enregistrée et de générer seize voix de synthèse à partir d'un seul enregistrement. Les paramètres de l'aléatoire de la transformation des formants sont : `d_std=250 cents`

1. Cabines audiométriques de standard *I.A.C.*.
2. Un premier micro est omnidirectionnel, référencé *SCHOEPS CMC5–U*, placé à 1/2 mètre de la chaise où était assis le locuteur, derrière un filtre anti-pop. Le second micro employé est de type *DPA d :fine Earset Omni Tan*. La carte son utilisée est *Fireface UCX* ; l'enregistrement s'est fait avec le logiciel Audacity. La fréquence d'échantillonnage est de 44100Hz et les fichiers audio sont écrits en 16 bits au format WAV.

(*std* pour *standard deviation*, écart-type) ; `str_mode = 'formants'`. Et ce, pour les sons monosyllabiques comme pour les bisyllabiques.

Constitution du sous-corpus de sons non-vocaux

Le logiciel *Modalys* permet, entre autres choses, de générer des sons à partir d'un modèle source-filtre physique. Pour constituer ce corpus de sons non-vocaux, les résonateurs sont : des modèles de membranes circulaires ou rectangulaires, de tubes aux conditions aux limites ouvert/ouvert, ouvert/fermé ou fermé/fermé, de barres rectangulaires, de plaques circulaires aux conditions aux limites encastré ou libre, ou encore de cordes. Les enregistrements de voyelles monosyllabiques remplacent les excitateurs physiques plus « classiques » tels le plectre, l'archet, ou la simple impulsion. Il en résulte un son qui, harmoniquement, a les propriétés des sons vocaux, mais auquel on ajoute, par ce filtrage, une sonorité instrumentale. Les paramètres physiques des résonateurs (fréquence de résonance, constantes d'amortissement, module d'Young etc.) sont configurés de manière à effacer le caractère vocal du son de synthèse. Dans ce même objectif, les enregistrements subissent un second fondu en ouverture linéaire de 70ms pour réduire l'attaque du son vocal. On génère de la sorte 96 sons de synthèse à partir des 6 enregistrements monosyllabiques de départ (cf annexe B.2).

Traitements communs aux différents sous-corpus

Tous les sons comportent un silence de 5ms au début et à la fin de la piste audio. La durée totale de chaque piste audio est de 500ms. L'étirement temporel pour arriver à cette durée se fait *via AudioSculpt* dès les enregistrements initiaux pour les deux premiers sous-corpus, lors de la synthèse *Modalys* pour le dernier sous-corpus. Enfin, les sons sont normalisés en sonie avec l'algorithme *R-128* [21]. A ce stade, le corpus global est composé de 288 sons.

2.1.3 Expérience préliminaire de catégorisation

Objectif et procédure

Afin de s'assurer que les sons vocaux et non-vocaux ainsi constitués seront bien reconnus comme tels lors de l'expérience de corrélation inverse, le corpus est testé par une expérience d'écoute de type 2-AFC sur 12 participants [3]. Plus précisément, ils ont répondu par « oui » ou « non » à la question : « le son entendu est-il un son vocal ? » après l'écoute de chaque son de la base [4].

Après cette pré-expérience de catégorisation, les sons trop ambigus sont écartés du corpus, c'est-à-dire les sons vocaux trop transformés (par l'étape d'agrandissement du corpus, par ex.) et les sons non-vocaux encore perçus comme vocaux.

Résultats

12 personnes ont passé la pré-expérience de catégorisation. Ces participants ont rejeté en moyenne $3.6\% \pm 4.2\%$ des sons non-vocaux, et toujours moins de 16%. Quant aux sons vocaux (comprenant monosyllabiques et bisyllabique), les rejets sont de $12.9\% \pm 8.1\%$ (cf figure 2.2).

Les trois participants ayant un taux de rejet de sons vocaux supérieur à 16% (pourcentage maximum de rejets des sons non-vocaux) sont écartés de l'analyse, leur jugement étant alors considéré comme trop sévère. Ainsi faisant, le taux de rejet pour les non-vocaux descend à $9.4\% \pm 4.2\%$.

Le critère d'acceptation des sons pour le corpus final est le suivant : les sons vocaux doivent être reconnus par la majorité des participants, soit par au moins cinq d'entre eux (sur les neuf considérés). Pour les sons non-vocaux, seuls ceux qui ont été bien reconnus comme tels par tous les participants sont gardés. Puis les trois sous-corpus ont été égalisés de façon à garder le même nombre de voyelles dans chaque sous-corpus et au maximum la parité voix masculines - féminines. Cette sélection est manuelle. 48 sons de chaque classe (monosyllabique, bisyllabique, de synthèse) sont finalement gardés, pour un total de 144 fichiers audio. L'équirépartition entre les voyelles [a], [i], [ou] et équivalents est conservée, ainsi que l'égalité homme/femme pour les classes de sons vocaux sauf pour le sous-corpus [ou], où la répartition est de 13 voix masculines et 3 féminines à cause des rejets des autres voix féminines de cette sous-classe.

3. L'expérience s'est déroulée dans les mêmes cabines insonorisées de l'équipe PDS ; le logiciel servant d'interface est « PsyExp ». L'écoute s'est fait au casque de modèle *Sennheiser HD650*. Le volume sonore imposé à tous les participants a été mesuré à 68,5 dBA pour un bruit blanc de fréquence entre 0 et 2 200 Hz, d'amplitude 0,5 et généré par *Audacity*. Il a été mesuré par un sonomètre Brüel & Kjaer BZ-5503.

4. La consigne se trouve en annexe B.1.

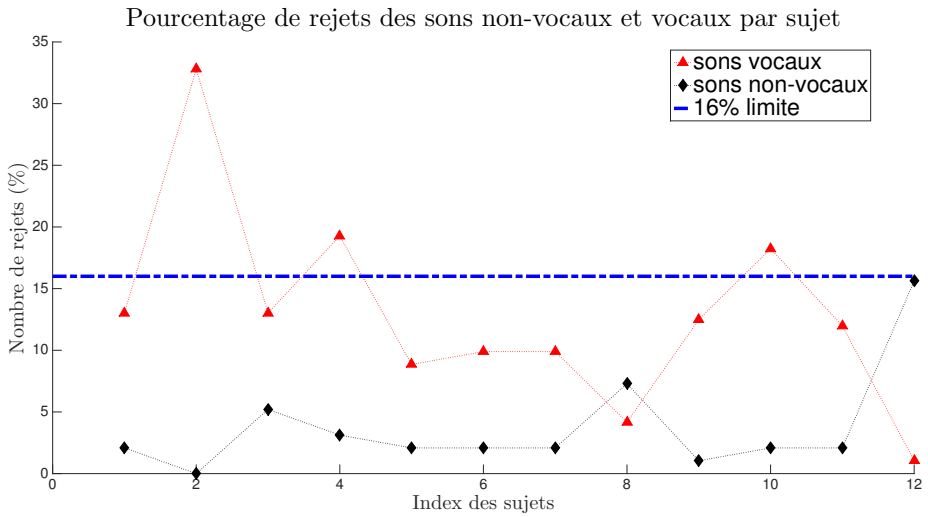


FIGURE 2.2 – Les trois participants rejetant plus de 16% (pourcentage maximum de rejets des sons non-vocaux) des sons vocaux sont écartés.

Ajustement du sous-corpus de sons de synthèse

Après cette réduction du corpus de 288 à 144 sons, il apparaît que les sons de synthèse ont un contenu fréquentiel trop faible en hautes fréquences (HF). En effet, la tâche de normalisation des sons qui est le dernier traitement avant passage dans la fonction de corrélation inverse (cf section 2.2) a pour effet de diminuer le gain global, et donc de supprimer les composantes spectrales trop faibles. Or les résonateurs ont la propriété d'avoir des modes de résonance de plus en plus faibles en énergie. Il en résulte qu'après cette baisse de niveau, les sons de synthèse ont un contenu fréquentiel quasi-nul en HF. Certains n'ont plus d'harmoniques à partir de 500Hz, ce qui les rend inutilisables pour la suite.

Ainsi, ces sons ont dû être ajustés de manière à augmenter leur contenu HF. Pour cela, on augmente le nombre de modes de la structure résonante et/ou augmente la fréquence fondamentale de la structure (lorsque celle-ci augmente, les modes de résonances sont plus espacés et plus hauts en fréquence). Dans ce nouveau sous-corpus de synthèse, 39 fichiers audio sont créés par association à un ancien (même excitateur, même résonateur). Les autres sont générés avec d'autres valeurs de fréquence fondamentale du système résonant (cf annexe B.2).

Ces nouveaux sons de synthèse ont été validés en interne mais pas par l'expérience préliminaire de catégorisation. On peut cependant estimer qu'ils seront bien reconnus comme non-vocaux par la similarité de construction.

2.1.4 Homogénéisation du corpus

Objectif

Cette étape sert à s'assurer que les sons vocaux et non-vocaux ont le même contenu fréquentiel et à éventuellement corriger par un filtrage des différences trop importantes entre les trois classes.

En effet, la notion de chaleur peut être biaisée par le contenu spectral moyen de la classe de sons à partir de laquelle la notion de chaleur est extraite. Par exemple, si les sons vocaux sont en moyenne pauvres en BF, riches en HF et inversement pour les sons non-vocaux, on s'imagine que la chaleur ne sera pas la même pour ces deux classes à cause de cette différence : le filtre de « chaleur » pour les voix serait sensible aux BF mais pas aux HF, au contraire du « chaud » des sons non-vocaux. Cette variabilité inter-corpus est à éviter. De même, on homogénéise les sons dans la deuxième dimension manipulée, le rapport harmoniques sur bruit.

L'égalisation du contenu spectro-temporel et du HNR se limite à la bande fréquentielle [60 ; 5000] Hz. Cette réduction se justifie par rapport au contenu spectral des sons de la base qui contiennent beaucoup moins d'énergie en-dehors de cet intervalle. Cette limitation se retrouvera dans la fonction de corrélation inverse et donc aussi dans les résultats. Dans cette partie, les catégories en jeu ne sont pas les trois corpus de sons monosyllabiques, bisyllabiques et de synthèse mais bien les deux catégories vocaux et non-vocaux.

Egalisation du HNR

On définit le rapport HNR comme étant le rapport des spectres d'amplitude des parties harmonique et bruitée du son (supposé harmonique). On calcule ceux-ci à l'aide de l'outil `superVP` d'*AudioSculpt*. Le calcul et la modification du HNR sont détaillés dans la section 2.2.

Pour égaliser le corpus initial dans la dimension du HNR, on calcule chacun des HNRs pour les sons des deux catégories en échelle de Mel (cette échelle est plus proche de la manière dont les sons sont perçus par l'oreille humaine). Les spectres d'amplitude sont en dB [5]. On réduit la matrice de spectrogramme à une matrice de taille plus réduite en moyennant le spectrogramme dans trois fenêtres temporelles et dix fenêtres fréquentielles (valeurs arbitraires). On applique cette réduction de dimensions car la précision n'est pas ici cherchée : on s'intéresse davantage à un résultat global (critère qui a régit le choix arbitraire du nombre de fenêtres temporelles et fréquentielles précisé ci-contre). Aussi, le souci de ne pas détériorer les signaux est encore présent. C'est pourquoi on applique des modifications lisses en fréquence et en temps pour éviter de trop fortes discontinuités.

5. Lorsque ce n'est pas précisé, les spectres seront exprimés et appliqués en échelle de Mel et en dB.

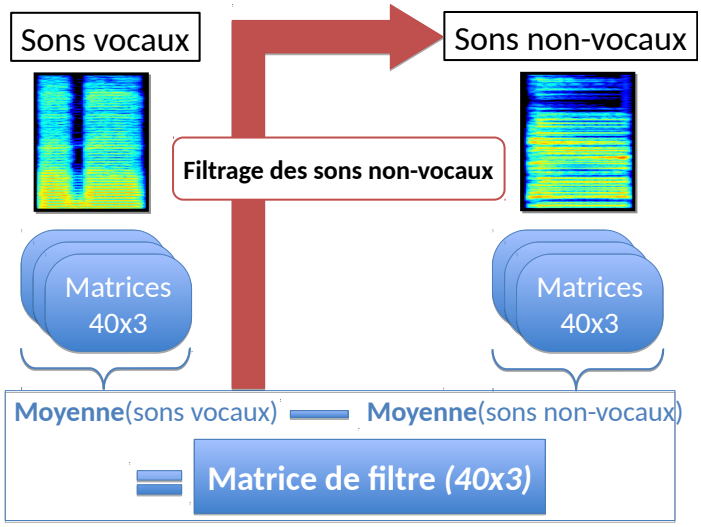


FIGURE 2.3 – Egalisation du contenu spectro-temporel : pour chaque élément du corpus, calcul du sonogramme, moyennage en fréquence et en temps pour obtenir les matrices 40x3, calcul de la différence des moyennes de chaque catégorie, application du filtre aux sons non-vocaux. Le principe est le même pour l'égalisation du HNR, à la différence qu'on n'agit pas directement sur le spectre par un EQ.

La différence des deux moyennes des matrices de HNR constitue le filtre qu'on applique dans un second temps à la partie harmonique des sons non-vocaux.

Le résultat de l'égalisation est représenté sur la figure 2.4 où l'on compare la différence des moyennes du HNR entre les sons vocaux et non-vocaux, avant et après filtrage d'égalisation. On s'attend à ce que les valeurs soient nulles après l'étape de normalisation. Ce n'est pas le cas. Cela s'explique par la manière dont on a homogénéisé les sons non-vocaux avec les vocaux : le filtrage est grossier et est destiné à gommer de grandes différences générales entre les les voix et les sons non-vocaux (chaque point du graphe représente la différence moyenne des amplitudes sur une fenêtre temporelle, sur toutes les fréquences, entre tous les sons vocaux et non-vocaux). De plus, les matrices 3x10 sont interpolées linéairement pour être appliquées sur l'ensemble des temps et des fréquences (les trois colonnes de la matrice correspondent aux milieux des trois fenêtres temporelles couvrant tout le signal, de même pour les lignes et fenêtres fréquentielles). Cette interpolation est faite par `superVP` et rend inévitable une différence entre le filtrage théorique et celui réellement appliqué.

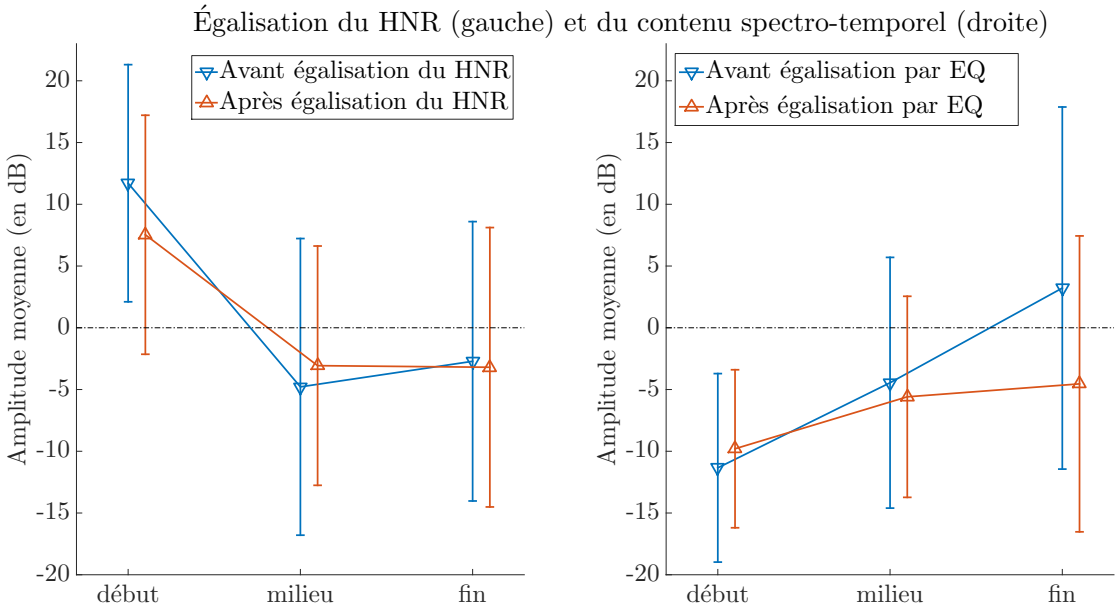


FIGURE 2.4 – Différence des moyennes avec écart-type des matrices pour l'EQ (à droite) et pour la modification du HNR (à gauche) sur toutes les fréquences et dans trois fenêtres temporelles, entre les sons vocaux et non-vocaux.

Egalisation du contenu spectro-temporel

Après l'égalisation du HNR, on fait de même pour ajuster le contenu spectro-temporel des sons non-vocaux en fonction du spectrogramme-moyen des deux catégories (la vue schématique de cette égalisation est en figure 2.3). Les résultats sont en figure 2.4. Comme pour la partie précédente, on s'attend à ce que les valeurs tendent vers zéro. La justesse de l'opération se vérifie expérimentalement en itérant le processus. Après quelques filtrages successifs, le spectrogramme moyen des fichiers audio non-vocaux égale bien celui des sons vocaux. Seulement, les artéfacts dus aux filtrages successifs sont trop forts pour pouvoir retenir cette procédure. On se contente donc du résultat obtenu après le premier filtrage.

Egalisation en volume sonore

Les pistes audio sont normées en RMS (moyenne quadratique, *Root Mean Square*). Cette normalisation ne garantit donc plus la normalisation en sonie faite, dont l'objectif essentiel était de garantir une uniformité de niveau sonore pour l'expérience de catégorisation (cet objectif n'a plus lieu d'être car les filtrages par la fonction de reverse correlation font changer aléatoirement les niveaux des *stimuli*).

Puis le niveau sonore maximum est réglé de sorte que le passage par la fonction de corrélation inverse ne puisse pas faire saturer le son.

2.2 Stimuli

Pour chaque participant, 18 fichiers audio sont pris aléatoirement parmi les 144 du corpus détaillé à la section 2.1, en respectant une répartition égale entre chaque sous-corpus et entre voix masculines et féminines. Chaque son est passé 3588 fois dans la fonction de *reverse correlation*. La fonction de corrélation inverse prend en argument d'entrée un fichier audio parmi les 18 sélectionnés, modifie aléatoirement le spectre et le rapport harmoniques-sur-bruit (cf figure 2.6) et ce, deux fois. Ces deux signaux ainsi filtrés forment une paire. Un silence de 0.5s sépare les deux extraits audio, et on applique un fade-in et fade-out de 10ms à chacun des deux sons de la paire. Cette opération n'est pas audible mais évite de créer un clip (causé par une discontinuité entre le signal et le silence) qui lui serait gênant.

Génération des matrices aléatoires

Pour faire varier aléatoirement les échantillons dans les dimensions du HNR et de l'EQ, on filtre par des matrices aléatoires dont on va maintenant détailler la construction.

On crée une matrice de taille 101x100 dont chaque point est tiré dans une distribution normale. On s'assure de plus que chaque point ne dépasse pas une valeur maximale valant 2.5 fois l'écart-type de la distribution normale (auquel cas on retire l'échantillon). Ensuite, on filtre la matrice par convolution 2D avec une matrice gaussienne de taille 40x20 pour lisser la matrice. Ainsi, on peut appliquer des gains plus élevés sans créer d'artéfacts à cause d'une irrégularité dans la suite des valeurs de la matrice (cf figure 2.5). La première dimension de la matrice correspond au temps, la seconde aux fréquences. Le choix des tailles (101x100 et 40x20) se fait de manière à obtenir une matrice aux variations ni trop lentes ni trop rapides. L'unité des gains de cette matrice est le décibel (dB). L'unité de σ_{ext} s'obtient en divisant les gains en décibel par σ_{ext} (*i.e.* (1 unité de σ_{ext}) $\leftrightarrow$ (5.46 dB pour l'EQ) $\leftrightarrow$ (8.16 dB pour le HNR)).

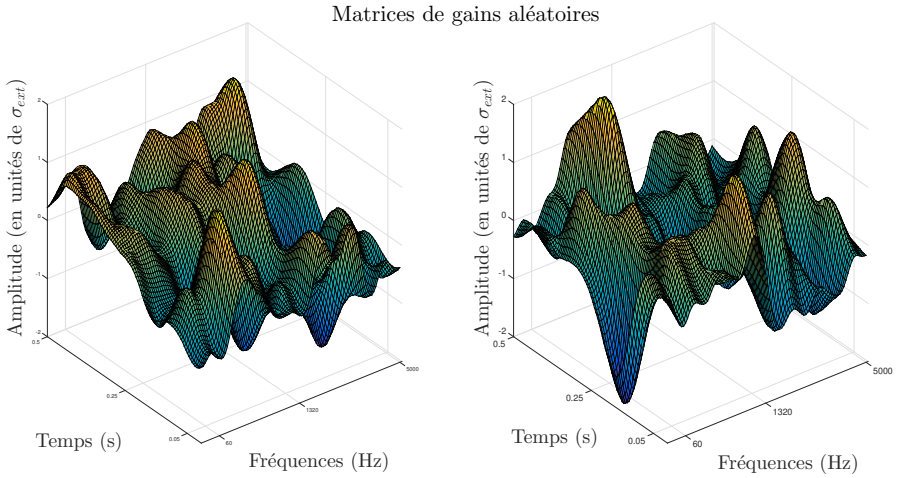


FIGURE 2.5 – Exemples de matrices aléatoires qui servent de filtre d'EQ ou pour le HNR. On retrouve les valeurs de gain en dB réellement appliquées en multipliant les amplitudes ci-contre par σ_{ext} ($\sigma_{ext,EQ} = 5.46$ dB et $\sigma_{ext,HNR} = 8.19$ dB).

Variation du HNR

Pour modifier le HNR d'un signal audio, on le décompose en une partie harmonique et une partie bruit à l'aide de `SuperVP`. On applique un EQ aléatoire à la partie hamonique puis recompose la partie bruitée avec la partie harmonique ainsi filtrée. Cette étape se fait également par appel à `superVP` qui filtre un fichier audio à partir d'une matrice dont il interpole linéairement les points. La matrice de filtre de la partie harmonique se génère comme expliqué ci-dessus. La valeur de l'écart-type de la distribution normale est de 15 dB avant le filtrage par la matrice gaussienne et de 8.19 dB après. On fait commencer le filtre à 60 Hz et s'arrêter à 5000 Hz. Il est appliqué en échelle de Mel. On garde l'échelle linéaire pour le temps et on filtre sur toute la durée du signal.

Variation du contenu spectro-temporel

Une fois la variation du HNR faite, on passe le signal dans un filtre spectro-temporel similaire. Les gains de la matrice aléatoire sont tirés dans une distribution normale d'écart-type de 10dB initialement. Après le filtrage convolutif, l'écart-type descend à 5.46 dB.

CreamDBgenerator

On utilise la *toolbox* `creamDBgenerator`[6] qui facilite la génération de telles matrices aléatoires et le filtrage du spectre. Cette *toolbox* est issue des travaux de J.J. Bur-

6. Outil fonctionnant sous Matlab.

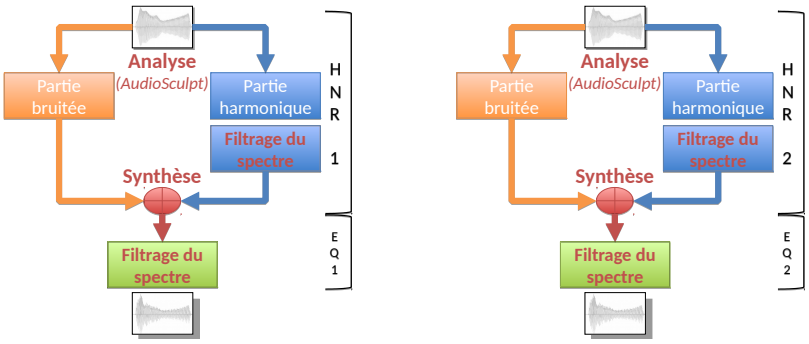


FIGURE 2.6 – Construction d'un *stimulus* à partir d'un son de la base initiale. On filtre aléatoirement dans les dimensions du HNR puis de l'EQ. Cette opération se répète une deuxième fois à partir de la même source pour créer une paire de sons. Cette paire est finalement présentée au participant qui juge lequel des deux sons est le plus « chaud ».

red, E. Ponsot, J.-J. Aucouturier et P. Belin sur des expériences de corrélation inverse en audio. Elle permet d'appliquer des transformations aléatoires sur des paramètres de sons vocaux comme le "time stretching", le "pitch shifting", le "time-varying equalization", le "time-varying gain". La dimension du HNR est une nouveauté pour les expériences de corrélation inverse en audio. Aussi, cette expérience est l'occasion de développer une fonction similaire pour cette dimension et de l'intégrer à la *toolbox*. On la complète par l'option qui permet de filtrer la matrice aléatoire de gains par le passe-bas gaussien. Les *stimuli* sont enregistrés avant les sessions d'écoute [7].

2.3 Procédure

Questionnaire et consigne

Après signature de la feuille de consentement, le participant répond à un questionnaire :

— Nom / Prénom / Âge / Sexe
— Avez-vous un trouble auditif ?
— Avez-vous déjé participé à un test en psycho-acoustique ? - Si oui, donnez-en une description rapide.
— Avez-vous une connaissance de la thématique de l'expérience ? Des travaux menés par l'équipe Perception et Design Sonores ?
— Quelle expérience avez-vous dans le domaine de l'ingénierie du son (nombre d'années d'études/de pratique) ?

7. Enregistrement au format WAV par l'appel au moteur de synthèse superVP, fréquence d'échantillonnage de 44100Hz et au format 16 bits.

— Donnez une définition/des aspects d'un son « chaud » en quelques lignes.

La consigne de l'expérience est lue par le participant (cf figure 2.7). Ensuite trois exemples de sons « chauds » sont écoutés. Ces exemples sonores sont issus des *workshops* de l'équipe PDS et ont été créés par le musicien y participant. Si le participant n'a pas de question, il commence l'écoute.

Segmentation de l'expérience

L'expérience est découpée en 4 sessions d'une heure chacune. Chaque session est composée de 6 parties d'une durée de 10 minutes environ. Chaque partie est formée de 162 paires de sons. Dans chaque session de 972 paires, il n'y a qu'en réalité 897 paires différentes car 75 paires sont répétées à un autre moment de la session (le bloc de paires répétées est placé aléatoirement dans la session). Ces répétitions serviront à estimer le bruit interne des participants, comme le fait E. Ponsot [15] (cf partie 3.1.4 pour plus de précisions).

A chaque fin de partie, un « score lié à la consistance du jugement [du participant] (...) est affiché. » Ce score est un nombre aléatoire entre 70% et 90%, qui se veut être un stimulant pour garder l'attention des participants sur cet exercice long, répétitif et pénible, et ainsi leur permettre d'avoir un jugement consistant, c'est-à-dire cohérent.

Matériel

L'écoute se fait au casque [8] et dans les mêmes cabines qu'utilisées lors de la pré-expérience de catégorisation. Le volume sonore est réglé de façon à ce qu'il varie entre 47.7 et 80.3 dBA [9].

L'interface logiciel est *Matlab* qui passe les sons, enregistre les réponses sous la forme de 1 et de 0 dans un fichier au format MAT, au rythme d'un fichier par partie (soit 6 fichiers par session, et au total,24 par participant). Parmi ces fichiers, l'un d'entre eux est illisible par l'ordinateur, donc inutilisable pour la suite. La proportion de données perdues (162 paires sur 3888 pour un participant) n'est pas significative pour exclure de l'analyse.

Les participants ont passé les deux premières sessions un premier jour et sont revenus un autre jour finir les deux autres heures. A la fin de l'écoute, ils ont rendu compte de leurs impressions et répondu à différentes questions posées par l'expérimentateur (par ex. sur la difficulté de la tâche, la pénibilité de l'expérience). Parfois, lors de leurs pauses, certains ont posé des questions portant sur les méthodes utilisées ou sur les hypothèses testées : on ne leur a répondu qu'après l'écoute, lors du bilan avec l'expérimentateur.

8. Le casque est de modèle Sennheiser HD250 II-"Linear".
9. Les niveaux sont enregistrés par le même sonomètre que celui utilisé lors de la pré-expérience de catégorisation.

EXPÉRIENCE D'ÉCOUTE

A. **Merci** d'avoir accepté de participer à cette expérience. Elle rentre dans le cadre d'une étude portant sur la caractérisation du code acoustique de l'attribut « chaud ».

B. Consigne

Vous allez entendre au casque des paires de sons de courte durée.
Après avoir écouté une paire, vous devrez répondre à la question suivante :

« **Quel son est le plus CHAUD** ?»

Les **réponses** possibles sont : **« Le premier » ou « Le second ».** La réponse se fait au clavier ('A' = le 1er ; 'P' = le 2nd). Après avoir répondu, une autre paire sera présentée.

C. Définition d'un son « chaud »

Pour vous aider dans la tâche de jugement, nous vous proposons auparavant d'**écouter trois exemples de sons « chauds »** qui seront répétés une fois chacun, ainsi que des **caractéristiques d'un son « chaud »** : « Un son chaud est un son plutôt grave qui donne une impression de largeur, de déploiement et d'une certaine richesse qui le rend plutôt agréable. »
Soyez spontané dans votre réponse.

D. Déroulement de l'expérience

- L'expérience est divisée en **2 séances** ; une séance est découpée en **12 sessions**. Une **session** dure **10 min**. A la fin de chaque session, un score lié à la consistance de votre jugement sur la session est affiché. Essayez de le maintenir à au moins 75%.
- Entre chaque session, vous pouvez faire une pause de quelques instants.
- **Au bout de la 6e session (1h)**, vous êtes invité à prendre une **pause de ~5-10 min** en dehors de la cabine.

- Si vous rencontrez un quelconque problème, n'hésitez pas à en faire part à l'expérimentateur.

Merci de votre participation

FIGURE 2.7 – Consigne de l'expérience donnée aux participants.

2.4 Participants

La notion de son « chaud » est peu commune. En réalité, elle est surtout employée dans les milieux professionnels du son, par exemple dans les métiers d'ingénieur son, de designer sonore, parmi les compositeurs. C'est pourquoi les participants de cette expérience ont été recrutés dans ces environnements-là. Plus précisément, ils ont été sélectionnés parmi les étudiants en spécialité « son » de l'école nationale supérieure Louis-Lumière ou de la Formation Supérieure aux Métiers du Son (FSMS) du CNSMDP (conservatoire de Paris). Avec ce choix de sélection, on espère qu'ils auront une idée précise de ce qu'est un son « chaud », donc que leur jugement sera plus fiable.

15 participants (9 hommes et 6 femmes, entre 21 et 31 ans et de moyenne d'âge de 23 ans) ont passé l'expérience. Aucun n'a reporté de problème auditif qui aurait gêné l'écoute. Ils ont tous mentionné une expertise en ingénierie du son. Ils ont donné leur consentement à l'expérience par écrit avant de débuter l'expérience.

Un participant a abandonné l'expérience au bout de 20 minutes à cause des conditions trouvées trop « oppressantes » . Un autre a, par erreur de manipulation, passé deux fois la première partie de l'expérience. Ces deux participants ont été retirés de l'analyse. Cette dernière portera donc sur 13 personnes (7 hommes et 6 femmes). 8 d'entre eux ont suivi ou suivent la formation FSMS du CNSMDP et 5 l'école Louis-Lumière (ils sont soit en formation soit diplômés depuis au plus 4 ans).

Chapitre 3

Résultats : analyse et discussion

Dans un premier temps, les résultats sont exposés et on indique la manière dont ils ont été obtenus. Ils sont ensuite commentés, interprétés, discutés et suivis de conclusions.

3.1 Résultats

3.1.1 Filtres moyens du son « chaud »

Savoir si la représentation mentale de « chaleur sonore » est physiquement caractérisable constitue la première partie de la problématique. Pour répondre à cette question, chaque participant a comparé la « chaleur » de deux sons appairés 3888 fois. La différence des matrices de filtre appliquées sur les deux échantillons d'une paire renseigne sur ce que le participant a jugé comme étant plus « chaud » selon sa réponse (cf figure 1.5). Et donc on peut construire la représentation que le participant s'est fait sur la chaleur sonore en moyennant les différences enregistrées pour chaque paire.

Nommons $EQ_1(k)$, $EQ_2(k)$, $HNR_1(k)$ et $HNR_2(k)$ les matrices de filtre générées par la fonction de corrélation inverse pour bruiter dans les dimensions du spectre et du HNR les deux échantillons de la k-ième paire (voir les correspondances de notation avec la figure 2.5). Posons $\delta(k)$ la réponse à la question posée (« quel son est le plus chaud ? »), valant 0 si la réponse est « le premier », 1 si c'est « le second ». Notons σ_{EQ} et σ_{HNR} les écarts-types des distributions normales utilisées ($\sigma_{HNR} = 8.19$ dB et $\sigma_{EQ} = 5.46$ dB, cf section 2.2) [1].

Alors $\mu_{EQ,1participant}$, le filtre moyen véhiculant la chaleur chez un participant - que l'on cherche à déterminer par l'expérience - s'exprime par :

$$\mu_{EQ,1participant} = \frac{1}{K.\sigma_{EQ}} \sum_{k=1}^{K=3888} (1 - 2\delta(k)).(EQ_1(k) - EQ_2(k)) \tag{3.1}$$

1. Par la suite, l'écart-type pourra être désigné par SD (*Standard Deviation*) ou « σ ».

De même pour $\mu_{HNR,1participant}$ avec les matrices de filtre correspondant. La figure 3.1 représente la moyenne sur tous les participants des $\mu_{EQ,1participant}$ (à gauche) et des $\mu_{HNR,1participant}$ (à droite). Ce tracé permettra de voir s'il y a une tendance générale de représentation entre les participants (deuxième question de la problématique). De plus, comme la troisième question de la problématique porte sur une différence ou ressemblance entre les trois catégories de sons (monosyllabiques, bisyllabiques et de synthèse), un filtre-moyen est tracé pour chacun des trois corpus. Les matrices de filtre pour chaque participant sont disponibles en annexe C.1.1. La normalisation par la SD permet d'exprimer les résultats en unités de σ_{ext} (unités de SD de bruit externe) [22, 15].

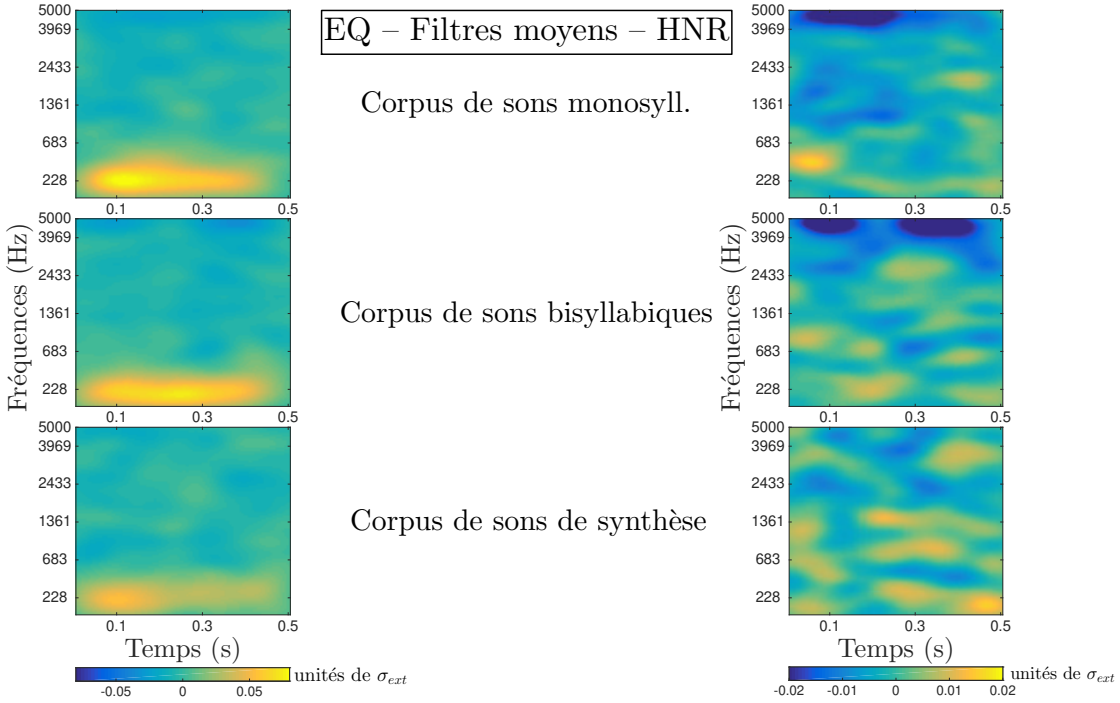


FIGURE 3.1 – Filtre moyen de « chaleur » selon les trois corpus de sons monosyllabiques, bisyllabiques et de synthèse (de haut en bas), dans les dimensions de l'EQ (à gauche) et du HNR (à droite) en unités de SD de bruit externe.

3.1.2 Z-scores

La représentation précédente est la moyenne sur tous les participants des filtres-moyens calculés pour chaque participant. Elle a le défaut de ne pas montrer les différences entre les filtres individuels. Aussi, la figure 3.2 représente les mêmes filtres sous la forme de z-scores. Le z-score est calculé à partir des moyennes normalisées en RMS : $zscore_{EQ} = \frac{\mu_{EQ,norm}}{SD_{EQ}}$ avec $\mu_{EQ,norm}$ le filtre-moyen en EQ normalisé par sa valeur RMS

et SD l'écart-type du vecteur $[(\mu_{EQ,1participant})_{1...N_{participants}}]$. Le $zscore_{HNR}$ se calcule de manière similaire.

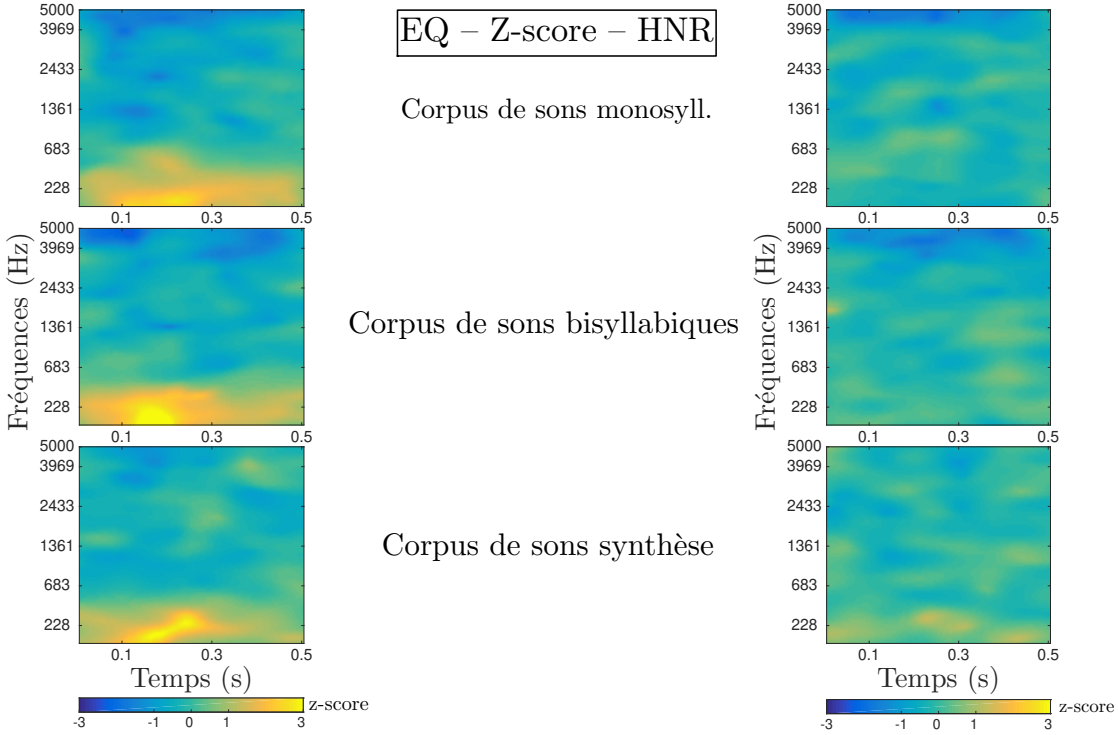


FIGURE 3.2 – Z-score du filtre moyen de « chaleur » selon les trois corpus de sons monosyllabiques, bisyllabiques et de synthèse (de haut en bas), dans les dimensions de l'EQ (à gauche) et du HNR (à droite).

Le z-score est un écart-type qu'on peut relier à la *p-value*, probabilité que le modèle observé soit généré à partir d'un processus aléatoire. Une faible valeur de p indique une faible probabilité que le résultat soit aléatoire. Une valeur de $|zscore| > 2$ équivaut à $p < 0.05$ (c'est-à-dire qu'il y a 95% de chances pour que le résultat ne soit pas issu d'un processus aléatoire). De même, $|zscore| > 1.65$ équivaut à $p < 0.10$ (c'est-à-dire à un niveau de confiance de 90%).

Les matrices de la figure 3.2 sont tracées une deuxième fois sur la figure 3.3. On met en évidence les zones de $|zscore| > 2$ (contours rouges) pour les matrices d'EQ, et de $|zscore| > 1.65$ (contours magenta) pour les matrices de filtre du HNR (pour ces dernières matrices, on prend un niveau de confiance plus faible que pour les matrices d'EQ à caude des zones de $|zscore| > 2$ trop réduites).

Les figures 3.1, 3.2 et 3.3 montrent que l'EQ-moyen donnant un son « chaud » (figures de gauche) comporte des BF accentuées. On peut localiser ce maximum autour de [60 ; 400] Hz.

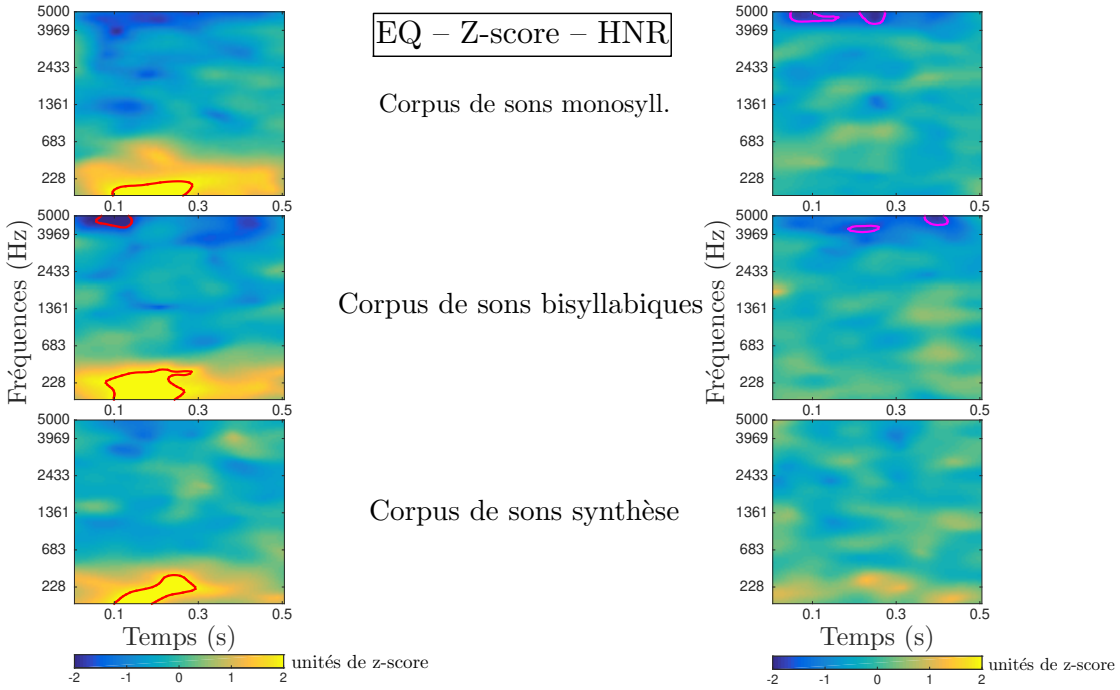


FIGURE 3.3 – Figure 3.2 à laquelle on a superposé les contours correspondant à des valeurs de $|zscore| > 2$ (contours rouges, figures de gauche) et de $|zscore| > 1.65$ (contours magenta, figures de droite). Ici, l'échelle des couleurs est de [-2 ; 2].

Par ailleurs, il semble y avoir une évolution temporelle. Plus précisément, la zone accentuée semble être davantage dans la première moitié du son et cet effet semble plus important pour le corpus de sons de synthèse.

Le tracé des z-scores (cf figure 3.2) fait apparaître une pente négative dans le haut du spectre pour le corpus de sons vocaux et une accentuation du HNR dans les BF pour le corpus de sons non-vocaux (figures de droite). Ce résultat est cependant moins probant car les zones de contour magenta correspondent à $p < 0.10$.

Ces premiers résultats vont être analysés plus précisément par l'étude des moyennes marginales.

3.1.3 Moyennes marginales

Les filtres manipulés sont des matrices 2D, dont les colonnes représentent le temps, les lignes les fréquences. On trace les moyennes marginales (cf figure 3.4) des matrices précédentes, c'est-à-dire qu'à une matrice 2D correspond un vecteur où chaque point correspond à la moyenne des lignes ou des colonnes, selon la dimension que l'on garde.

On retrouve les tendances fréquentielles pour l'EQ et le filtre du HNR (figures du bas). Le pic pour l'EQ est atteint autour de 230 Hz. Les tracés des moyennes marginales

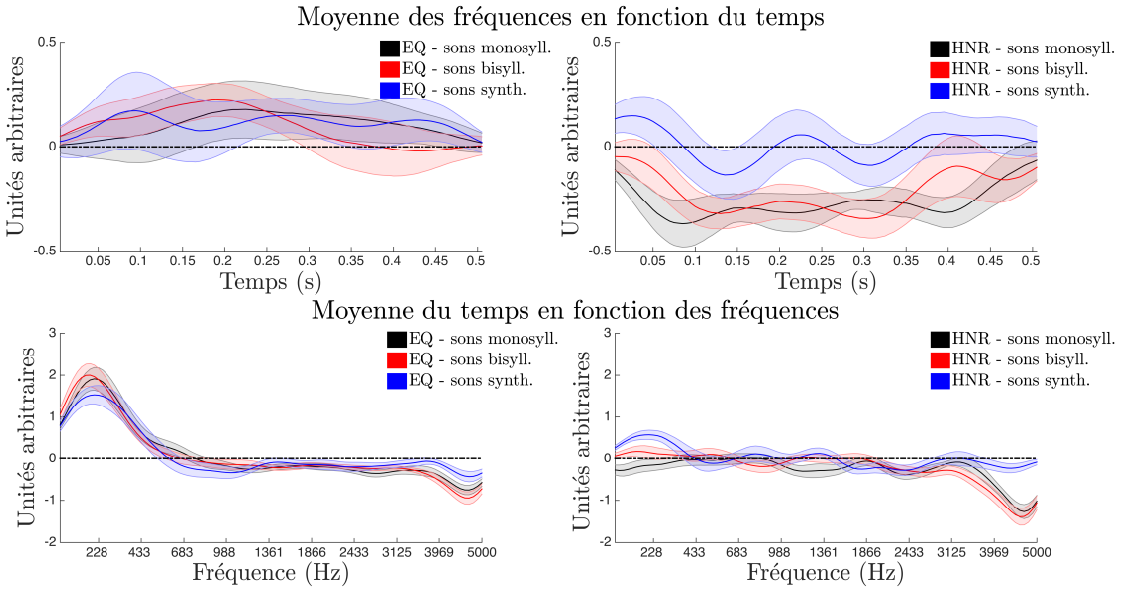


FIGURE 3.4 – Tracé des moyennes marginales et de leurs erreurs-type (SEM : *Standard Error of the Mean*). En haut, tracé des amplitudes en fonction du temps (moyenne sur les fréquences) pour les dimensions de l'EQ (à gauche) et du HNR (à droite) à partir des matrices de filtres-moyens normalisés en RMS. En bas, amplitudes en fonction des fréquences (moyenne sur le temps). Les corpus de sons monosyllabiques, bisyllabiques et de synthèse sont représentés en noir, rouge et bleu respectivement.

en fonction du temps ne font pas apparaître de tendance générale. Ce n'est pas étonnant car les moyennes sont faites sur l'ensemble des fréquences, et non sur la plage fréquentielle concernée ([60 ; 400] Hz par ex.).

En première approximation, on essaye de corréler linéairement les moyennes marginales avec l'axe des abscisses qui leur correspondent. On trouve ainsi que la dépendance linéaire de la moyenne marginale du temps en fonction des fréquences est significative à $p < 0.002$ pour l'EQ. Les autres corrélations linéaires ne sont pas significatives ($p > 0.05$). On calcule une deuxième fois les mêmes moyennes marginales, cette fois à partir des matrices normalisées en RMS (Joosten et Neri font de même pour une expérience de détection de hauteur [23]. C'est d'ailleurs à partir de ces matrices normalisées que l'on trace les moyennes marginales de la figure 3.4.). On trouve cette fois-ci une dépendance linéaire de la moyenne marginale du temps en fonction des fréquences significative à $p < 0.002$ pour le filtre du HNR.

On observe à présent la moyenne marginale des fréquences en fonction du temps pour l'EQ à partir des matrices normalisées mais en ne prenant que les fréquences comprises entre 60 et 683 Hz, pour observer l'évolution temporelle de la « bosse » BF. La valeur de 683 Hz est arbitraire. Elle correspond à la fréquence pour laquelle la moyenne marginale

des temps en fonction de la fréquence coupe l'axe des fréquences (cf figure 3.4 en bas à gauche). On peut comparer ce tracé à un « zoom » fréquentiel du tracé en haut à gauche de la figure 3.4. La figure 3.5 fait ressortir l'évolution temporelle en BF.

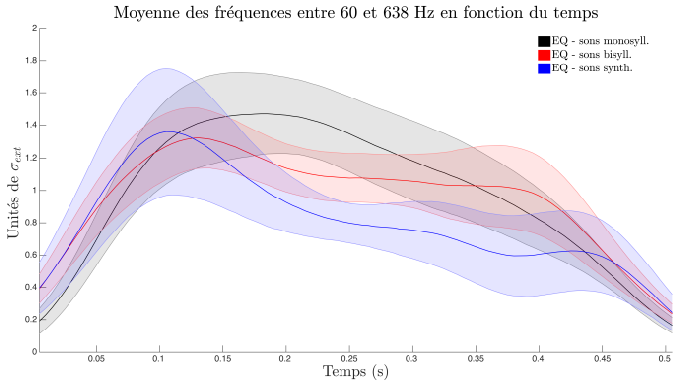


FIGURE 3.5 – Tracé correspondant au graphe en haut à gauche de la figure 3.4 où l'on n'a moyenné les fréquences que sur l'intervalle [60 ; 683] Hz (intervalle correspondant à la localisation fréquentielle de la « bosse »). Les corpus de sons monosyllabiques, bisyllabiques et de synthèse sont représentés en noir, rouge et bleu respectivement.

3.1.4 Consistance des réponses et bruit interne.

En plus du bruit externe présent dans les *stimuli* (qui est le bruit expérimental, celui introduit par la fonction de corrélation inverse et quantifié par la SD des gains des matrices aléatoires), il existe un bruit propre au participant. Il s'agit d'une autre source de variabilité dans les résultats et peut s'expliquer par plusieurs facteurs : bruit neuronal, distraction pendant l'expérience, incertitude dans le jugement. Il se calcule à partir d'un modèle théorique de détection du signal. On utilise la méthode de double passage ("double pass") [15, 22, 24]. Le bruit interne est un écart-type exprimé en unités de SD de bruit externe (σ_{ext}) . La figure 3.6 illustre le bruit interne calculé pour chaque participant. Plus la variabilité dans les réponses des paires répétées est grande, plus l'écart-type de la distribution représentatif du jugement est grand, plus le bruit interne est grand.

Cette mesure permet d'évaluer la consistance des réponses des participants. Pour effectuer cette mesure, on se sert des paires qui ont été répétées au cours des sessions, sans que les participants en aient été prévenus. Ainsi, s'ils donnent les mêmes réponses aux mêmes *stimuli*, cela montrera qu'ils ont une tactique de réponse (et sinon, qu'ils ont répondu aléatoirement). On ne pourra pas vérifier que les participants ont fait exactement la tâche demandée (à savoir juger de la « chaleur » sonore), néanmoins on saura qu'ils n'ont pas répondu au hasard et jugé selon un critère qu'ils se sont donné.

Neri calcule le bruit interne des participants dans des expériences de détection et de discrimination psychophysique en audition et en vision [22]. Pour ces expériences, il y

a une bonne et une mauvaise réponse à chaque essai (la tâche porte sur la détection de cibles) à la différence de cette expérience-ci, où le critère est subjectif. L'écart-type de 1.3 ± 0.75 correspond à ce que Neri trouve pour les expériences de 2-AFC qu'il a menées. Cette SD est désignée par « aire-référence » sur la figure 3.6 : la ligne horizontale bleue représente la moyenne, la zone bleutée la SEM.

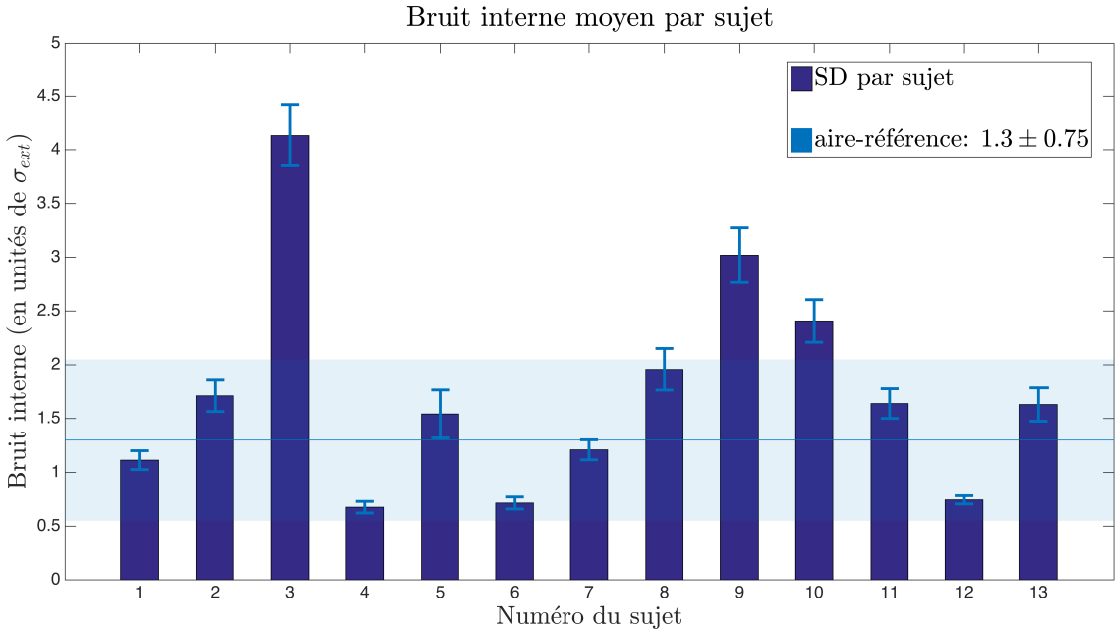


FIGURE 3.6 – Pour chaque participant, on calcule le bruit interne en comparant les réponses à une même paire répétée deux fois pendant l'expérience. On compare les résultats à la mesure du bruit interne faite par Neri [22] : 1.3 ± 0.75. Les barres d'erreur sont en SEM.

10 des 13 participants de l'expérience de corrélation inverse ont un bruit interne inclus dans la plage de valeurs issue des expériences de Neri.

3.1.5 Bruit interne par corpus

Alors qu'au paragraphe précédent on a calculé le bruit interne par participant en moyennant sur les corpus, on le calcule ici par corpus en moyennant sur les participants (cf figure 3.7).

On remarque que le bruit interne des corpus de synthèse est en moyenne plus grand que celui du corpus de voix (cf figure 3.7). On réalise un t-test entre les distributions de bruits internes pour chaque corpus, pour confirmer cette observation (donc au total trois t-tests). On obtient un $p < 0.05$ uniquement lorsque l'on compare le corpus de sons bisyllabiques et celui de synthèse ($p = 0.016$).

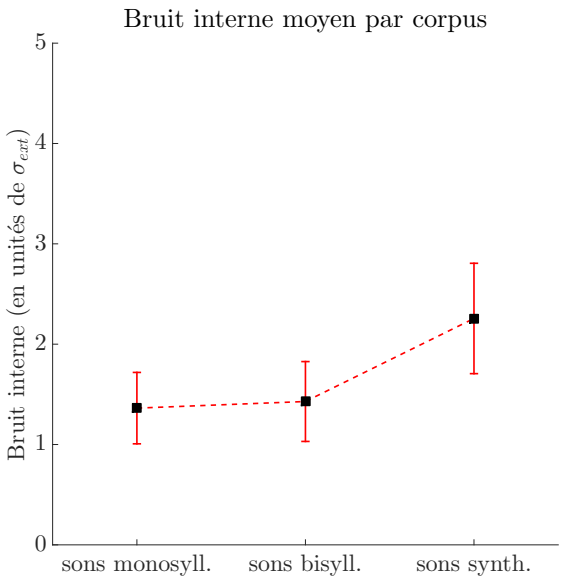


FIGURE 3.7 – Moyenne de bruit interne par corpus avec SEM.

Une autre question que l'on se pose est si la source de bruit interne est la même pour les catégories de sons vocaux et non-vocaux. Par des tests de corrélation de Pearson, on trouve que seuls les corpus de sons monosyllabiques et bisyllabiques sont corrélés (à $p = 0.005$), ce que l'on observe graphiquement à la figure 3.8.

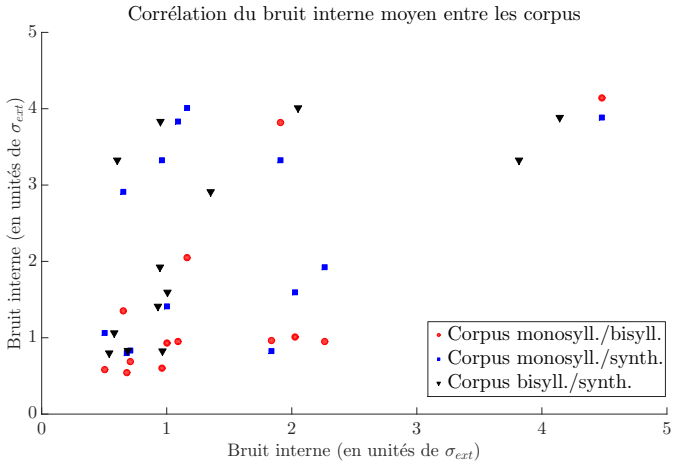


FIGURE 3.8 – Corrélation des bruits internes entre les 3 corpus. Les résultats confirment que seuls les corpus de sons monosyllabiques et bisyllabiques sont corrélés (en rouge).

3.1.6 Retours d'expérience des participants

A la fin de l'expérience, les participants ont répondu aux questions de l'expérimentateur afin d'avoir un retour sur l'expérience : leurs ressentis, les difficultés, leurs avis sur la tâche demandée etc. Ce paragraphe rapporte les vues globales sur 4 points : la définition du son « chaud » qu'ils ont donnée, la pénibilité de l'expérience, la tâche de jugement, le schéma de l'expérience.

Définition du concept

A propos de la définition de la « chaleur », 10 sur les 15 ont mentionné un renforcement des BF, 5 le côté agréable, 4 un rapport avec les harmoniques, 4 avec l'enveloppe temporelle. On trouvera une transcription de toutes les réponses en annexe C.2.

Tâche de jugement

En majorité, les participants n'ont pas trouvé la tâche de jugement difficile. Certains ont été troublés par les *stimuli*, trouvés courts et isolés de tout contexte auquel le concept est intimement lié, selon eux.

Pénibilité de l'expérience

Les participants ont avoué avoir trouvé l'expérience pénible. L'abandon d'un participant en cours d'épreuve est révélateur. De plus, un autre a dit avoir hésité à revenir à la deuxième séance d'écoute. Quand l'expérimentateur a demandé à dix d'entre eux d'évaluer la pénibilité sur une échelle de 1 (pas pénible) à 10 (très pénible), ils ont répondu en moyenne 6.8 ± 1.5 sur 10 (les deux participants dont il est question au début de paragraphe ne sont pas parmi ces 10). Les réponses telle « 8/10, parce que je suis gentille ! » confirment ce fait.

Il revient aussi que leur concentration a fluctué au cours du temps. Parfois même, leur tactique de réponse a changé : deux ont relevé qu'au début leur jugement était analytique (par rapport à la définition qu'ils ont donné) et avec la fatigue, leur jugement est devenu plus instinctif (tout en jugeant la même notion de « chaleur », rapportent-ils).

Schéma de l'expérience

Notons que l'indice affiché à la fin des parties, « lié à la consistance » (score aléatoire destiné à les tenir concentrés) a perturbé certains (au moins deux) qui ont fait un lien entre « leurs scores » et leur tactique de jugement : lorsqu'ils répondent instinctivement, leur score est plus élevé. Pour d'autres, il varie « logiquement » selon leur état de fatigue.

3.2 Discussion

3.2.1 Contenu spectro-temporel du filtre-moyen de « chaleur »

Des matrices en z-score et des résultats statistiques (cf figure 3.2, parties 3.1.2 et 3.1.3) on peut conclure que :

- **l'expérience a montré que la notion de « chaleur » s'explique par une répartition fréquentielle de l'énergie, plus précisément par une concentration de l'énergie en basses fréquences autour de 60-400 Hz**.
- Egalement, **le « chaud » est davantage localisé sur la première moitié temporelle du son ; cet effet est plus marqué pour les sons de synthèse que les sons vocaux**.
- **Il n'apparaît pas de différence de représentation de la « chaleur » entre les sons monosyllabiques et bisyllabiques.**

Des études de corrélation plus poussées mériteraient d'être faites. On peut notamment penser à une étude de corrélation de la « bosse » fréquentielle et temporelle avec une fonction non-linéaire (polynomiale par ex.). Une meilleure précision sur la localisation temporelle pourrait permettre de distinguer les représentations mentales des corpus vocaux et non-vocaux de manière plus significative. On pourrait aussi rapprocher l'évolution temporelle de l'enveloppe temporelle des sons de la base. Par exemple, on pourrait dans un premier temps caractériser les enveloppes des courbes de la figure 3.5 selon les paramètres de l'enveloppe ADSR (*Attack Decay Sustain Release*, Attaque Chute Entretien Extinction en français), puis la comparer à l'enveloppe ADSR moyenne des sons des trois sous-corpus.

3.2.2 Contenu bruité du filtre-moyen de « chaleur »

D'après les résultats de la partie précédente (cf parties 3.1.2 et 3.1.3), **il semblerait que la « chaleur » sonore soit liée pour les sons vocaux à un HNR qui décroît dans les hautes fréquences (à partir de 4000 Hz), pour les sons de synthèse à un HNR plus élevé dans les basses fréquences (autour de 60-300 Hz). Aucune dépendance temporelle entre l'attribut sonore et le HNR n'est trouvée**. Le lien fréquentiel qui distinguerait le son « chaud » vocal du non-vocal est à confirmer. Comme pour l'EQ, des études supplémentaires mériteraient d'être faites, en particulier pour corréler les effets trouvés à des fonctions non-linéaires.

3.2.3 Consistance des réponses et bruit interne

On a trouvé que 10 des 13 participants de l'expérience de corrélation inverse ont un bruit interne inclus dans la plage de valeurs issue des expériences de Neri. **Cela montre d'abord que les participants n'ont pas répondu au hasard**, et ensuite **que la tâche de cette expérience n'a pas ajouté de bruit supplémentaire**.

3.2.4 Bruit interne par corpus

On conclut des résultats trouvés (cf section 3.1.4 que **le bruit interne est significativement plus élevé quand il s'agit de juger de la « chaleur » des sons non-vocaux que des sons vocaux.** L'hypothèse que l'on peut proposer pour expliquer ce fait est que les sons vocaux sont plus naturels que les sons non-vocaux (le corpus est construit ainsi, et les participants l'ont remarqué). Et comme l'idée de son « chaud » se construit empiriquement chez les personnes, le concept est moins précis pour les sons qu'ils n'ont pas l'habitude d'écouter.

3.2.5 Retours d'expérience

Les avis des participants sur l'expérience valident certaines hypothèses et font entrevoir certaines limites expérimentales de la méthode de correlation inverse.

Définition du concept

Les définitions qu'ils ont données convergent vers celle qu'on leur a proposée (qui est, pour rappel, le fruit du *workshop* de l'équipe PDS). On remarque que l'aspect temporel revient régulièrement dans les définitions, ce qui est cohérent avec les tracés de la figure 3.1. Leur grande cohérence dans leur définition vient sûrement du fait qu'ils sortent des mêmes écoles, et donc ont eu les mêmes professeurs qui ont dû leur donner une même définition (ce que les participants ont affirmé).

Tâche de jugement

Les commentaires des participants sur la faisabilité de la tâche de jugement et la mesure du bruit interne font penser qu'ils ont bien fait la tâche demandée, et donc qu'elle était bien faisable. Cela veut dire que **les dimensions manipulées ont bien influé sur l'aspect « chaud » des sons** de la base.

Pénibilité de l'expérience

L'expérience est assez conséquente : 4h d'écoute, sur 2 séances, écoute de 3888 paires de sons. L'aspect répétitif, routinier de la tâche n'est pas négligeable. Cela milite en

faveur d'une réduction du temps de passation. Or la longueur de l'expérience dépend du nombre de dimensions testées et de la précision des résultats voulue. Il y a donc un compromis à faire si l'on veut garder cette méthode expérimentale. De plus, on pourrait optimiser la durée de l'expérience en calculant la vitesse de convergence des matrices de chaleur. Cela revient à s'intéresser à la convergence de la suite $(\mu_{EQ,1participant})_{K=1..3888}$ où l'indice est K le nombre de paires écoutées (cf équation 3.1). On saurait ainsi à partir de combien de paires écoutées la matrice moyenne de filtre est assez semblable à la matrice-limite (selon un critère à préciser).

Schéma de l'expérience

Quant au choix du corpus du sons, on peut remettre en question la distinction monosyllabique/bisyllabique étant donné que les résultats ne font pas ressortir de différence entre ces deux sous-corpus. Peut-être que rien n'est apparu parce que les sons bisyllabiques sont trop simples et que l'hypothèse sur l'assimilation de ces sons comme des mots réels (mais sans signification) n'est pas validée. Peut-être qu'une simple distinction vocal/non-vocal aurait suffi.

3.2.6 Comparaison avec l'expérience de Sabin

Pour rappel, l'expérience a été menée à la suite du travail de Sabin (cf section 1).

L'expérience de Sabin présente l'avantage de tester plusieurs attributs acoustiques haut-niveau en même temps. Il réussit à limiter la durée de l'expérience en choisissant 75 courbes d'EQ parmi 1000 les plus dissemblables entre elles, là où la méthode de corrélation inverse n'a aucun contrôle sur l'aléatoire (mis à part celui sur l'écart-type et sur la vitesse de variations). Les résultats de Sabin sont certes moins consistants entre les participants mais ceux-ci ne sont pas choisis selon leur expertise dans le domaine du son.

Par contre, les résultats de Sabin ne sont pas généralisables car les *stimuli* ne sont créés qu'à partir d'un seul fichier audio par catégorie. L'expérience de corrélation inverse se distingue surtout par l'étude de nouvelles dimensions de l'attribut acoustique : la variation l'EQ au cours du temps et le bruitage du HNR des *stimuli* sont des nouveautés.

Chapitre 4

Conclusion et perspectives

L'expérience est partie d'une étude lexicale d'attributs sonores, d'une expérience tâchant de caractériser l'attribut « chaud » par un EQ, de différents études montrant que l'oreille humaine traite différemment les sons naturels des sons synthétiques, les sons vocaux des non-vocaux.

On a mis en place cette expérience utilisant la méthode de corrélation inverse pour déterminer comment les sons « chauds » sont physiquement perçus par les professionnels du son. Rappelons la problématique qui a guidé ce travail :

1) La représentation mentale de « chaleur sonore » est-elle physiquement caractérisable par la méthode de corrélation inverse ?
2) Existe-t-il une représentation commune entre différentes personnes ?
3) Existe-t-il une représentation commune entre les sons vocaux et les sons non-vocaux ?

Résultats de l'expérience Les résultats ont montré que les professionnels du son ont bien une représentation commune des sons « chauds » en terme d'enveloppe spectro-temporelle : il ressort qu'un son « chaud » est lié à une certaine répartition de l'énergie, en particulier à une concentration de BF autour de 60-400 Hz. Ce qui répond à la deuxième question de la problématique et donc aussi à la première.

Les différents tracés tendent à montrer que cette répartition de l'énergie dépend aussi du temps : il semblerait que la concentration en BF se localise sur la première moitié temporelle du son ; cet effet serait plus marqué pour les sons de synthèse que les sons vocaux, même si les calculs statistiques réalisés n'ont pas vérifié cela. L'aspect « bruité » (qui se rapporterait au côté « agréable ») serait lié à une diminution du HNR dans les HF (4000-5000 Hz) pour les sons vocaux, et à une accentuation dans les BF (60-300 Hz) pour les sons non-vocaux. Enfin, le calcul de bruit interne montre que l'idée de son « chaud » est plus confuse pour les sons non-vocaux.

Amélioration de l'expérience Suite à ce travail, on peut penser à certaines améliorations et à plusieurs pistes de recherche. Tout d'abord, on pourrait valider le descripteur par une autre expérience qui filtrerait des exemples sonores par les matrices de filtre obtenues, avec des gains variables (pour mimer un son plus ou moins « chaud »). Les participants auraient à juger de la « chaleur » et on pourrait corréler les réponses avec la valeur du gain appliqué (ce que fait d'ailleurs Sabin).

Les courbes de moyennes marginales pourraient de plus être plus finement décrites, par exemple par une corrélation à des fonctions non-linéaires. On pourrait aussi calculer la vitesse de convergence des matrices de filtre. Cela pourrait aider à optimiser la durée de l'expérience. Dans cette même optique, on pourrait peut-être se rapprocher de la méthode utilisée par Sabin, consistant à choisir les courbes aléatoires selon leur dissemblance pour couvrir l'ensemble de la dimension plus rapidement, c'est-à-dire avec moins d'essais.

Au niveau du corpus de sons, on pourrait simplifier la division monosyllabique, bisyllabique et de synthèse en vocal/non-vocaux ou en naturel/synthétique. Si l'on poursuit la présente division, on pourrait choisir des sons bisyllabiques plus longs. Le caractère « chaleureux » d'une voix serait lié à la prosodie, donc augmenter la durée des sons de voix permettrait de jouer sur ce paramètre et donc pourrait éventuellement faire transparaître l'aspect « chaud » du fait qu'on infère sur la personnalité à partir de la voix.

A présent que l'on sait que la notion est physiquement caractérisable, on pourrait élargir le recrutement des participants, quitte à séparer les résultats selon l'expertise des participants en matière sonore.

Perspectives sur l'étude des attributs acoustiques haut-niveau Pour continuer le travail sur la caractérisation du son « chaud », l'aspect subjectif « agréable » est intéressant mais serait à préciser avant de pouvoir l'étudier en terme d'acoustique. Il se pourrait que les attributs « chaud » et « rond » soient parfois confondus (cf les définitions données par les sujets, serait-ce cette confusion qui fait apparaître une dépendance temporelle dans nos résultats ?), on pourrait les étudier en les associant. Il serait intéressant de voir si une expérience sur l'adjectif « froid », opposé de « chaud » donnerait le résultat opposé. On pourrait même refaire passer l'expérience en posant la question : « quel son est le plus *froid* ? ».

La méthode de corrélation inverse trouve aussi ses limites dans les possibilités de manipulation des variations. Par exemple, à partir de cette base de sons, il n'est pas possible de tester l'hypothèse sur les harmoniques paires parce que les sons n'ont pas la même hauteur donc les harmoniques ne sont pas aux mêmes fréquences, et seraient trop rapprochées pour espérer les faire varier indépendamment avec la matrice de filtre aléatoire ainsi construite.

Cette expérience montre que l'on peut caractériser physiquement un attribut sonore haut-niveau dans plusieurs dimensions par corrélation inverse. La méthodologie pourra s'étendre à d'autres descripteurs sonores relevant du discours sémantique.

Annexe A

Introduction

OCC.	ENG. WORD	FRENCH WORD	OCC.	ENG. WORD	FRENCH WORD	OCC.	ENG. WORD	FRENCH WORD
29	Soft	Doux	9	Light	Léger	6	Uneven*	Irrégulier*
28	Dull	Sourd, mat	9	Noisy	Bruité	6	Deep	Profond
21	High	Aigu	9	Muffled	Feutré	6	Narrow	Etriqué
21	Loud	Fort	9	Large	Large	6	Tonal	Tonal
19	Low	Grave	9	Strong	Puissant	6	Cold	Froid
19	Sharp	Aiguisé, incisif	9	Resonant*	Résonant*	6	Near	Proche
19	Rough	Rugueux	8	Thin	Mince	5	Piercing	Perçant
18	Bright	Brillant	8	Long*	Long*	5	Strident	Strident
16	Smooth	Lisse	8	Continuou	Continu*	5	Irregular*	Irrégulier*
15	Clear	Clair	8	Dark	Sombre	5	Vibrating	Vibrant
15	Round	Rond	8	Quiet	Calme	5	Constant*	Constant*
15	Rich	Riche	8	Clean	Net	5	Aggressive	Agressif
14	Nasal	Nasal	8	Calm	Calme	5	Heavy	Lourd
14	Full	Plein	8	Harsh	Rêche	5	Complex	Complexe
13	Hard	Dur	7	Shrill	Criard	5	Dynamic*	Dynamique
11	Weak	Faible	7	Short*	Court*	5	Natural	Naturel
10	Slow*	Lent*	7	Powerful	Puissant	5	Empty	Creux
10	Fast*	Rapide*	7	Metallic	Métallique	5	Far	Lointain
10	Even*	Régulier*	7	Open	Ouvert	5	Edged	Tranchant
10	Warm	Chaud	6	Ringing	Sonnant			

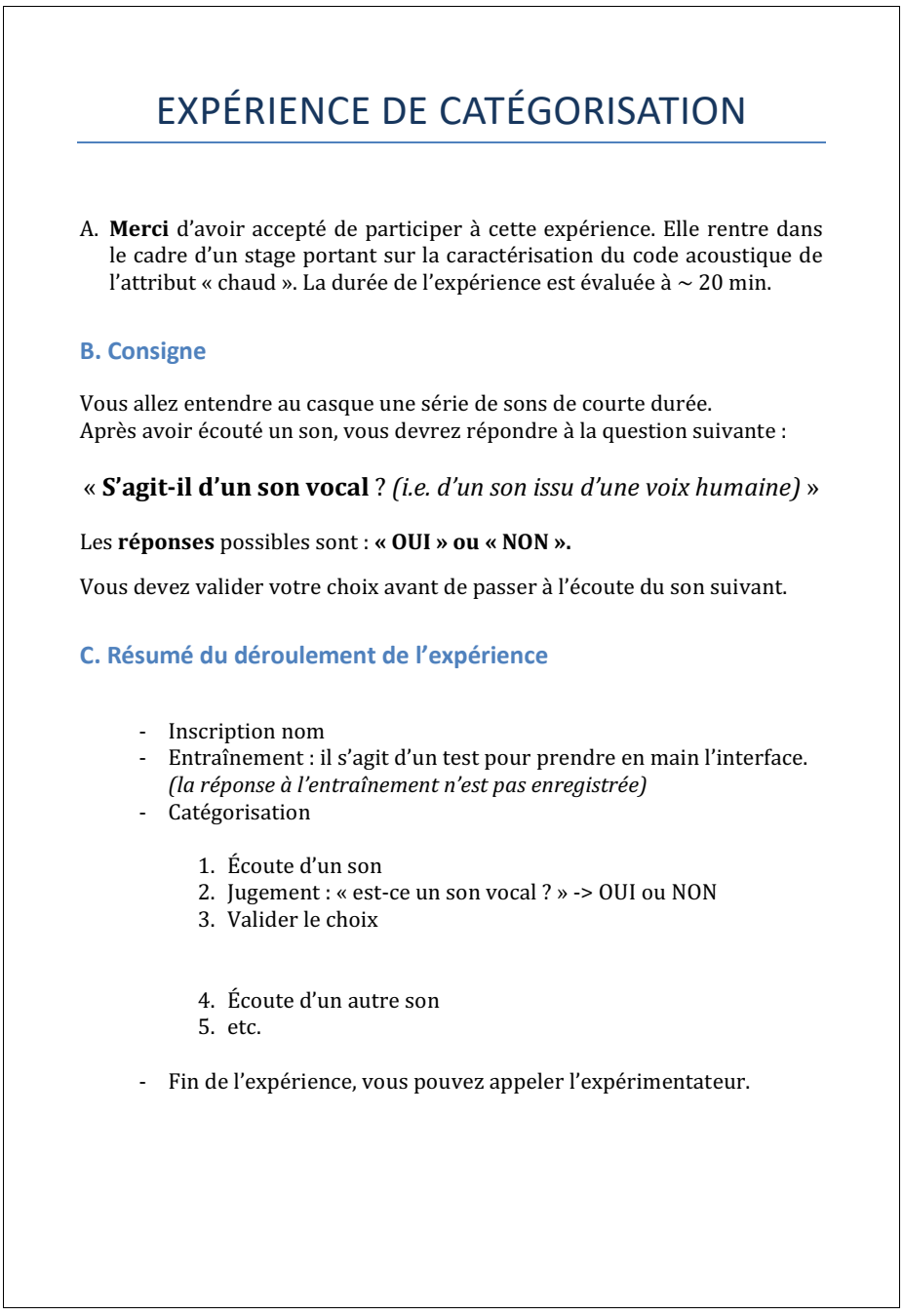
Texte

FIGURE A.1 – Tableau contenant les trente-cinq mots de vocabulaire sonore par ordre d'occurences (occ.) décroissant. « * » désigne les mots portant sur l'aspect temporel du son [3].

Annexe B

Présentation de l'expérience

B.1 Consigne de l'expérience de catégorisation



EXPÉRIENCE DE CATÉGORISATION

A. **Merci** d'avoir accepté de participer à cette expérience. Elle rentre dans le cadre d'un stage portant sur la caractérisation du code acoustique de l'attribut « chaud ». La durée de l'expérience est évaluée à ~ 20 min.

B. Consigne

Vous allez entendre au casque une série de sons de courte durée.
Après avoir écouté un son, vous devrez répondre à la question suivante :

« **S'agit-il d'un son vocal** ? *(i.e. d'un son issu d'une voix humaine)* »

Les **réponses** possibles sont : **« OUI » ou « NON ».**

Vous devez valider votre choix avant de passer à l'écoute du son suivant.

C. Résumé du déroulement de l'expérience

- Inscription nom
- Entraînement : il s'agit d'un test pour prendre en main l'interface.
  (la réponse à l'entraînement n'est pas enregistrée)
- Catégorisation

  1. Écoute d'un son
  2. Jugement : « est-ce un son vocal ? » -> OUI ou NON
  3. Valider le choix

  4. Écoute d'un autre son
  5. etc.

- Fin de l'expérience, vous pouvez appeler l'expérimentateur.

FIGURE B.1 – Consigne de l'expérience préliminaire de catégorisation donnée aux participants.

B.2 Origine des sons de synthèse

Liste des résonateurs Modalys utilisés

— Plaque circulaire aux conditions aux limites encastrées et libres : `free-circ-plate` et `clamped-circ-plate`
— Tube à perce conique ouvert/ouvert (O/O), fermé/ouvert (C/O) et fermé/fermé (C/C) : `open-open-tube`, `closed-open-tube` et `closed-closed-tube`
— Barre rectangulaire `rect-free-bar`
— Membrane circulaire et rectangulaire `circ-membrane` et `rect-membrane`.

Paramètre	Plaque circulaire	Tube O/O	Tubes C/O et C/C	Barre rect.	Membrane circ.	Membrane rect.
`modes`	150	200, 255, 500	150, 300, 330, 350, 380, 700	150, 300	1000	1000, 1300
`radius`	*	0,7–0,07 et 1–0,7	0,7–0,07 1–0,7 et 1–0,3	-	*	-
`thickness`	4	-	-	1	-	-
`density`	10^7	-	-	200	-	-
`poisson`	$0,45$	-	-	3	-	-
`young`	$1,36.10^8$	-	-	7.10^{30} et 7.10^3	-	-
`freq-loss`	5	7, 10	7, 10	0,7	0,2	0,2
`const-loss`	$0,1$	3,35	1,35	1,5	0,02	0,2
`width`	-	-	-	-	-	-
`length`	-	*		*	-	*
`tension`	-	-	-	-	10^5	10^5

TABLE B.1 – Paramètre des résonateurs Modalys utilisés pour générer les sons de synthèse. « * » signifie que le paramètre est variable selon la hauteur (*pitch*) imposée au résonateur, « - » que le paramètre n'est pas spécifié.

Voyelle	Résonateur	Hauteur (*pitch*) du résonateur (Hz)	Nouveau corpus
[a] homme	Plaque circulaire encastrée	10, 15, 40, 70 et 110	70, 110, 200, 400
	Plaque circulaire libre	30, 45	50, 70
	Membrane rectangulaire	90	336
[a] femme	Tube C/C	17, 20 et 30	350
	Tube C/O	80, 120	430
	Tube O/O	7	255, 150, 300
	Plaque circulaire encastrée	13	110
	Membrane rect.	50	336
[i] homme	Tube C/C	13, 40	90, 130
	Tube O/O	17	200
	Plaque circulaire encastrée	13, 15	70, 400
	Plaque circulaire libre	30	50
	Membrane rectangulaire	30, 50	500
[i] femme	Tube C/C	40, 50, 90	90, 500
	Tube O/O	10, 20	300
	Tube C/O	-	250
	Plaque circulaire encastrée	15, 30 et 40	70, 110, 250
	Barre rectangulaire	-	209

TABLE B.2 – Liste des fréquences fondamentales imposées aux structures résonantes en fonction des voyelles utilisées comme excitateur dans Modalys, avant et après ajustement du corpus.

Voyelle	Résonateur	Hauteur (*pitch*) du résonateur (Hz)	Nouveau corpus
[ou] homme	Tube C/C	70, 110	70, 110
	Tube O/O	10, 30	17, 30
	Plaque circulaire encastrée	10, 70	70, 350
	Plaque circulaire libre	40	50
	Barre rectangulaire	30	7
[ou] femme	Tube C/C	10	110, 15, 180, 250
	Tube O/O	-	150
	Tube C/O	10, 17	-
	Plaque circulaire encastrée	10, 50	70, 110
	Plaque circulaire libre	-	70
	Membrane rectangulaire	30	-

TABLE B.3 – (suite) Liste des fréquences fondamentales imposées aux structures résonantes en fonction des voyelles utilisées comme excitateur dans Modalys, avant et après ajustement du corpus.

Annexe C

Résultats : analyse et discussion

C.1 Filtres moyens du son « chaud »

C.1.1 Filtre moyen d'EQ par participant et par corpus

Les matrices composant la colonne de gauche de la figure 3.2 ont été obtenues par moyennage sur les sujets des matrices de filtre. On trace ces matrices individuelles, normalisées en RMS.

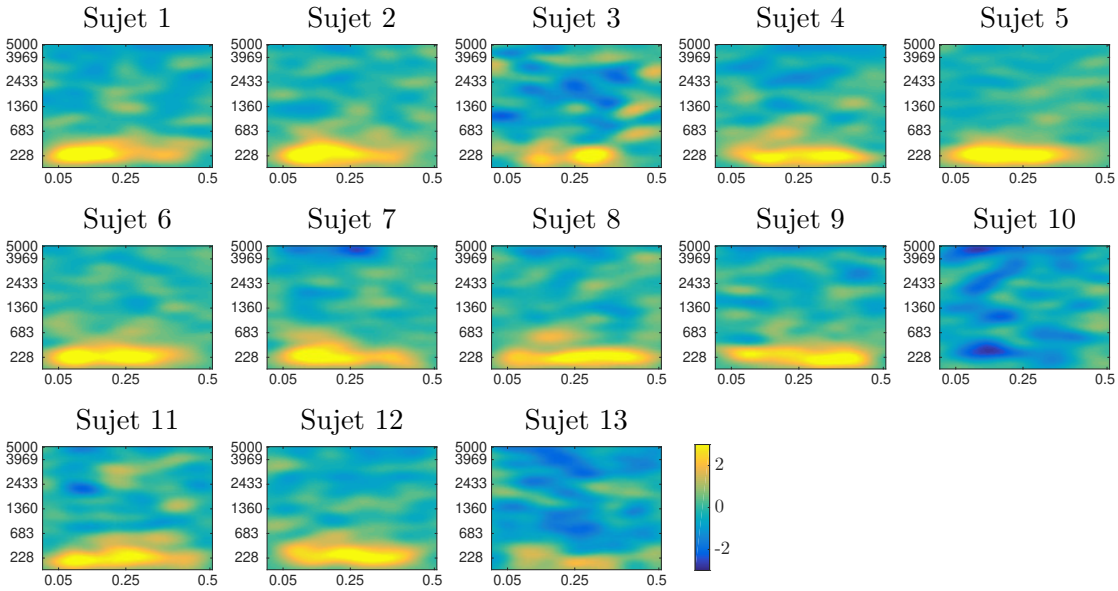


FIGURE C.1 – Filtres-moyens de « chaleur » d'EQ par participant pour le corpus de sons monosyllabiques normalisés en RMS.

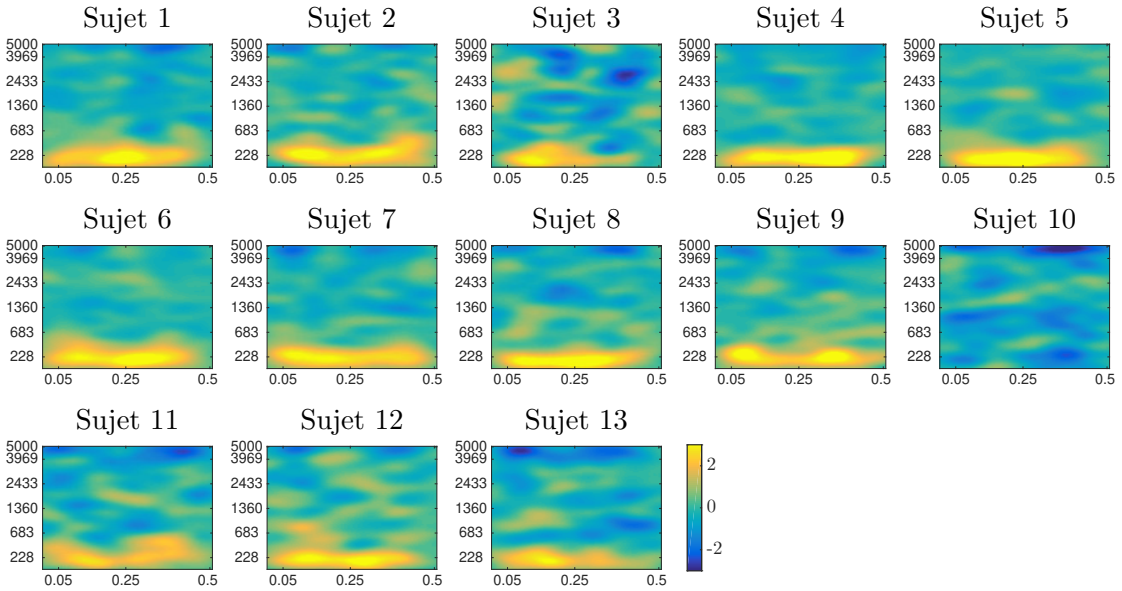


FIGURE C.2 – Filtres-moyens de « chaleur » d'EQ par participant pour le corpus de sons bisyllabiques normalisés en RMS.

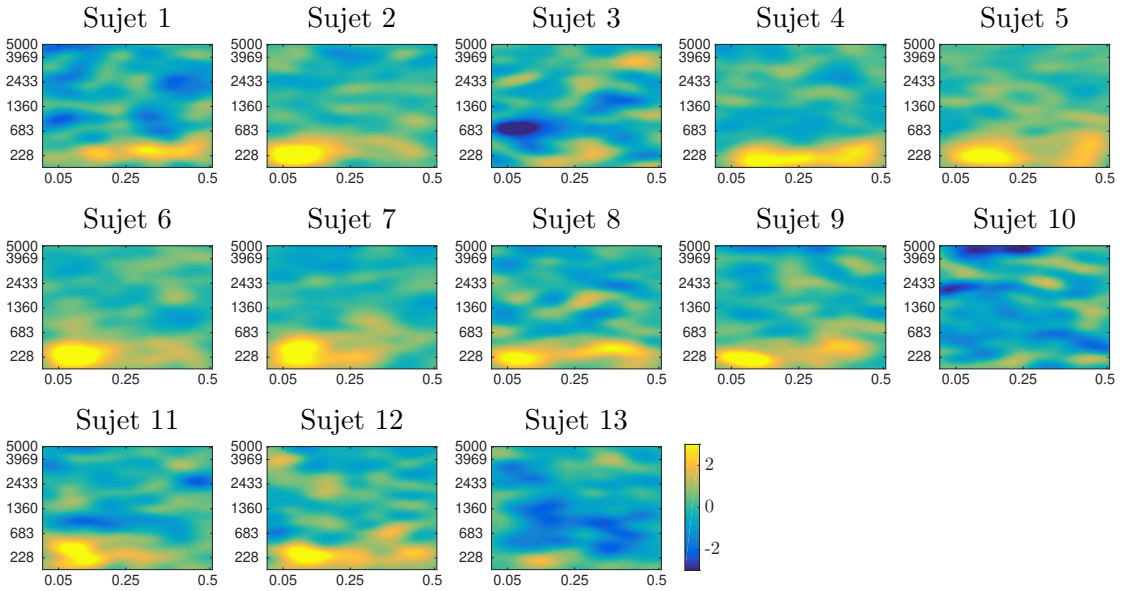


FIGURE C.3 – Filtres-moyens de « chaleur » d'EQ par participant pour le corpus de sons de synthèse normalisés en RMS.

C.1.2 Filtre moyen dans la dimension du HNR par participant et par corpus

Les matrices composant la colonne de droite de la figure 3.2 ont été obtenues par moyennage sur les sujets des matrices de filtre. On trace ces matrices individuelles, normalisées en RMS.

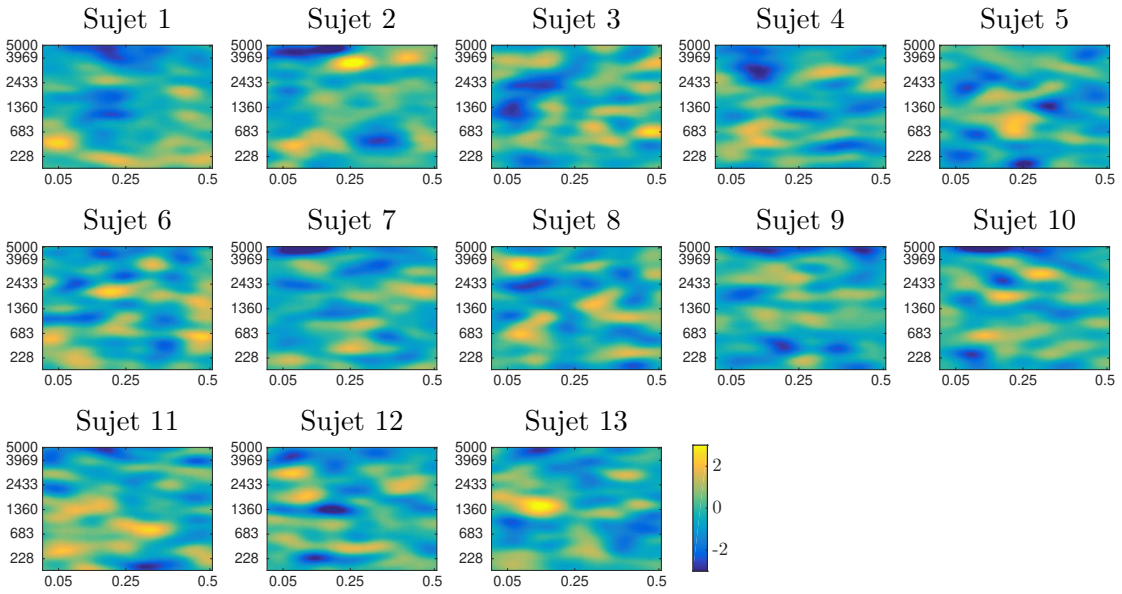


FIGURE C.4 – Filtres-moyens de « chaleur » dans la dimension du HNR par participant pour le corpus de sons monosyllabiques normalisés en RMS.

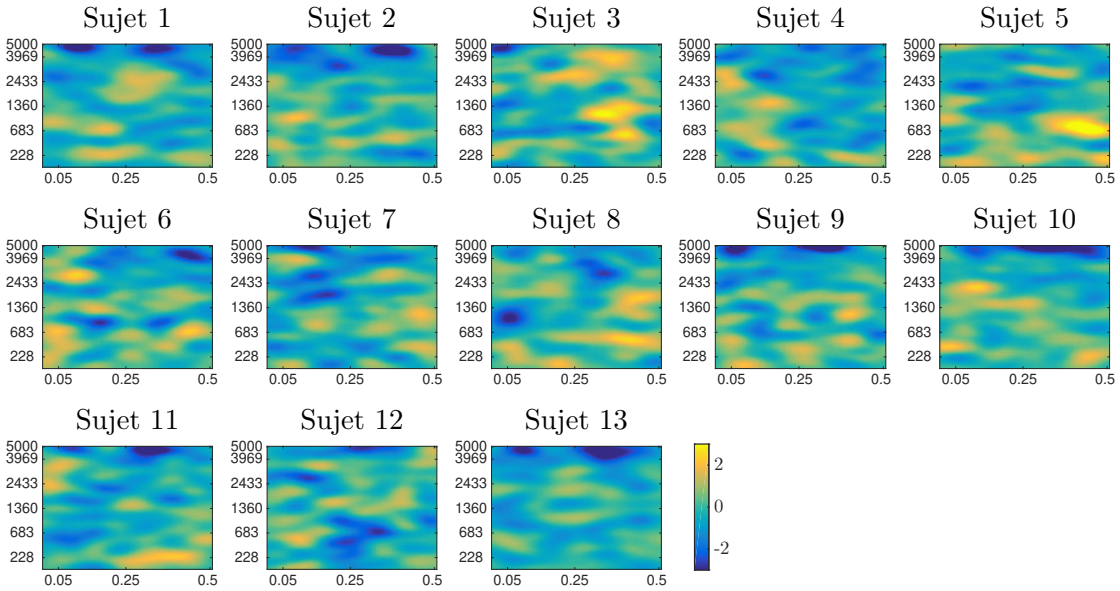


FIGURE C.5 – Filtres-moyens de « chaleur » dans la dimension du HNR par participant, corpus de sons bisyllabiques normalisés en RMS.

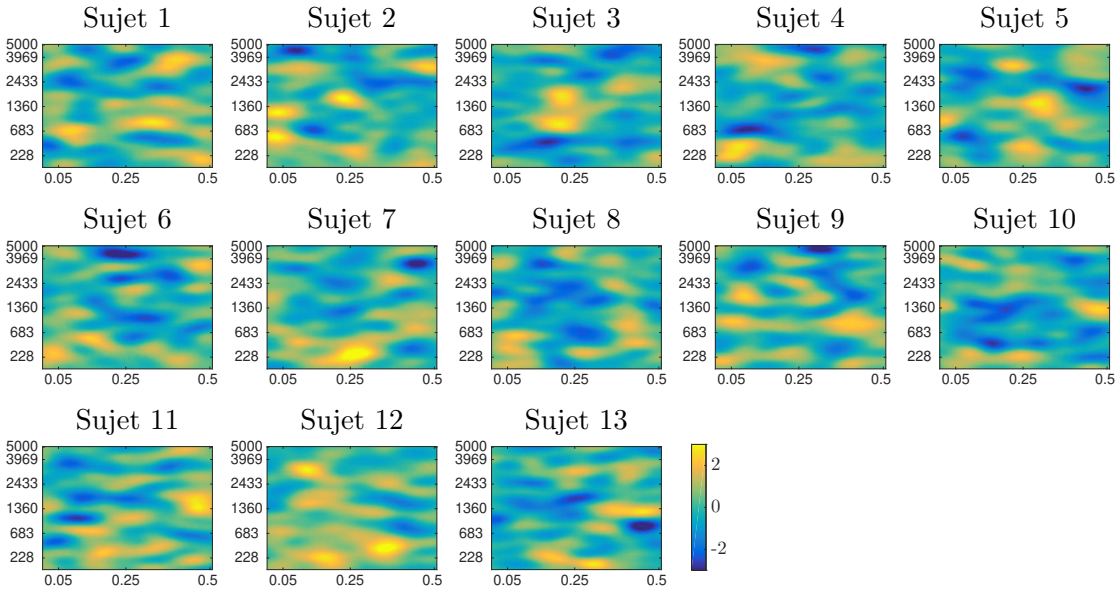


FIGURE C.6 – Filtres-moyens de « chaleur » dans la dimension du HNR par participant, corpus de sons de synthèse normalisés en RMS.

C.2 Définitions d'un son « chaud », données par les participants

Les participants ont écrit leur définition d'un son « chaud » avant la lecture de la consigne (cf section 2.3). Ces définitions sont retranscrites telles quelles.

- Un son « chaud » est un son riche en harmoniques. C'est un terme emprunté au visuel pour traduire un ressenti perceptif sonore. Son chaud $\leftrightarrow$ son analogique (pas pour moi, mais selon les audiophiles).
- Un son agréable à l'écoute, enveloppant, riche en basses fréquences et en harmoniques paires...
- Un son dont le spectre est assez fort autour des basses-mediums. Pas de formes d'ondes carrées, triangulaires... se rapprocherait d'un son pur ?
- Un son « chaud » peut être un son riche à certaines fréquences, qui le rendent "particulier", pas "transparent".
- « chaud » implique agréable. « chaud » est lié à une forme de proximité ou en tout cas à un effet de proximité non commun (comme quand on murmure à l'oreille). Il y a donc une surexpression des basses. Une certaine dose de saturation, harmoniques dans les basses fréquences.
- C'est un son qui n'est pas agressif, qui donne une sensation « naturelle » avec des graves « beaux » : bien définis, pas brouillons, et qui apportent une certaine énergie à l'écoute.
- Un son « chaud » est riche en bas mediums et en harmoniques aiguës (tout dépend du type de chaleur ?)
- Son généreux dans le bas médium et avec peu d'aigus.
- Son rond, présence de fréquences basses et médiums, son pas agressif, agréable à écouter, enveloppant.
- Chargé dans les graves, bas médium et milieu aigu ; agréable à écouter ; pas agressif ; continu.
- Souvent assimilé à « rond » ; agréable à écouter ; fréquences aiguës atténuées.
- Pour moi, un son « chaud » est un son accueillant, rond, avec une enveloppe assez douce. Il peut être musicalement harmonieux.
- Une bande fréquentielle accentuée entre 100Hz et 2kHz.
- Avec beaucoup d'harmoniques paires, sensation de son « plein » et chargé, assez pesant.
- Prédominance des basses et moyennes fréquences, saturation (légère) de celles-ci.

Table des figures

Liste des tableaux

Bibliographie

[1] P. Schaeffer, *Traité des objets musicaux*. Le Seuil, 1966.

[2] M. Carron, T. Rotureau, F. Dubois, N. Misdariis, and P. Susini, "Speaking about sounds : a tool for communication on sound features," *Journal of Design Research*, vol. 15, no. 2, pp. 85–109, 2017.

[3] M. Carron, *Méthodes et outils pour définir et véhiculer une identité sonore : application au design sonore identitaire de la marque SNCF*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2016.

[4] A. T. Sabin, Z. Rafii, and B. Pardo, "Weighted-function-based rapid mapping of descriptors to audio processing parameters," *Journal of the Audio Engineering Society*, vol. 59, no. 6, pp. 419–430, 2011.

[5] K. R. Scherer, "Judging personality from voice : A cross-cultural approach to an old issue in interpersonal perception," *Journal of personality*, vol. 40, no. 2, pp. 191–210, 1972.

[6] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[7] P. McAleer, A. Todorov, and P. Belin, "How do you say 'hello' ? personality impressions from brief novel voices," *PloS one*, vol. 9, no. 3, p. e90779, 2014.

[8] T. R. Agus, S. Paquette, C. Suied, D. Pressnitzer, and P. Belin, "Voice selectivity in the temporal voice area despite matched low-level acoustic cues," *Scientific reports*, vol. 7, no. 1, p. 11526, 2017.

[9] V. Isnard, *L'efficacité du système auditif humain pour la reconnaissance de sons naturels*. PhD thesis, Paris 6, 2016.

[10] J. Krimphoff, S. McAdams, and S. Winsberg, "Caractérisation du timbre des sons complexes. ii. analyses acoustiques et quantification psychophysique," *Le Journal de Physique IV*, vol. 4, no. C5, pp. C5–625, 1994.

[11] J. M. Grey and J. W. Gordon, "Perceptual effects of spectral modifications on musical timbres," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1493–1500, 1978.

[12] N. Obin, "Cries and Whispers - Classification of Vocal Effort in Expressive Speech," in *Interspeech*, (Portland, United States), pp. –, Sept. 2012.

[13] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," *Ircam*, 2004.

[14] E. Ponsot, J. J. Burred, and J.-J. Aucouturier, "Cracking the social code of speech prosody 63 3 65 cracking the social code of speech prosody using reverse correlation," *PNAS*, January 2018.

[15] E. Ponsot, P. Arias, and J.-J. Aucouturier, "Uncovering mental representations of smiled speech using reverse correlation," *The Journal of the Acoustical Society of America*, vol. 143, no. 1, pp. EL19–EL24, 2018.

[16] A. Ahumada Jr and J. Lovell, "Stimulus features in signal detection," *The Journal of the Acoustical Society of America*, vol. 49, no. 6B, pp. 1751–1756, 1971.

[17] L. Brinkman, A. Todorov, and R. Dotsch, "Visualising mental representations : A primer on noise-based reverse correlation in social psychology," *European Review of Social Psychology*, vol. 28, no. 1, pp. 333–361, 2017.

[18] C. F. Hellwag, *De formatione loquelae*. Heilbronn, 1781.

[19] N. Ellis, J. Bensoam, and R. Caussé, "Modalys demonstration," in *Proceedings of International Computer Music Conference*, pp. 101–102, 2005.

[20] J. J. Burred, E. Ponsot, J.-J. Aucouturier, and P. Belin, "Cleese ("ministry of silly speech")," March 2018.

[21] E. Ponsot, H. Dejardin, and E. Roncière, "Controlling program loudness in individualized binaural rendering of multichannel audio contents," in *Audio Engineering Society Convention 140*, Audio Engineering Society, 2016.

[22] P. Neri, "How inherently noisy is human sensory processing ?," *Psychonomic Bulletin & Review*, vol. 17, no. 6, pp. 802–808, 2010.

[23] E. R. Joosten and P. Neri, "Human pitch detectors are tuned on a fine scale, but are perceptually accessed on a coarse scale," *Biological cybernetics*, vol. 106, no. 8-9, pp. 465–482, 2012.

[24] A. Burgess and B. Colborne, "Visual signal detection. iv. observer inconsistency," *JOSA A*, vol. 5, no. 4, pp. 617–627, 1988.