Les émotions sont-elles exprimées de la même façon en musique que dans la voix parlée ?

RAPPORT DE STAGE

AUTEUR :
Daniel BEDOYA

Encadrants :
Jean-Julien AUCOUTURIER
Louise GOUPIL

Master ATIAM

Équipe Perception et Design Sonores
STMS UMR9912 (IRCAM/CNRS/Sorbonne Université)

05 février - 06 juillet

Année Universitaire : 2017 - 2018

Remerciements

Tout d'abord, j'adresse mes remerciements à mes encadrants : Jean-Julien Aucouturier et Louise Goupil qui m'ont donnée l'opportunité de participer dans ce projet, m'ont guidé pendant les phases de conception et passation de l'expérience, puis m'ont aidé avec l'analyse de données.

Je tiens à remercier l'équipe Perception et Design Sonores (notamment Nicolas Misdariis et Patrick Susini) de l'IRCAM pour son accueil ; pour m'avoir donné l'occasion d'acquérir de nouvelles connaissances et compétences, et en particulier au reste de membres de l'équipe CREAM (Lou Seropian, Laura Rachman, Pablo Arias, Marco Liuni et Vasiliki Zachari).

Je voudrais adresser ma gratitude aux membres de l'INSEAD (Hoai Huong Ngo, Jean-Yves Mariette, Germain Dépetasse) et aux participants de l'expérience.

Un grand merci à tous les stagiaires de l'équipe PDS : Rosalie Ollivier, Lucie Marignier, Pierre Rampon, Paul Saint-Aubert et Nicolas Haezebaert pour les conversations intéressantes et la bonne ambiance de travail.

Finalement, je tiens à remercier à toutes les personnes qui ont aidé avec la relecture et correction de ce rapport de stage, en particulier Bernarda Requena et encore une fois Jean-Julien Aucouturier.

Résumé et abstract

Résumé

La question de la relation entre la voix parlée et la musique a été vastement abordée dans la littérature. Cependant, on sait très peu sur la correspondance des indices émotionnels vocaux, appliqués à la musique. Plusieurs travaux ont montré que certaines caractéristiques acoustiques des signaux de voix parlée, par exemple, la fréquence fondamentale et les formants, sont impliquées dans la communication émotionnelle. On s'est donc intéressé pendant ce stage, à l'existence d'un lien similaire entre indices acoustiques et réponse émotionnelle, dans le cas de la musique.

L'étude est basée sur l'utilisation de 3 algorithmes de traitement du signal, *DAVID*, *ZIGGY* et *ANGUS*, à partir desquels on applique des modifications sur des dimensions perceptives du son : un décalage de hauteur (vers l'aigu, le grave et un vibrato), un décalage de formants (vers l'aigu et le grave), et une transformation de rugosité (modulation d'amplitude et addition d'harmoniques). Pour les employer, on construit un ensemble de 476 sons divisés dans des conditions musicales (voix chantée, voix chantée avec accompagnement musical, instrument avec accompagnement musical) et non-musicales (voix parlée, cris). Une expérience d'évaluation de l'*affect* perçu a été réalisé sur 66 sujets qui jugeaient les transformations sur des paires de stimuli. L'analyse statistique a été faite en utilisant la méthode rmANOVA (analyse de la variance avec mesures répétées) et les résultats respectifs nous permettent de montrer l'existence du lien cherché entre indices acoustiques et perception émotionnelle des auditeurs, avec des effets statistiquement significatifs sur toutes les transformations acoustiques.

Mots-clés : Émotion, musique, voix parlée, transformations acoustiques.

Abstract

The question of a relationship between speech and music has been broadly studied in literature. However, very little is known about the correspondence of vocal emotional indicators, applied to music. Several scientific works have shown that certain acoustical characteristics in speech signals, such as fundamental frequency and formants, are involved in emotional communication. During this internship, attention has been given to the existence of a similar link between acoustic properties and emotional response, in the particular case of music.

The present study uses 3 signal processing algorithms, *DAVID*, *ZIGGY* and *ANGUS*, with which modifications on perceptive sound dimensions were applied : pitch shifting (upwards, downwards and vibrato), formant shifting (upwards and downwards) and roughness transformation (amplitude modulation and added harmonics). To employ them, an ensemble of 476 sounds was created and divided into musical conditions (singing, singing with musical background, instrument with musical background) and non-musical conditions (speech, screams). An experiment of perceived affect was conducted on 66 subjects who judged the transformations over pairs of stimuli. The statistical analysis was done by using rmANOVA (repeated measures analysis of variance) and the respective results show the existence of the link between acoustic properties and emotional perception, with statistically significant effects, throughout every acoustical transformation.

Keywords : Emotion, music, speech, acoustic transformations.

Table des matières

Chapitre 1

Introduction

1.1 Présentation générale

La coexistence de la musique et du langage, tous deux uniques chez l'humain, a inspiré de nombreuses thématiques de recherche scientifique, allant de débats sur les origines évolutive du langage et de la musique chez l'être humain (Patel, 2010) jusqu'au traitement neuronal de l'information des signaux linguistiques dans une mélodie chantée (Schön, Gordon, & Besson, 2005). La question des émotions dans la musique et la voix est centrale à beaucoup de ces travaux. La recherche sur l'émotion dans la voix parlée [1] analyse l'information affective qui est communiquée par sa partie non-verbale. On peut se focaliser sur l'analyse des paramètres acoustiques du signal et comment leur variation entraîne des perceptions émotionnelles discernables. La relation entre la perception émotionnelle de la musique et celle de la voix parlée a été étudié par plusieurs auteurs (cf. 2.5).

Ce stage, dont la méthodologie emprunte aux domaines du traitement du signal audio, de l'acoustique et de la psychologie cognitive, a comme objectif principal d'examiner l'hypothèse selon laquelle **les indices acoustiques signalant des émotions vocales entraînent des perceptions émotionnelles similaires sur des signaux musicaux**.

Pour ce faire, nous avons utilisé une série d'outils de transformation de voix développés dans l'équipe d'accueil de ce stage, et les avons appliqués pour créer un ensemble de stimuli vocaux et musicaux finement contrôlés. Nous avons ensuite réalisé une expérience de perception acoustique sur N=66 participants (musiciens et non-musiciens), puis analysé ses résultats avec des outils statistiques afin de comparer les effets de ces transformations sur la voix parlée et la musique. Les résultats obtenus confirment l'hypothèse principale, et ouvrent plusieurs pistes de recherche supplémentaires, que nous détaillerons à la fin de ce rapport.

1.2 Organisme d'accueil

Destiné à la recherche scientifique et à la création musicale, l'Institut de recherche et coordination acoustique/musique (IRCAM) est une institution où convergent à la fois des chercheurs, des artistes, des enseignants et des étudiants. Le laboratoire de recherche scientifique qu'il heberge, STMS (Sciences et Technologies de la Musique et du Son), est une unité mixte de recherche (UMR9912), en co-tutelle entre l'IRCAM, le CNRS et Sorbonne Université.

Ce stage a été réalisé grâce à l'accueil de l'équipe Perception et Design Sonores (PDS), dans le cadre du Projet européen *Cracking the Emotional Code of Music* (CREAM) sous la direction de Jean-Julien Aucouturier. L'équipe PDS travaille dans le champ de la psychoacoustique, de la psychologie cognitive et du design sonore. Plus spécifiquement, le projet CREAM analyse les mécanismes cérébraux d'induction émotionnelle activés par la stimulation avec des signaux musicaux.

1. Speech emotion analysis

1.3 Contexte du stage

Lors des différents étapes du travail de recherche du projet CREAM, certains algorithmes qui manipulent le signal acoustique ont été développés (Rachman et al., 2018 ; Arias et al., 2018 ; Liuni, Ardaillon, Lou, Vasa, & Aucouturier, 2018), ayant pour but d'induire une réponse émotionnelle chez l'auditeur. Ces outils ont été principalement utilisés pour étudier des sons comme de la voix parlée ou des cris. Cependant, jusqu'à maintenant, on n'a pas examiné en détail une application purement musicale.
La littérature sur les mécanismes de codification perceptive de l'émotion dans la voix parlée et la musique nous permet d'envisager cette recherche, dans une approche expérimentale. On se concentre ici sur une correspondance générale de ces réponses émotionnelles dans la musique quelque soit son genre/style (et en pratique ici, dans le registre de la musique populaire occidentale), d'une façon qui est typique des approches dites de 'biologie de la musique' (Patel, 2010). Cette première approche pourrait être complétée ensuite avec un questionnement plus musicologique ou anthropologique, examinant les spécificités de certains genres ou de cultures musicales, mais ces approches sont en dehors du cadre de ce travail de stage.

1.4 Objectif du stage

Ce travail a pour objectif principal d'appliquer des outils de traitement de signal, développés lors du projet CREAM, à des sons musicaux et d'examiner une possible correspondance entre les mécanismes d'induction émotionnelle de la voix parlée et ceux de la musique. Plus spécifiquement, le travail du stage examine les indices acoustiques liés à la hauteur de la voix (une hauteur plus ou moins élevée), son inflexion (le vibrato), le timbre (le son d'une voix 'souriante') et les modulations d'amplitudes (la rugosité d'un cri de colère). Nous comparons ici ces indices entre la voix parlée et la voix chantée, avec ou sans accompagnement musical, mais aussi des pistes purement instrumentales. Une contribution méthodologique additionnelle de ce travail est la production d'un ensemble de stimuli sonores dont les caractéristiques acoustiques sont finement contrôlées pour son utilisation dans des contextes de recherche similaires.

1.5 Plan du rapport

Le présent rapport décrit l'ensemble des activités réalisées pour atteindre cet objectif. Nous décrivons tout d'abord les travaux de l'état de l'art qui abordent la question de la relation entre voix parlée et musique. Cette étape fournira un point d'entrée et de repère auquel on reviendra à la fin lors de l'interprétation des données. Le rapport s'attache ensuite à définir les différents paramètres acoustiques étudiés et les algorithmes utilisés pour modifier les sons. Nous détaillons ensuite chaque étape de la conception expérimentale et de la passation de l'expérience, puis les résultats obtenus et les analyses statistiques pertinentes. La dernière partie portera sur les conclusions obtenues lors de cette étude et les perspectives de recherche qui en découlent.

Chapitre 2

État de l'art

Ce chapitre ouvre sur quelques définitions acoustiques de base utilisées dans le reste du rapport, puis présente une synthèse de l'état de l'art sur le thème de l'émotion dans la voix et la musique. Ces éléments de support nous permettront de formaliser le problème à étudier.

2.1 Quelques définitions préliminaires

Un son donné (et en particulier les sons vocaux et musicaux qui nous intéressent ici) peut être caractérisé par des paramètres acoustiques, qui décrivent la physique de la vibration dans l'air, ainsi que des paramètres perceptifs, qui montrent comment ce son est perçu par un auditeur. La relation entre les caractéristiques physiques du son et ses attributs perceptifs est étudiée par la discipline scientifique de la psychoacoustique.

2.1.1 Définitions acoustiques

L'exemple plus simple est celui d'un son périodique simple, dit 'pur', qui n'est composé que d'une fréquence. Ce signal peut s'écrire comme une fonction du temps :

$$x(t) = A\cos(\omega t + \phi) \tag{2.1}$$

où A est l'**amplitude** en [Pa] ou [dB], t le temps en [sec], ϕ la **phase** en [rad] et $\omega = 2\pi f$ est la pulsation en [rad/s], d'où l'on obtient la **fréquence** f en [Hz]. La période du signal $T = 1/f$ en [sec], dépend alors de sa fréquence. Il existe aussi des sons périodiques complexes qui ont plusieurs composants fréquentiels. Le théorème de Fourier nous permet de montrer que l'on peut représenter ces signaux comme :

$$g(t) = \sum_{i=0}^{+\infty} A_i \cos(2\pi f_i t + \phi_i) = \sum_{i=0}^{+\infty} g_i(t) \tag{2.2}$$

La fréquence en $i = 0$ est appelé **fréquence fondamentale** et est souvent notée F0. Une notion importante à retenir est celle de **fréquences harmoniques**, les multiples entiers f_i de F0, qui jouent un rôle important en perception. Dans ce cas, chaque fonction $g_i(t)$ est appelée harmonique de rang i, d'amplitude A_i, de fréquence f_i et de phase ϕ_i.

Dans le cas de sons vocaux et de certains sons musicaux, le mode de production du son crée certaines résonances visibles dans l'enveloppe spectrale du son. Dans le cas de la voix, ces résonances sont appelées **formants**, et sont liées à la taille et la forme du conduit vocal au moment de la phonation. La fréquence des premiers formants F1 et F2 nous donne une information suffisante pour caractériser quelle voyelle est prononcé par un sujet. Dans la voix masculine et féminine, on trouve ces fréquences de F1 et F2 entre 300 Hz et 2500 Hz environ (Rossing, 2007).

2.1.2 Définitions psychoacoustiques

Les principales dimensions psychoacoustiques étudiées dans ce document sont la hauteur, le timbre et la rugosité.

La **hauteur** est la dimension perceptuelle qui correspond au paramètre acoustique de la fréquence, qui prend en compte une organisation des sons sur un continuum allant de grave à aigu. En musique, l'auditeur utilise la hauteur pour discriminer entre la fréquence de notes musicales différentes, même si on n'a pas besoin d'identifier une fréquence ou note spécifique pour discerner un changement relatif de hauteur dans une mélodie. Dans la voix parlée, la F0 aide à inférer l'intention, l'émotion ou à reconnaître l'identité de la voix d'un locuteur (McPherson & McDermott, 2018).

Le **timbre** est un attribut perceptif multidimensionnel. Il est principalement lié au spectre du signal et permet de distinguer entre deux sons qui ont la même amplitude, la même durée et la même fréquence fondamentale (Rossing, 2007).

Il y a plusieurs définitions de **rugosité**. On associe souvent ce concept au percept évoqué quand des composants fréquentiels proches créent des battement rapides, car ils stimulent des régions contiguës sur la membrane basilaire dans la cochlée (Plack, 2010). Dans cette étude, nous définissons la rugosité comme une sensation de fluctuation que l'on peut imiter artificiellement en modulant un signal en amplitude avec une intensité donnée. Parizet (2006), signale trois aspects importants sur la recherche de Zwicker et Fastl :

— La force de fluctuation est maximale pour une fréquence de modulation d'amplitude de 4 Hz.
— La rugosité est maximale pour une fréquence de modulation entre 30 Hz et 70 Hz, selon la fréquence du son modulé.
— Cette rugosité est d'autant plus perçue que la fréquence du son modulé est proche de 1 kHz.

2.2 Psychologie des émotions

L'étude des émotions humaines dans les sciences cognitives est un domaine de recherche qui remonte au XIXe siècle (Darwin, 1872 ; James, 1884), et qui a donné lieu à de nombreuses définitions et de nombreux débats sur la notion même d'émotion. Sans prétendre à l'exhaustivité, nous décrivons ici quelques-uns des axes qui organisent ce domaine.

2.2.1 Affect central, émotion et humeur

Il existe de multiples termes pour parler d'émotion, et la communauté des sciences cognitives s'accorde aujourd'hui à différentier 3 états affectifs principaux (Ekkekakis, 2013).

Affect central

L'affect central ou '*core affect*' est l'ensemble des processus émotionnels élémentaires accessibles de manière consciente dans l'expérience subjective de tous les jours. C'est la dimension la plus générale, et qui enveloppe donc la plupart de phénomènes émotionnels ressentis consciemment. La psychologie cognitive s'accorde à lui prêter 2 dimensions :

— Valence. Peut être définie comme le spectre entre une sensation de plaisir ou de déplaisir.
— Activation. Est associée à l'étendue des sensations d'excitation entre faible et forte.

Il faut remarquer qu'une expérience neutre est aussi une réponse affective. On peut alors utiliser les deux dimensions d'affect pour avoir une idée du type des sensations d'une personne à un moment donné avec la carte conçue par Russell, Weiss, et Mendelsohn (1989), voir Figure 2.1.

Émotion

Contrairement au '*core affect*' qui la sous-tend, l'émotion n'est pas un processus unique mais le produit de plusieurs composants ; c'est une réponse à un élément déclencheur (dans l'exemple donné par James, je tombe nez à nez avec un ours) qui est évaluée par rapport à lui. La réponse émotionnelle est complexe, courte mais intense, et entraîne : des expériences affectives (ex. j'ai peur) ; des processus cognitifs (ex. mon attention est

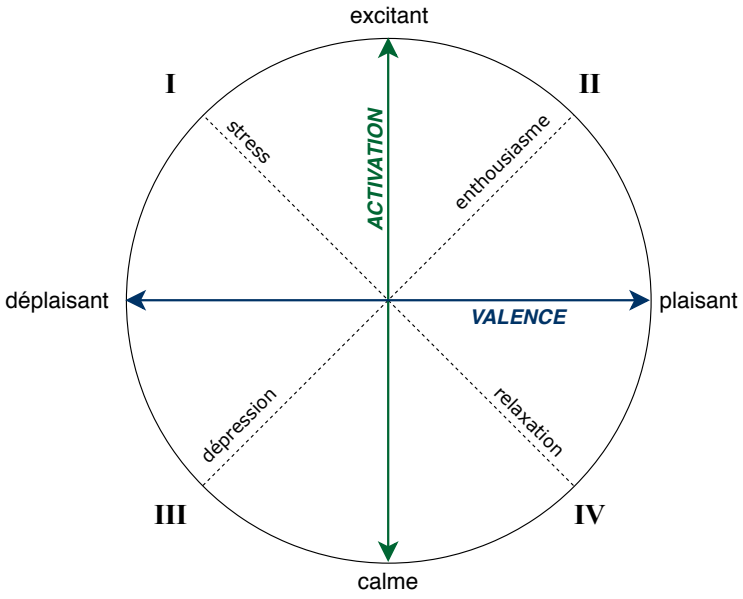


FIGURE 2.1 – Cercle des dimensions affectives, adapté de Russell et al. (1989) et Watson et Tellegen (1985). Les quadrants représentent les différentes combinaisons d'affect dans un continuum : **I.** Valence faible, Activation élevée, **II.** Valence et Activation élevées, **III.** Valence et Activation faibles, **IV.** Valence élevée, Activation faible. Le centre illustre une sensation affective neutre.

augmentée, je cherche un moyen de m'échapper) ; des réglages physiologiques pour répondre aux circonstances (ex. mon rythme cardiaque augmente) ; et des réponses comportementales expressives, adaptatives et orientées vers un objectif (ex. je crie, je m'enfuis) (Ekkekakis, 2013).

Humeur

La différence principale entre émotion et humeur est sa durée. Alors que les émotions sont des événements ponctuels, le concept d'humeur est lié à une composante affective qui, au contraire, dure plus longtemps (quelques heures ou quelques jours). Une humeur est souvent assez vague et on ne peut pas cibler une cause spécifique qui la déclenche, sauf pour des événement majeurs.

2.2.2 Approche catégorielle vs dimensionnelle

Il existe un débat entre les différentes manières d'étudier l'émotion, ce qui rend difficile la mise en place de définitions communes pour la recherche dans ce domaine. Deux tendances principales s'opposent, qui sont à la fois des propositions théoriques (sur comment les émotions sont représentées cognitivement) et méthodologiques (sur comment on doit les mesurer expérimentalement).

L'organisation catégorielle (Ekman, 1992) des émotions suppose un ensemble de quelques émotions de base, inhérentes à l'être humain ; reconnaissables et présentes de manière universelle. Ces émotions sont la joie, la tristesse, la colère, la peur et le dégoût (auxquels on ajoute parfois la surprise). Il s'agit d'une théorie simple qui présuppose que ces catégories ont émergé au cours de l'évolution à cause d'évènements récurrents (chercher de la nourriture, défendre un territoire, etc.), qui appellent des réactions et des expressions relativement stéréotypées (Juslin & Laukka, 2003). Dans le cadre catégoriel, les émotions sont mesurées souvent en choix forcé (ex. joie vs tristesse), et reliées à des actions expressives stéréotypées (ex. la présence ou non d'un sourire sur un visage).

L'approche dimensionnelle (et/ou constructionniste, auquel elle est très liée) (Russell, 1980 ; Russell & Barrett, 1999) affirme de son côté que l'idée des empreintes émotionnelles est un mythe (Barrett, Lewis, & Haviland-Jones, 2016) ; les visages et corps seuls ne peuvent communiquer aucune émotion spécifique de manière consistante, et que nous avons besoin d'un contexte pour interpréter les signes qui peuvent être contradictoires, au premier abord, à l'émotion exprimée. Par exemple, on peut être heureux et pleurer en même temps. Dans ce cas, une personne qui ne connaît pas le contexte ne va pas être capable de 'reconnaître' l'émotion correcte dans l'abondance d'informations extériorisées. En fait, Scherer (2005) et Frijda et Scherer

(2009), cités dans Ekkekakis (2013) précisent qu'une émotion peut être le résultat d'une interaction complexe entre cinq composantes : le traitement cognitif de l'information, la neurophysiologie, l'exécutif, l'expressif et l'expérientiel. Dans le cadre dimensionnel, les émotions sont souvent évaluées sur des échelles continues (par exemple, valence et activation).

2.2.3 Émotion perçue vs émotion ressentie

Il n'est pas la même chose de ressentir une émotion, et de percevoir ou identifier une émotion, bien que parfois il peut il y avoir un recouvrement des deux. Le premier processus est automatique, et entraîne une réaction émotionnelle plus ou moins intégrée (un comportement, une modification physiologique, etc.). Le second processus est limité au traitement de l'information d'un stimulus émotionnel, souvent communicatif (je reconnais un visage comme étant joyeux ou triste...), mais ce traitement n'évoque pas nécessairement une réponse émotionnelle (... mais "ça me laisse froid"). Les rapports entre les 2 types de processus sont débattus, allant de l'indépendance complète (on peut ressentir une émotion sans avoir conscience du stimulus la déclenchant, Taschereau-Dumouchel et al., 2018) à l'identité (on ne peut pas identifier de stimulus émotionnel sans d'abord ressentir l'émotion correspondante, Stel & van Knippenberg, 2008).

2.2.4 Approche utilisée dans ce document

D'un point de vue théorique, nous nous intéresserons dans ce travail à la notion de '*core affect*', plutôt que sur celles d'émotion et d'humeur. D'une part, les stimuli vocaux et musicaux utilisés dans cette étude sont nécessairement d'une courte durée (quelques secondes ou dizaines de secondes) et il semble donc peu pertinent de comparer leur effet à l'échelle de l'humeur. D'autre part, il existe un débat sur le fait que les émotions musicales sont des émotions au même sens que les émotions 'écologiques' dans la voix et les visages (Putman, 1987). Il est possible qu'une musique joyeuse ne soit pas 'joyeuse' au même sens qu'un autre être humain puisse l'être : Qui exactement exprime la joie dans la musique ? L'instrumentiste, le compositeur, la musique elle-même ? De même, il est possible que les réactions émotionnelles engendrées par la musique ne soient pas de même nature que celles déclenchées par des stimuli vocaux : Dois-je avoir 'peur' d'un violoncelle de la même façon que j'aurais peur d'un ours ou d'un congénère ? Notre étude vise à comparer l'expression émotionnelle dans la voix et la musique, sans avoir à trancher ces débats. Il semble donc risqué de chercher à comparer voix et musique sur une base aussi spécifique que la notion d'émotion, et préférable de se concentrer sur la dimension générale d'affect central, plus à même de se généraliser entre la voix et la musique. Notons que, par simplicité et par convention, nous garderons néanmoins les termes 'émotion' ou 'émotionnel' pour parler de cette notion d'affect central par la suite.

D'un point de vue méthodologique, nous allons utiliser l'approche dimensionnelle, et mesurer les émotions perçues sur des échelles continues de valence et d'activation.

Enfin, notre étude s'intéresse à la notion d'émotion perçue, et non ressentie. En effet, notre question est principalement une question de traitement d'information : il s'agit d'examiner si les indices acoustiques porteurs d'émotion dans la voix communiquent ces mêmes émotions quand ils sont appliqués à des sons musicaux. Il serait intéressant de voir si la transposition de ces indices à des sons musicaux entraîne plus, ou moins, d'induction émotionnelle, par exemple en comparant les réponses physiologiques des auditeurs, mais cette question dépasse le cadre de notre étude.

2.3 Émotion dans la voix

Les travaux qui parlent d'émotion dans la voix peuvent être catégorisés selon la valeur linguistique des vocalisations. Certaines études portent sur les expressions émotionnelles dans les vocalisations non-linguistiques (ou non-verbales), par exemple les cris, rires, pleurs ou grognements (Laukka et al., 2013 ; Sauter, Eisner, Ekman, & Scott, 2010). D'autre études portent sur les vocalisations linguistiques, quelles soient simples (phonèmes ou syllabes) (Fan & Cheng, 2014) ou complexes (mots, phrases) (Bryant & Barrett, 2008).

Ces études sont souvent menées en analysant les propriétés acoustiques du signal vocal. Bien que l'on puisse employer beaucoup d'attributs acoustiques de la voix pour la recherche, généralement on n'a de résultats consistants que pour très peu d'entre eux. En particulier, la fréquence fondamentale F0, et l'amplitude moyenne, qui permettent de reconnaître des caractéristiques importantes de la voix du locuteur, ont été largement

étudiées. Au niveau technique, le modèle source filtre a été très utile pour étudier ceci dans la voix parlée (comme on le verra en 3.2).

Il existe un lien bien établi entre l'activation physiologique et les traits acoustiques du signal vocal (Bachorowski, 2008, cité dans Lewis, Haviland-Jones, & Barrett, 2008), et notamment une corrélation positive entre l'activation et l'amplitude RMS de la vocalisation. D'autres résultats indiquent que les traits acoustiques de la voix parlée reflètent également la valence de l'état émotionnel du locuteur (Laukka, Juslin, & Bresin, 2005). Par exemple, une étude récente où l'on manipule le signal a montré une relation entre une fréquence fondamentale plus aiguë et une valence positive (Rachman et al., 2018).

D'une manière similaire, Ma et Thompson (2015) ont examiné la relation entre des états émotionnels (valence et activation) et les propriétés acoustiques du signal à partir des modifications de la fréquence, l'intensité et la vitesse, dans 24 sons environnementaux. Leurs résultats suggèrent que les émotions humaines sont très sensibles aux fluctuations dans leur environnement acoustique et permet de faire le lien avec la question d'un processus émotionnel commun entre la musique et la parole. En plus, cet étude a une méthode expérimentale avantageuse dont on se servira par la suite pour creuser cette relation.

2.4 Émotion dans la musique

La musique est une activité sociale dans toutes les cultures et il y a des preuves très vastes (par exemple la littérature scientifique de nombreux auteurs cités dans Barrett et al., 2016 et Juslin & Sloboda, 2011) qui indiquent que la musique provoque une réponse émotionnelle chez les auditeurs. Néanmoins, l'étude scientifique de ce sujet est assez complexe car il existe une grande quantité de facteurs intervenants et différentes approches pour aborder la question.

En premier lieu, on peut s'interroger sur comment la musique entraîne des émotions. À ce propos, il y a deux postures qui diffèrent sur la délimitation de l'influence émotionnelle de la musique et sont encore le sujet de débats. La position cognitiviste postule qu'on peut dissocier entre reconnaître une émotion particulière dans un morceaux de musique et ressentir cette émotion. Par contre, la position émotiviste présuppose qu'une émotion est invariablement induite quand on écoute de la musique (Juslin & Sloboda, 2011). Quelque soit le cas, pour tenter de trouver la base des représentations émotionnelles évoquées par la musique, les sciences cognitives s'intéressent aux types de traitement de l'information qui provoquent une réponse émotionnelle, appelés mécanismes psychologiques. Juslin et Västfjäll (2008) et Juslin et Sloboda (2011) en ont proposé 7 :

- Contagion émotionnelle : La musique créerait des émotions en imitant les caractéristiques principales du comportement émotionnel, par exemple de la voix parlée ou des gestes expressifs. Ceci est en rapport avec la théorie mimétique (Scherer, 1986 ; Davis, 1994 ; Juslin & Laukka, 2003). Un partie de cette théorie présuppose que la musique a un pouvoir expressif très élevé car elle utilise les mêmes caractéristiques acoustiques que la voix en les manipulant d'une manière qui serait impossible pour la voix humaine (*super-expressive voice theory* ; Juslin, 2001).
- Réponse du tronc cérébral : la musique créerait des émotions en déclenchant des réponses prédéterminées pour certaines propriétés musicales (hauteur et intensité). Ceci est en accord avec la proposition d'une évaluation cognitive rudimentaire de la musique due à une basse 'puissance de calcul' associée (Lewis et al., 2008).
- Mémoire épisodique : Les émotions créées sont en relation avec des souvenirs de l'auditeur.
- Imagerie visuelle : Les émotions créées sont dues à des images mentales visuelles déclenchées par la musique.
- Conditionnement évaluatif : Les émotions résultent d'un appariement de la musique avec un autre stimulus émotionnel par le passé.
- Attentes musicales : Les émotions résultent de la réussite ou de l'échec des prédictions mentales sur l'évolution d'une structure musicale (ex. une résolution harmonique surprenante).
- Évaluation cognitive : Évaluation du déroulement musical par rapport aux aspirations subjectives de l'individu.

Tous ces mécanismes peuvent interagir dans une réponse émotionnelle à un morceau de musique donné.

En second lieu, on peut s'interroger sur quelles émotions sont accessibles à quel type de musique. D'après une interprétation de la théorie communicative des émotions [1] (Oatley & Johnson-Laird, 1987, 1996, 2011 dans

1. Les émotions ont l'objectif de permettre la communication à l'intérieur et entre groupes de personnes.

Barrett et al., 2016), la musique pure (hors d'un autre contexte ou stimulus) n'a pas de contenu propositionnel et donc ne peut pas exprimer autre chose que des émotions générales comme la joie, la tristesse, l'anxiété et la colère (voir par exemple l'essai '*Why music has no shame*' Putman (1987)). En plus, les émotions évoquées par la musique ne sont pas les mêmes que celles évoquées par d'autres formes d'art ou activités en général. Par exemple, la tristesse ressentie quand on écoute de la musique n'a toujours pas une correspondance avec un sentiment 'aversif' (sombre ou d'abattement) qui est présent avec cette émotion la plupart du temps. Selon cette approche, c'est seulement quand les facteurs comme les paroles et le contexte interviennent, que la musique est capable d'induire des réponses émotionnelles complexes ou bien de les intensifier. En partant de ce principe, des efforts pour déclencher des réponses émotionnelles simples, à partir de la configuration d'éléments musicaux, ont abouti aux critères suivants (d'après Bunt & Pavlicevic, 2001 ; Juslin, 2001 ; Johnson-Laird & Oatley, 2008, cités dans Barrett et al., 2016) :

- Joie : Tempo moyen, forte amplitude, intervalles de hauteur assez larges, gamme majeure, consonant.
- Tristesse : Tempo lent, faible amplitude, hauteur grave avec intervalles assez courts, légèrement dissonant.
- Anxiété : Tempo rapide, amplitude modérée, hauteur grave, gamme mineure, dissonant.
- Colère : Tempo rapide, amplitude forte, hauteur élevée, gamme mineure, dissonant.

Finalement, même si ces références sont utiles pour beaucoup d'applications, le problème avec ce point de vue est que les mécanismes psychologiques décrits au-dessus, sont présents pour n'importe quel type de musique et empêcheraient l'étude de la musique à son état 'pur' ; on a donc encore un obstacle dans les définitions. À présent, on peut constater que la recherche est encore en cours de développement et elle est pleine d'opinions contraires qui ont ses points forts et faibles ; en espérant que les travaux à venir apporteront des preuves plus convaincantes.

2.5 Émotion dans la voix parlée et la musique

Dans cette section, on attire l'attention sur les parallèles des travaux présentés en 2.3 et 2.4, en se concentrant sur quelques études spécifiques qui examinent le lien entre la voix parlée et la musique, afin d'illustrer que leurs points communs constituent une base solide pour la recherche réalisée pendant ce stage.

Juslin et Laukka (2003) montrent, dans un article très cité, un lien fort entre la musique et la parole dans une importante méta-analyse de 104 études d'expression vocale et 41 études de performance musicale. On peut souligner les aspects suivants :

- La communication d'émotions discrètes est possible et efficace en expression vocale comme en expression musicale. Il y a des ensembles spécifiques de propriétés acoustiques qui sont utilisées pour communiquer des émotions.
- L'expression vocale des émotions est précise même de manière inter-culturelle. Cependant, les données sur la performance musicale sont insuffisantes.
- La capacité de décoder des émotions basiques en expression vocale et en performance musicale est développée assez tôt, au moins dès l'enfance.
- La performance musicale utilise largement les mêmes motifs spécifiques émotionnels des traits acoustiques que l'expression vocale.

Dans une approche expérimentale complémentaire à Juslin et Laukka (2003), Ilie et Thompson (2006) ont comparé l'effet émotionnel des transformations de paramètres acoustiques (fréquence, intensité et vitesse) dans des enregistrements de voix parlée et des morceaux de musique classique. Ils ont varié, par exemple, la hauteur sur tout le signal musical en 2 demi-tons vers le haut et vers le bas. Leur résultats indiquent un effet contraire à Juslin et Laukka sur les notes de valence pour la voix parlée et la musique (valence plus positive pour une musique diminuée en hauteur). Leur interprétation est basée sur des associations comme celle entre une musique avec des hauteurs graves et un sentiment plaisant.

Il y a d'autres efforts qui essaient d'élucider la ressemblance entre la musique et la voix parlée en regardant les particularités des gammes musicales. On a le cas de Curtis et Bharucha (2010), qui mettent en avant des correspondances entre les intervalles musicaux comme la tierce mineur (connue en musique pour être perçue comme triste) et sa présence plus importante dans la parole triste. Ils mentionnent que leurs résultats pourraient être des signes de l'existence d'une origine évolutive d'un proto-langage qui précéderait à la fois la musique et la parole. D'autres ont analysé la relation entre les propriétés spectrales des vocalisations humaines et la construction des intervalles dans les gammes musicales (Ross, Choi, & Purves, 2007 ; Bowling, Purves,

& Gill, 2017). De même, Escoffier, Zhong, Schirmer, et Qiu (2013), ont mesuré par imagerie par résonance magnétique fonctionnelle (IRMF) l'activité cérébrale à l'écoute de vocalisations et de la musique. Lors de tâches de reconnaissance d'expression émotionnelle, ils ont trouvé une superposition de cette activité cérébrale pour la musique et les vocalisations.

En conclusion, on constate donc que la littérature s'oriente vers la confirmation d'une connexion profonde, à plusieurs niveaux, de la musique et la voix parlée. Cette connexion est même postulée comme l'un des mécanismes fondamentaux selon lequel la musique créé des émotions (le mécanisme de 'contagion émotionnelle' de Juslin & Västfjäll, 2008). Toutefois, la plupart des preuves expérimentales apportées en soutien de cette théorie sont de nature corrélationnelles, établies en analysant d'une part des corpus linguistiques, d'autre part des corpus musicaux. La seule étude basée sur une manipulation du signal de stimuli musicaux, de nature à apporter une preuve causale (par exemple on monte la hauteur, et la valence ressentie augmente), est restée contradictoire, en particulier sur les liens entre hauteur et valence Ilie et Thompson (2006).

2.6 Problème de recherche

La présente étude vise donc à établir s'il existe une concordance mesurable entre les indices acoustiques exprimant des émotions dans la voix parlée et ceux exprimant des émotions dans la musique. Pour ce faire, à la manière de Ilie et Thompson (2006), nous utiliserons une méthode de manipulation du signal acoustique, de manière à établir une preuve de nature causale. Cependant, notre étude diffère de l'état de l'art sur plusieurs points importants :

1. Contrairement à Ilie et Thompson (2006), nous ne manipulerons pas l'intégralité du signal musical (par exemple transposer toute la texture musicale d'une tierce vers le haut), mais seulement la partie musicale considérée comme expressive (la partie de voix chantée ou de solo instrumental) et sur lequel portera le jugement de l'auditeur.
2. Nous n'utiliserons pas des transformations acoustiques de bas-niveau (hauteur, vitesse, énergie) communes à tous les signaux acoustiques (voix, musique, mais aussi sons de l'environnement Ma et Thompson (2015)), mais plutôt des transformations conçues pour être spécifiques à la voix parlée (par exemple le son d'un sourire, la rugosité d'un cri).
3. Nous testerons chaque participant à la fois sur des sons vocaux et des sons musicaux, afin de vérifier que les relations entre manipulations acoustiques et perceptions émotionnelles sont valides au niveau individuel. Ceci nous permettra également de comparer l'influence émotionnelle d'une transformation donnée sur la voix et la musique, et de tester expérimentalement la théorie de la '*super-expressive voice*'.
4. Enfin, nous comparerons plusieurs niveaux de 'transposition musicale' du signal vocal, d'abord à la voix chantée *a cappella*, puis à la voix accompagnée, puis enfin à la musique instrumentale.

Chapitre 3

Méthodes : Traitement du signal

Ce chapitre décrit les outils de traitement du signal utilisés pour manipuler nos signaux pendant ce stage. On décrira 3 algorithmes différents développés par l'équipe PDS à l'IRCAM, dans le cadre du projet CREAM. Étant donné que ces outils de traitement du signal ont été conçus pour la voix parlée, elles permettent de manipuler les paramètres acoustiques dont on a besoin pour évaluer la question principale du stage.

3.1 Algorithme de transformation de hauteur

DAVID (*Da Amazing Voice Inflection Device*) est un outil de transformation de voix en temps-réel qui permet de manipuler la voix d'un locuteur pour contrôler l'affect perçu à partir de paramètres acoustiques de la voix comme la fréquence fondamentale, l'inflexion et le spectre. Cet algorithme fonctionne sous forme de 'patch' du logiciel Max (Cycling'74).

DAVID a plusieurs paramètres qui sont utilisés pour modifier le signal. On ne décrit ici que la partie qui concerne le décalage de hauteur, car on n'utilise pas le reste des fonctionalités pour l'expérience.

3.1.1 Fonctionnement

L'algorithme, utilise le principe de lignes de retard (*delay lines*) décrit par Puckette (2007) d'un signal échantillonné $x[n]$. Ce signal est la représentation d'un sinusoïde (équation 2.1) de manière discrète comme une fonction de chaque échantillon n au lieu du temps t. On peut utiliser un décalage de longueur variable dans le temps $d[n]$ avec une longueur maximale D.

Si on dénote la position de l'échantillon d'entrée

$$y[n] = n - d[n] \tag{3.1}$$

alors, la sortie de la ligne de retard est :

$$z[n] = x[y[n]] \tag{3.2}$$

où le signal x est évalué au point $y[n]$, avec une interpolation si $y[n]$ n'est pas entier.

On ne peut pas décaler le signal au-delà de la longueur maximale D. Cette limite s'écrit donc :

$$n > y[n] > n - D \tag{3.3}$$

Par ailleurs, la formule de transposition [1] utilisée est :

$$\tau[n] = y[n] - y[n-1] = 1 - (d[n] - d[n-1]) \tag{3.4}$$

1. Appelée par Puckette *Momentary Transposition Formula* ; utilisée pour calculer la transposition des wavetables.

où τ est le multiple de transposition. Quand le décalage incrémente en fonction de n on transpose vers les graves, inversement, on transpose vers les aiguës quand $d[n]$ diminue. Ceci est connue comme l'Effet Doppler. L'utilisation de ce principe en combinaison avec les lignes de retard variables est appliqué pour décaler la hauteur du signal.

Il existe deux façon d'utiliser ce principe. D'une part, une modulation en fréquence du signal qui est toujours entre les bornes de l'équation 3.3 donne un vibrato. Elle utilise une fonction de décalage définie comme :

$$d[n] = d_0 + A\cos(\omega n) \tag{3.5}$$

où d_0 est le décalage moyen, A est l'amplitude et ω est la pulsation. La formule de transposition s'écrit de façon approximative :

$$\tau = 1 + A\ \omega\cos(\omega n - \pi/2) \tag{3.6}$$

et le décalage est entre $1 - A\omega$ et $1 + A\omega$.

D'autre part, si l'on veut maintenir une transposition constante on doit utiliser une fonction, par exemple un oscillateur de dent de scie, qui contrôle le décalage par intervalles de temps et aura deux objectifs : en premier lieu, fixer le temps de décalage en échantillons permettant de rester dans les limites d_0 et $d_0 + s$ où s est la taille de la fenêtre ; en deuxième lieu, modifier le gain de l'enveloppe, qui varie entre 0 et 1 tous les f/F_e (où F_e est la fréquence d'échantillonnage) selon la fréquence définie f en Hz et la position sur la forme d'onde. Si l'on note la sortie de l'oscillateur $x[n]$, on a :

$$x[n+1] - x[n] = \frac{f}{F_e} \tag{3.7}$$

et si on borne la sortie avec la fenêtre s, on obtient une nouvelle pente, où la constante d_0 n'a pas d'effet :

$$s \cdot x[n+1] - s \cdot x[n] = \frac{sf}{F_e} \tag{3.8}$$

La formule de transposition s'écrit maintenant :

$$\tau = 1 - \frac{sf}{F_e} \tag{3.9}$$

Finalement, on devrait ajouter une copie de l'oscillateur déphasée de $\pi/4$ et pour contrôler le changement de hauteur on modifie la fréquence avec s fixe. Pour un intervalle de transposition t, on écrit :

$$f = \frac{(\tau - 1)F_e}{s} \tag{3.10}$$

Des transpositions importantes risquent de produire des modulations d'amplitude audibles, qui peuvent être gérées en augmentant la taille de la fenêtre.

3.1.2 Validation expérimentale

Les travaux de (Rachman et al., 2018) ont validé qu'une manipulation de hauteur positive sur la voix parlée la fait paraître comme plus positive/joyeuse, et une hauteur moyenne plus faible comme plus négative/triste. D'autre part, l'utilisation du vibrato fait percevoir une voix comme plus anxieuse/effrayée. Ces relations sont stables dans au moins 4 langues (français, anglais, suédois et japonais), et les transformations sont perçues avec un degré de naturel comparable à celui de voix non-manipulées.

3.2 Algorithme de transformation des formants

ZIGGY (*Zygomatic induction*) est un outil de traitement du signal qui, en utilisant quelques modules d'analyse et de synthèse du logiciel Audiosculpt, permet de décaler les formants de la voix, imitant le changement des propriétés acoustiques générées lors de la production d'un son chez une personne souriante. Cet algorithme fonctionne sous la plate-forme Python 2.7.

3.2.1 Fonctionnement

Arias et al. (2018) ont développé cette approche qui applique une transformation directement sur de la voix parlée et non de la voix synthétisée ; en plus, elle préserve les partiels harmoniques originaux pour éviter des distorsions. On peut résumer le comportement de l'algorithme ainsi (voir aussi 3.1a) :

— Calculer l'enveloppe spectrale et obtenir la fréquence fondamentale F0. La méthode utilisée est le *True envelope* (Villavicencio, Robel, & Rodet, 2006), qui est une méthode améliorée de codage prédictif linéaire (LPC [2]), basée sur le lissage cepstral du spectre d'amplitude.

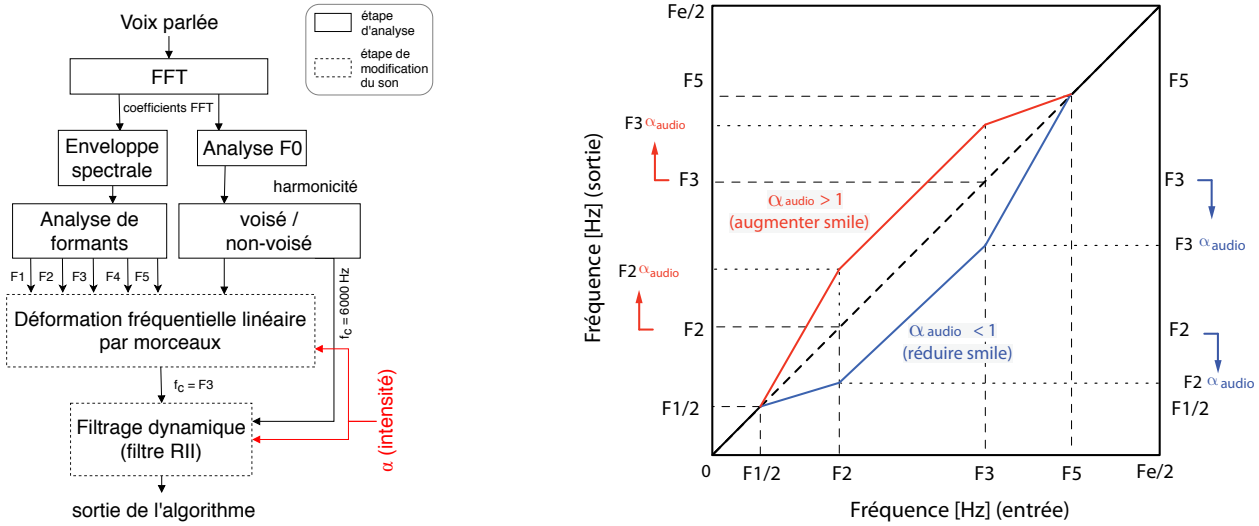
— Une fois que l'information de l'enveloppe est calculée, on peut faire l'analyse de formants (F_i) et obtenir la partie harmonique et non harmonique du signal, puis discerner entre du son voisé et non-voisé.

— On applique une fonction heuristique Φ de *frequency warping* (déformation fréquentielle) pour déplacer les formants vers le haut ou vers le bas avec l'objectif d'augmenter ou diminuer l'action de l'effet. La fonction suit une relation

$$f_{\text{sortie}} = \Phi(f_{\text{entrée}}, \alpha_{\text{audio}})$$

où α est le paramètre qui régule l'intensité de la transformation. On peut voir sur la Figure 3.1b que c'est une fonction par morceaux qui agit uniquement sur le signal contenu entre $F_1/2$ et F_5.

— Finalement, on effectue un filtrage dynamique et une re-synthèse aidée par la méthode du vocodeur de phase (Roebel, 2010).



(a) Étapes de l'algorithme de transformation de formants

(b) Correspondance de l'entrée et la sortie de la fonction de déformation fréquentielle

FIGURE 3.1 – Description du fonctionnement de l'algorithme Induction Zigomatique. Adapté avec permission de Arias et al. (2018).

Le paramètre α est assez sensible et ces changements ont des résultats variés, adaptables selon le signal vocal et par conséquent, les caractéristiques physiques du sujet. Néanmoins, il est important de noter 3 cas d'utilisation :

— Si $\alpha > 1$, l'algorithme déplace le spectre vers les fréquences aiguës, entraînant un effet de son souriant.

— Si $\alpha < 1$, l'algorithme déplace le spectre vers les fréquences graves, produisant l'effet contraire d'un son souriant.

— Si $\alpha = 1$, l'algorithme ne déplace pas le spectre. Alors $f_{\text{sortie}} = f_{\text{entrée}}$.

3.2.2 Validation expérimentale

Les travaux de Arias et al. (2018) ont validé que la transformation de sourire avec $\alpha > 1$ rend les voix manipulées plus souriantes et plus positives, et la transformation avec $\alpha < 1$ moins souriantes et plus négatives. Cette validation n'a été faite que sur des voix françaises, jugées par des auditeurs francophones.

2. *Linear Predicting Coding*. C'est une méthode de traitement du signal souvent utilisée comme un outil d'extraction de formants.

3.3 Algorithme de transformation de la rugosité

ANGUS ("*Highway to yell*") est un outil de transformation de voix en temps-réel qui permet de manipuler la rugosité de manière arbitraire à partir d'un modèle des plis vocaux qui génère des modes de vibrations instables. Ces vibrations produisent ensuite des sous-harmoniques et composants non-linéaires (absents dans le processus de phonation normal) perçus comme rugosité. Cet algorithme fonctionne sous forme de 'patch' du logiciel Max (Cycling'74).

3.3.1 Fonctionnement

Cet effet de rugosité est basé sur une modulation en amplitude du signal, piloté par la fréquence fondamentale F0, et un filtrage dans le domaine temporel pour sommer les sous-harmoniques. Par la suite, on décrit les étapes de la transformation, d'après Gentilucci, Ardaillon, et Liuni (2018).

On considère d'une part le signal porteuse $x_c(t) = A_c \cos(\omega_c t)$ avec une amplitude A_c et une pulsation ω_c et d'autre part le signal modulant $x_m(t) = 1 + h\cos(\omega_m t)$ avec une pulsation ω_m et une profondeur de modulation $h \in [0-1]$. Alors la modulation donne comme résultat :

$$y(t) = x_c(t) + y_+(t) + y_-(t) \tag{3.11}$$

où on trouve le signal original $x_c(t)$ et deux nouveaux signaux $y_+(t)$ et $y_-(t)$ définis comme :

$$y_+(t) = \frac{A_c h}{2}\cos((\omega_c + \omega_m)t) \quad \text{et} \quad y_-(t) = \frac{A_c h}{2}\cos((\omega_c - \omega_m)t)$$

avec amplitudes $\frac{A_c h}{2}$ et fréquences $\omega_c + \omega_m$ et $\omega_c - \omega_m$.

Si maintenant $x_c(t)$ est un signal de voix, représenté comme une somme de N sinusoïdes harmoniques $x_c(t) = \sum_{i=1}^{N} A_i \cos(i\omega_0 t)$ avec $\omega_0 = 2\pi f_0$, la modulation sera simplement la somme de chaque harmonique modulé individuellement :

$$y(t) = x_c(t) + \sum_{i=1}^{N} \left(y_{+_i}(t) + y_{-_i}(t)\right) \tag{3.12}$$

avec

$$y_{+_i}(t) = \frac{A_i h}{2}\cos((i\omega_0 + \omega_m)t) \quad \text{et} \quad y_{-_i}(t) = \frac{A_i h}{2}\cos((i\omega_0 - \omega_m)t)$$

On peut utiliser une somme de sinusoïdes comme signal modulant, pour générer davantage des sous-harmoniques :

$$x_m(\omega_0, t) = 1 + \sum_{k=1}^{K} h_k \cos\left(\frac{\omega_0}{k}t\right) \tag{3.13}$$

Cependant, ces sous-harmoniques créent une relation d'amplitude complexe et l'amplitude du premier sous-harmonique est trop élevée. Il est donc souvent préférable de les affaiblir ; on note que l'isolation des sous-harmoniques du signal est faite en soustrayant le signal original du signal modulé $y_{sh}(t) = y(t) - x_c(t)$ puis, on applique un filtrage passe haut pour obtenir $y_{\text{sh}}^{\text{HP}}(t)$.

Finalement, on multiplie les sous-harmoniques par un facteur de mixage $\beta > 0$ et on les somme au signal original.

$$y_{\text{rug}}(t) = x_c(t) + \beta y_{\text{sh}}^{\text{HP}}(t) \tag{3.14}$$

Cette méthode permet de générer des sons rugueux avec des sous-harmoniques stables, mais il est possible d'introduire un régime de bifurcation avec des variations temporelles du nombre de sous-harmoniques ; on peut aussi créer des signaux plus chaotiques avec du bruit ajouté.

3.3.2 Validation expérimentale

Les travaux de Liuni et al. (2018) (en préparation) ont validé que la transformation de rugosité rend les voix manipulées plus négatives et plus activées. Cette validation n'a été effectuée que pour des cris isolés (et non de la voix parlée), sur des auditeurs francophones.

Chapitre 4

Méthodes : psychologie

Ce chapitre décrit les méthodes de psychologie cognitive utilisées dans cette étude : conception et passation de l'expérience, puis analyse statistique.

4.1 Construction des stimuli

4.1.1 Choix

Le premier choix a été de ne pas se restreindre à un type de musique en particulier, mais de garder un aspect de validité 'écologique' sur le choix des stimuli, plutôt que de fabriquer artificiellement les expressions de chaque morceau, par exemple en les composant nous-mêmes.

On a donc travaillé sur 14 morceaux de musique tonale en format multi-piste (cf. Tableau A.2) qui utilisent des échelles diatoniques, appartiennent à différents sous-genres de la musique populaire et ont une mélodie chantée (7 voix d'homme, 7 voix de femme). Pour faire la sélection, on a utilisé prioritairement la ressource gratuite Mixing Secrets For The Small Studio (MSFSS) où l'on trouve des fichiers en format sans compression (`.wav`) séparés par genre. Cependant, on a aussi utilisé d'autres sources, notamment des multi-pistes de groupes connus (de licence fermée) pour profiter de leur meilleure qualité de production musicale.

Il était très important d'utiliser un format multi-piste afin de pouvoir manipuler les éléments musicaux comme la mélodie et l'accompagnement de manière individuelle (par opposition à Ilie et Thompson (2006) qui transformait tout l'ensemble) et de maintenir une seule session dans le logiciel d'édition audio pour construire et écouter les stimuli plus facilement.

Holt (2007) pose l'argument que, même si les limites entre les étiquettes des genres sont arbitraires et ne seront pas représentatives pour tout le monde, on peut se permettre de classifier la musique en genres pour des raisons pratiques. La liste affichée sur le site MSFSS, similaire à celle proposé par Holt sous le nom de genres historiques, est utile pour repérer les styles utilisés :

— Acoustique / Jazz / Country / Orchestral
— Électronique / Dance / Expérimental
— Pop / Chanteur-compositeur
— Alt Rock / Blues / Country Rock / Indie / Funk / Reggae
— Rock / Punk / Métal
— Hip-Hop / R&B

Dans le cadre de ce travail, on a choisi entre différents styles musicaux qui sont dans la grande catégorie dite de 'musique populaire', dont la pop, le jazz et le pop-rock. Ce choix permet d'explorer une plage de tempo, de tonalité et de timbre assez varié. La durée de chaque extrait a été limitée à un intervalle entre 3 et 10 secondes, selon la logique musicale du morceau (égale à une phrase musicale, généralement).

Tous les stimuli sont accessibles à l'adresse : `https://nuage.ircam.fr/index.php/s/UitetvZ3hXGylSi`

4.1.2 Enregistrement de voix parlées et de violon

Afin d'examiner un continuum de 'transpositions musicales' de stimuli parlés, nous avons d'une part, profité des paroles des 14 morceaux de musique originaux pour créer 14 phrases en voix parlée qui contenaient exactement la même signification linguistique que les stimuli musicaux. On a enregistré 2 femmes et un homme dont une française, une américaine et un américain. Pendant l'enregistrement on s'est assuré de garder une inflexion neutre et de ne pas imiter la hauteur des fréquences fondamentales des mélodies.

D'autre part, nous avons réalisé un enregistrement de toutes les mélodies (cf. Tableau A.3) avec un instrument de musique qui pourrait imiter la plupart de variations de hauteur (imperfections, liaisons, intention) des voix humaines. De cette manière, on a produit des nouveaux stimuli avec le même contenu musical, mais ne portant pas le message linguistique des paroles. On a donc enregistré un violon capté par un microphone DPA 4061 placé au-dessus du chevalet. Étant donné que la tessiture de cet instrument n'est pas exactement la même que la voix humaine (la note la plus basse étant un G3-196 Hz), l'instrumentiste a octavié les mélodies les plus graves pour être capable de les jouer. Ce compromis nous a permis de reproduire avec une très bonne qualité les mélodies de tous les morceaux de l'ensemble.

Chaque morceau était donc disponible en 4 versions : une version parlée (voix seule, même paroles, même langues et locuteur de même sexe que l'original), une version chantée a cappella (piste chant seule), une version chantée mixée avec accompagnement musical (par exemple le morceau original) et une version instrumentale (même mélodie que le chant, transposée au violon, mixée avec le même accompagnement orchestral que l'original).

4.1.3 Édition et mixage

Les logiciels utilisés pendant l'étape d'édition et mixage sont : Audacity, Audition, Max, Live.

Les signaux ont été normalisés en amplitude au niveau maximum à -3 dBFS, et fondus en entrée et en sortie. L'exportation des fichiers a été faite en format `.wav`, en résolution de 16 bits et avec une fréquence d'échantillonnage de 44100 Hz, en stéréo ou mono selon le cas.

4.2 Protocole expérimental

4.2.1 Participants

Basé sur la taille de l'effet trouvé par l'étude de Ma et Thompson (2015), l'échantillon a été constitué de $N = 66$ personnes adultes, dont 35 femmes et 31 hommes âgés entre 18 et 31 ans ($M_A = 23.24$ et $SD_A = 3.37$) [1]. Nous avons de plus décidé d'ajouter un critère de sélection (auto-déclaré) en divisant les participants en deux groupes de 30 musiciens et 36 non musiciens de la manière suivante : pour être classé comme musicien, le sujet doit avoir suivi plus de 3 ans d'études musicales en conservatoire ou équivalent.

4.2.2 Types de son

Dans l'ensemble des stimuli, il y a 5 types de son en total : 3 de type musical et 2 de type non-musical.

- Types de son musicaux :
  - Voix chantée (mélodie seule ou a cappella).
  - Voix chantée avec accompagnement musical.
  - Instrument principal (mélodie de violon) avec accompagnement musical.
- Types de son non-musicaux :
  - Voix parlée (paroles des 14 morceaux).
  - Cris de personnes (Stimuli utilisés pour l'étude de Liuni et al. (en préparation), avec l'autorisation des auteurs).

1. M_A est la moyenne d'âge. SD_A est l'écart type d'âge.

4.2.3 Types de manipulation

On a appliqué 4 types de manipulations aux stimuli sonores, en suivant les dimensions psycho-acoustiques décrites en 2.1. Ce sont :

Hauteur

Il s'agit d'une modification sur la fréquence fondamentale pour chaque type de son. On a utilisé l'algorithme DAVID (cf. 3.1) pour l'appliquer et il y a deux transformations possibles.

— Hauteur montante, appelé '***up***'. Décalage de +30 cents [2].
— Hauteur descendante, appelé '***down***'. Décalage de −25 cents.

Ces valeurs de transformation ont été déterminées de façon heuristique : une diminution de 30 cents était considérablement fausse par rapport à l'accompagnement musical non modifié ; au contraire, une augmentation de 25 cents n'était pas suffisamment saillante pour tous les stimuli.

On a décidé de rester en dessous d'un écart de ~ 100 cents, représentant un intervalle de seconde mineure pour ne pas changer les notes de la composition musicale. Les intervalles en musique ont leurs propres connotations émotionnelles (Curtis & Bharucha, 2010), par exemple la seconde mineure est perçue en musique occidentale comme le plus dissonant des intervalles.

La figure 4.1 montre que la transposition fonctionne mieux quand il s'agit de sons voisés, harmoniques (4.1c ou 4.1d) que quand elle transforme des sons avec un contenu non-harmonique important comme la voix parlée (4.1b). Logiquement, la voix chantée a un comportement intermédiaire (4.1a).

Formants

Cette transformation altère et tend l'enveloppe spectrale du signal à partir de l'application de l'algorithme d'induction zygomatique ZIGGY (cf. 3.2). Il y a deux transformations possibles :

— Décalage des formants vers les fréquences aiguës, appelé '***smile***', avec un paramètre d'intensité $\alpha = 1.1$.
— Décalage des formants vers les fréquences graves, appelé '***unsmile***', avec un paramètre d'intensité $\alpha = 0.85$.

On peut voir comment les formants sont décalés par l'algorithme sur la figure 4.2. Comme dans le cas précédent, l'estimation et la représentation des fréquences de formants dépend des caractéristiques du signal ; la détection des formants en fonction du temps pour un signal de parole est meilleure quand il y a davantage d'harmonicité dans ce signal.

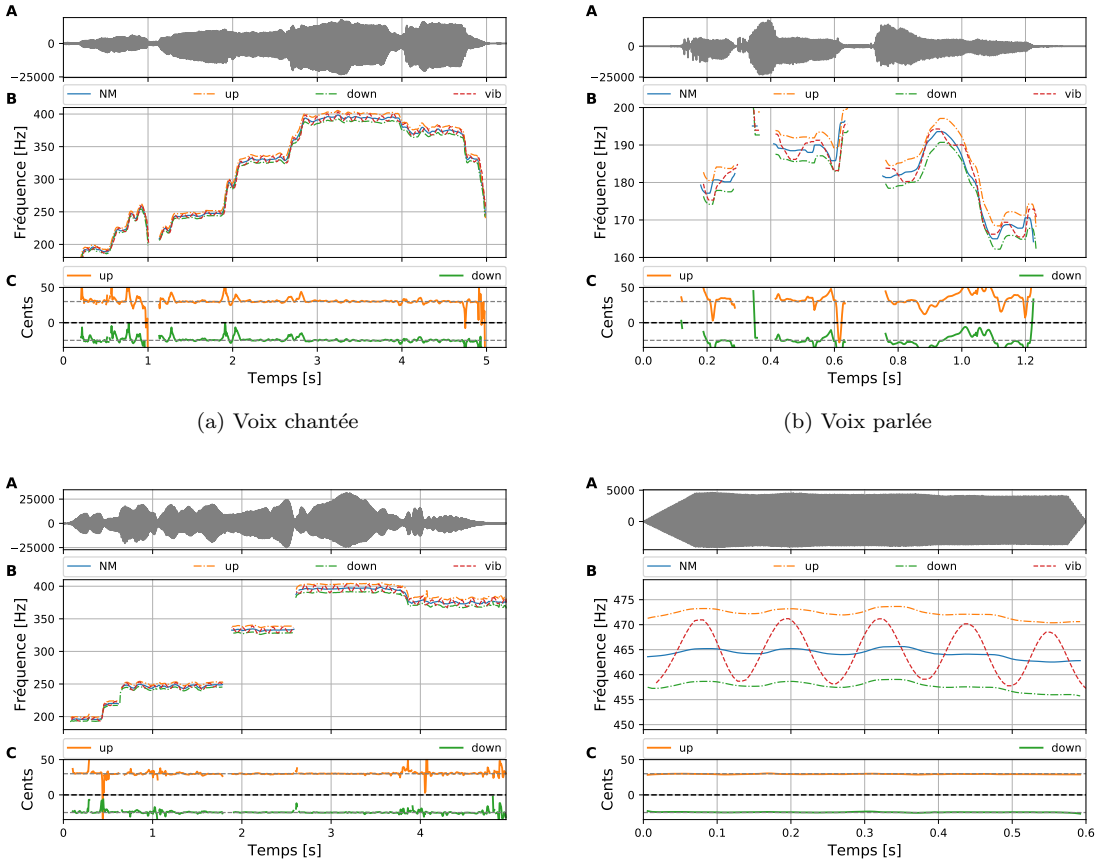
Vibrato

Modulation en fréquence d'une fréquence de modulation de 8 Hz et une amplitude de 25 cents autour de la fréquence fondamentale, avec des variations aléatoires de 20%. On s'est servi encore de DAVID et il y a une seule transformation possible appelée : '***vibrato***'. On peut visualiser la variation de la F0 du signal sur les traits discontinus en rouge de la figure 4.1.

Rugosité

Nous avons appliqué l'algorithme ANGUS pour créer un effet de modulation d'amplitude et fréquence (selon l'équation 3.11) avec 2 modulateurs, 0 sous-harmoniques, un lissage de 500 $[ms]$ et une durée de l'enveloppe de 7100 $[ms]$. Dans ce cas, il y a une seule transformation possible appelée : '***angus***'. Cette transformation peut être observée dans l'exemple de la figure 4.3 où l'on constate l'inclusion des harmoniques ajoutés par l'algorithme.

2. Un cent est définie comme la centième partie d'un demi-ton dans une échelle tempérée. On calcule cet écart entre les fréquences f_2 et f_1 à partir de la formule : $\text{cent} = 1200 \log_2 (\frac{f2}{f1})$.



(a) Voix chantée

(b) Voix parlée

(c) Violon

(d) Cri

FIGURE 4.1 – Changement de hauteur pour les 4 types de son (sans accompagnement musical). Les stimuli de voix chantée, voix parlée et violon correspondent aux paroles et aux pistes chantée et instrumentale du morceau N° 3 des stimuli. Le cri correspond au son N° 4.
A. Forme d'onde, **B.** Fréquence fondamentale, **C.** Variation de hauteur en cents.
Légende : NM (son non-modifié), *up* (hauteur montante), *down* (hauteur descendante), *vib* (vibrato).

4.2.4 Tâche

Les participants ont écouté des paires de sons, présentés avec un intervalle inter stimulus de 1 [s], et devaient juger l'émotion qu'ils exprimaient. Plus précisément, la tâche consistait à comparer l'émotion qui était exprimée dans l'enregistrement cible, par rapport à un extrait où l'on a appliqué la transformation contraire, par exemple la transformation de hauteur *up* vs *down*.
On a divisé les participants dans deux groupes : le groupe 1 devait évaluer le premier son par rapport au deuxième alors que le groupe 2 devait évaluer le deuxième son par rapport au premier. On a choisi de créer ces groupes et contrebalancer l'ordre de présentation de stimuli pour éviter un possible effet de préférence pour le son écouté le plus récemment.

Pour juger l'affect de ces sons, les sujets devaient indiquer leur réponse à l'aide de deux échelles de Likert (1932), gradées de 1 à 7, qui évaluaient la valence et l'activation et qui apparaissaient à l'écran après chaque écoute (cf. figure A.1). Avec la première échelle, ils devaient juger si l'enregistrement cible de la paire exprimait une émotion plus négative (triste/colérique/inquiet) ou plus plus positive (satisfait/joyeux/optimiste) que l'enregistrement modifié avec la transformation opposée. Avec la deuxième échelle, ils devaient juger si l'enregistrement cible de la paire était plus calme (serein/détendu/résigné) ou plus agité (excité/énervé/anxieux) que son opposé. S'ils ne trouvaient aucune différence entre les extraits, ils devaient choisir la valeur 4 (≪aucune différence≫). Cette dernière instruction prenait notamment en compte l'existence de la tâche contrôle qui consistait en présenter deux fois le même son non-modifié (NM), dans le but d'avoir un réponse de référence.

En total, il y avait 340 extraits à évaluer, avec la possibilité d'une pause de 5 minutes tous les 70 extraits.

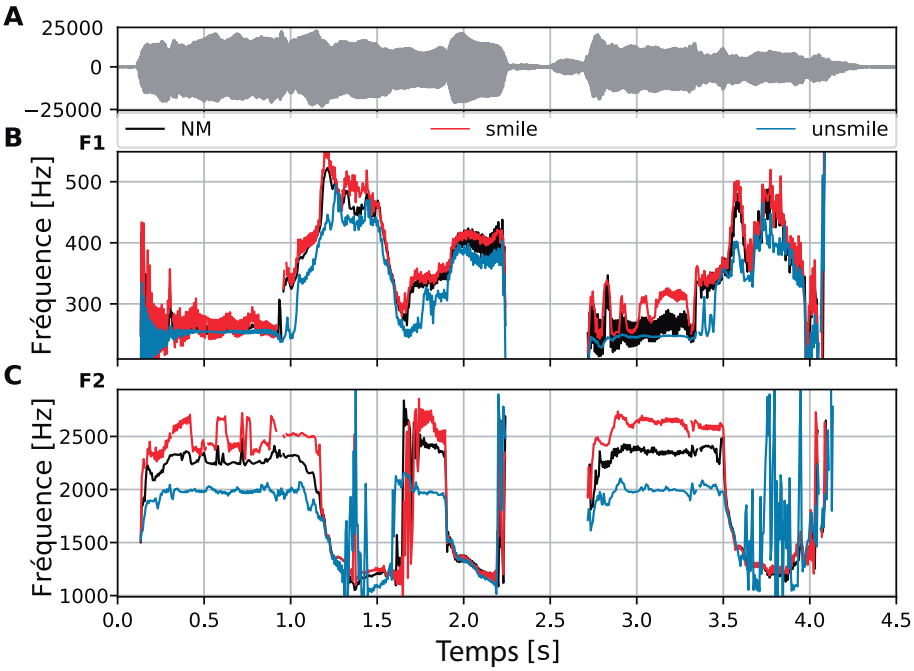


FIGURE 4.2 – Décalage de formants visualisé par l'analyse des formants F1 et F2 (réalisé avec Audiosculpt) du stimulus N° 1 de voix chantée.
A. Forme d'onde du signal original, **B.** F1 en fonction du temps, **C.** F2 en fonction du temps.
Légende : NM (son non-modifié), *smile* (décalage vers l'aigu), *unsmile* (décalage vers le grave).

L'expérience avait une durée d'environ 2 heures par sujet et était divisée en trois blocs (plus un entraînement) de la manière suivante :

- Bloc 0 : Entraînement (réponses non-enregistrées). Tous les types de son. 5 extraits.
- Bloc 1 : Types de son musicaux. 210 extraits.
- Bloc 2 : Voix parlée. 70 extraits.
- Bloc 3 : Cris. 60 extraits.

4.2.5 Ressources

La salle utilisée pour l'expérience était équipée avec plusieurs casques audio et ordinateurs fixes, nous avons installé sur ces derniers toutes les ressources informatiques nécessaires pour l'expérience. Ainsi, on a fait passer l'expérience à plusieurs sujets (10 maximum) à la fois. Toutes les ressources utilisées sont listées dans le Tableau A.1.

4.2.6 Récolte de données

Le script utilisé pour l'expérience produisait un fichier de texte (`.csv`) dans lequel on a enregistré différentes données de l'expérience qui sont détaillées dans le Tableau 4.1.

Note : À cause d'une erreur technique, on a eu seulement 294 sur 340 réponses (46 points du Bloc N°3 perdus) pour le participant numéro 14. On a donc éliminé ses réponses du Bloc N°3 et gardé le reste. Pour le cas de la rugosité, on a dû enlever les données du participant numéro 14 pour pouvoir exécuter la fonction `ezanova`, car cette fonction ne fonctionnait pas sans les données manquantes du bloc N°3.

4.2.7 Valeurs extrêmes

Les réponses correspondants aux participants qui ont fait des jugements d'affect de manière instable sont considérées comme valeurs extrêmes et ont été éliminées. Pour cela, on a fixé deux critères :

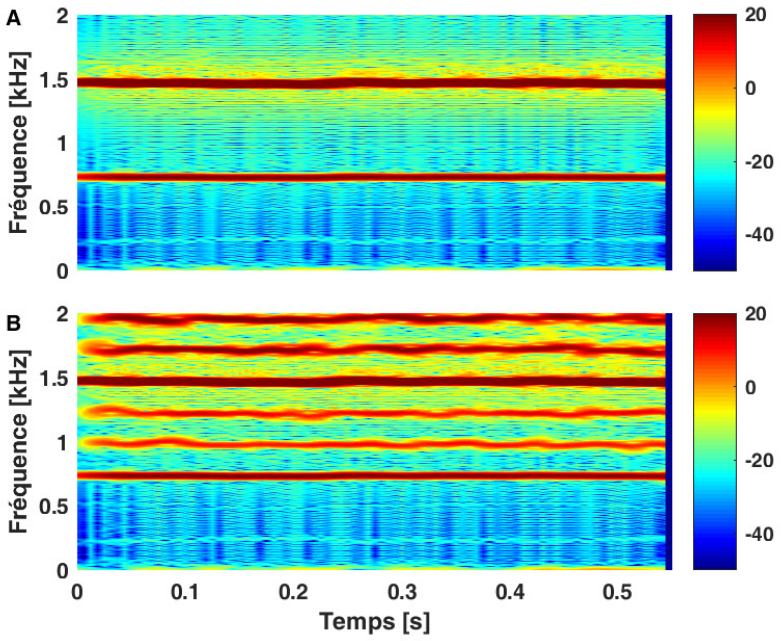


FIGURE 4.3 – Transformation de rugosité visualisée par le spectrogramme du cri N° 3.
A. Signal original **B.** Signal avec transformation *angus*.

1. Par temps de réponse (rt). Les temps de réponse de valence et/ou activation avec une variation de plus de 3 IQR (Écart interquartile) : rt $> 3 \times$ IQR. Par exemple, une personne qui a tardé 15 secondes pour donner une réponse sur son jugement affectif.
2. Par utilisation de l'échelle de valence et activation (rs). Les participants dont les notes de leurs réponses sur l'échelle de 1 à 7 ont un écart extrêmement réduit, définie comme : $\varepsilon = \big(max(\text{rs}) - min(\text{rs})\big) < 3$. Par exemple une personne qui répondait toujours 3 et 4.

4.3 Pré-enregistrement

L'expérience a fait l'objet d'un pré-enregistrement (*preregistration*) sur le site `http://aspredicted.org` où ont été consignées les hypothèses de départ avant d'avoir les données de l'expérience. En résumé, sur la voix parlée on attendait :

- Tâche de contrôle (NM comparée à NM) :
  - L'appréciation sur les deux échelles sera en moyenne autour du centre de l'échelle : NM vs NM : valence $= 4$; activation $= 4$.
- Hauteur - Les valeurs de valence des conditions '*up*' et '*down*' seront en moyenne :
  - Supérieures à 4 pour '*up*' vs '*down*' : Valence > 4.
  - Inférieures à 4 pour '*down*' vs '*up*' : Valence < 4.
- Induction zygomatique (Ziggy) - Les valeurs de valence des conditions '*smile*' et '*unsmile*' seront en moyenne :
  - Supérieures à 4 pour '*smile*' vs '*unsmile*' : Valence > 4.
  - Inférieures à 4 pour '*unsmile*' vs '*smile*' : Valence < 4.
- Vibrato :
  - Les valeurs de valence des conditions '*vibrato*' vs NM seront en moyenne inférieures à 4 : Valence < 4.
  - Les valeurs d'activation des conditions '*vibrato*' vs NM seront en moyenne supérieures à 4 : Activation > 4.
- Rugosité :
  - Les valeurs de valence des conditions '*angus*' vs NM seront en moyenne inférieures à 4 : Valence < 4.

Type de donnée	Nom du champ	Description
Identification	ID	Numéro de participant
Classification et repérage	Trial	Numéro d'essai
	Excerpt	Numéro du morceaux
	Stim 1	Premier extrait présenté
	Stim 2	Deuxième extrait présenté
	Order	Ordre dans la paire de l'extrait à évaluer
Manipulation du signal	Sound-type	Type de son présenté
	Manipulation	Type de manipulation appliquée au son évalué
	Transformation	Transformation évaluée sur le type de manipulation
Réponse cherchée	rs Valence	Note sur l'échelle de Valence
	rs Arousal	Note sur l'échelle d'Activation
	rt Valence	Temps de réponse sur l'échelle de Valence
	rt Arousal	Temps de réponse sur l'échelle d'Activation
Méta-données	Gender	Genre du son (voix parlée, chantée ou cri)
	Language	Langue dans laquelle les paroles sont écrites
	Duration	Durée en secondes du son évalué
Information démographique	Age	Âge du participant
	Sex	Sexe du participant
	Music experience	Musicien ou non-musicien
	Total time	Durée totale de l'expérience

TABLE 4.1 – Données récoltées

— Les valeurs d'activation des conditions '*angus*' vs NM seront en moyenne supérieures à 4 : Activation > 4.

On espérait retrouver une correspondance similaire pour les conditions d'écoute des sons musicaux. Pour plus de détails, le document est disponible sur le lien : `http://aspredicted.org/blind.php?x=2ur8pj`.

4.3.1 Aspect éthique

L'expérience a été réalisée au Centre de Recherche Multidisciplinaire Sorbonne Universités (INSEAD), les 3, 4 et 7 mai 2018 entre 10h00 et 17h30. Les participants ont donné leur consentement éclairé, et ont été rémunérés 10 €/h pour leur participation. Le protocole expérimental a été validé par le comité éthique de l'INSEAD.

4.3.2 Code

L'expérience a été entièrement codée avec le langage de programmation python 2.7, en utilisant principalement la librairie psychopy. Le code de l'expérience est accessible à travers l'adresse : `https://github.com/bedoya-daniel/affect_in_music`.

4.4 Analyses statistiques

Toutes les analyses statistiques ont été réalisées à l'aide du logiciel `R` et les figures ont été tracées avec la librairie python seaborn.

On utilise principalement la méthode d'analyse de variance à mesures répétées RM-ANOVA sur les échelles de valence et activation. On décrit dans l'annexe B la définition et caractéristiques du calcul des ANOVA. Chaque transformation est analysée de façon indépendante, selon 4 facteurs :

— Transformation : Hauteur, Formants, Vibrato, Rugosité (intra-sujet).

— Type de son : Voix parlée, Voix chantée, Voix chantée avec accompagnement musical, Instrument avec accompagnement musical, Cris (intra-sujet).
— Expérience musicale : Musicien, Non-musicien (inter-sujet).
— Ordre de présentation : Transformation cible évaluée en premier ou en deuxième place (inter-sujet).

La conception pour chaque analyse est visualisée dans le schéma de la figure 4.4.

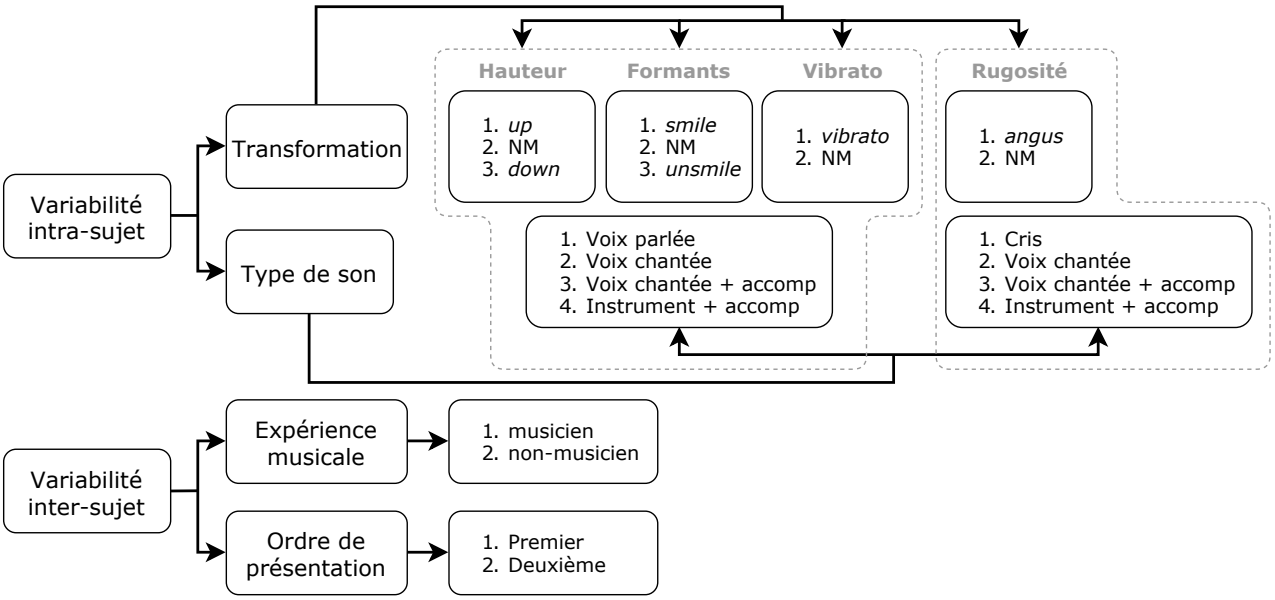


FIGURE 4.4 – Configuration des analyses ANOVA tant pour la note de valence que pour celle de l'activation. À chaque fois, seulement les éléments des niveaux de Transformation et de Type de son (entourés en trait discontinu) variaient pour selon le cas.

On reporte, dans le chapitre suivant, les résultats des ANOVA pour chaque effet, avec le format suivant :

$$\left[\mathrm{F}(DDL_B,\, DDL_W) = F, \quad \mathrm{p} = p, \quad \eta^2_G = \mathrm{ges}\right]$$

où F est le résultat du test de Fischer, DDL_B sont les degrés de liberté inter-sujet, DDL_W sont les degrés de liberté intra-sujet, p est la valeur statistique p, et ges est la taille de l'effet généralisé (*generalized eta-squared*, cf. B.14). Les corrections des valeurs p selon Huynh-Feldt epsilon ont été appliquées à partir du critère de sphéricité donné par le test de Mauchly. Les degrés de liberté sont reportés sans correction.

Chapitre 5

Résultats

Ce chapitre présente les résultats de l'expérience et des possibles interprétations de l'analyse des données.

5.1 Hauteur

La transformation de hauteur a un effet principal sur les notes de valence [F(2, 114)=38.74, p=5.14e-10, η^2_G=0.17], les stimuli manipulés vers le haut ('*up*') étant jugés plus positifs que ceux manipulés vers le bas ('*down*'). Cette relation est vraie pour tous les types de son (Fig. 5.1), ce qui valide d'une part l'hypothèse pré-enregistrée qu'une variation positive ou négative de hauteur a un effet de valence émotionnel clair sur la voix parlée, mais aussi que cet effet se transfère non seulement à la voix chantée, mais aussi à la musique instrumentale.

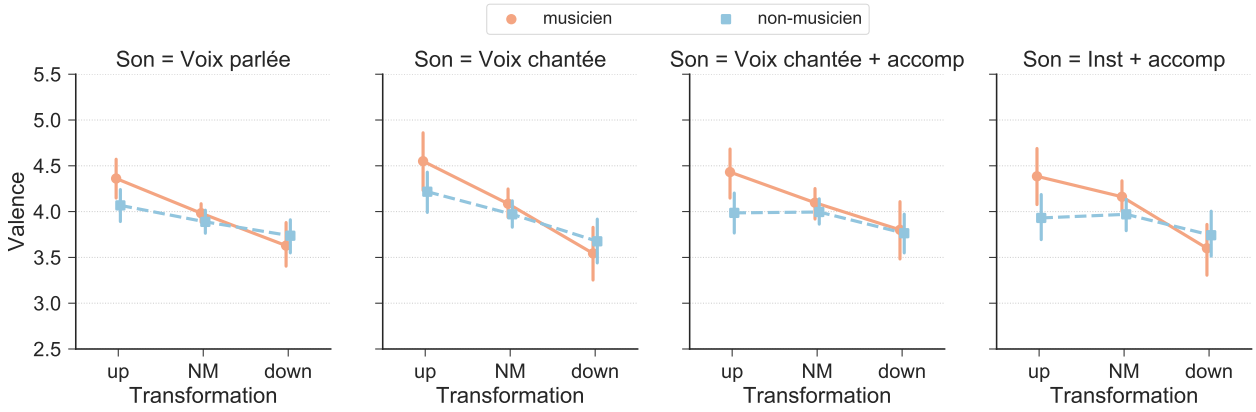


FIGURE 5.1 – Résultats de valence sur la hauteur pour tous les types de son, séparés par expérience musicale. Intervalles de confiance à 95%.

Deux facteurs interagissent avec cet effet de façon plus subtile, non-prédite par notre pré-enregistrement, néanmoins intéressante. D'une part, même si les transformations '*down*' sont perçues comme plus négatives que '*up*' pour tous les types de son, cet écart est plus grand pour la voix a cappella, ce qui se traduit par une interaction type de son/transformation significative [F(6, 342)=23.29, p=4.30e-02, η^2_G=0.01]. Une interprétation possible est que l'ambitus de la voix chantée est plus élevé que sur la voix parlée, ce qui pourrait favoriser la perception de changements de hauteur. En plus, la voix a cappella présente moins de masquage que les conditions avec accompagnement musical, ce qui peut également favoriser cette condition.

D'autre part, les musiciens perçoivent les sons comme plus positifs que les non-musiciens dans la condition '*up*', et ceci quelque soit le type de son, ce qui se traduit par une interaction expérience musicale/transformation significative [F(2, 114)=62.79, p=7.86e-03, η^2_G=0.03] (l'interaction expérience musicale avec transformation et type de son est non significative). Une interprétation possible est qu'il y a une meilleure sensibilité aux différences de hauteur chez les musiciens, par exemple dû à l'entraînement (Magne, Schön, & Besson, 2006).

Cependant, l'asymétrie '*up*'-'*down*' est difficile à expliquer, car on compare toujours le même intervalle au sein d'une paire (+30 cents et −25 cents) ; ce phénomène est peut-être à relier avec une asymétrie sur le jugement d'intervalles ascendants (par exemple juger le son '*up*' dans une paire '*down*'-'*up*') et descendants (par exemple juger le son '*down*' dans une paire '*up*'-'*down*') (Russo & Thompson, 2005). Par ailleurs, notre plage fréquentielle (cf. A.5) serait dans le registre grave ('*low-register*') utilisé par Russo et Thompson (2005) pour lequel on prédirait d'une part une sur-estimation des intervalles descendants plutôt qu'ascendants, et d'autre part l'absence d'interaction avec l'expérience musicale. Cette asymétrie est d'autant plus renforcée que les non-musiciens semblent évaluer moins positivement le '*up*' sur les 2 conditions avec accompagnement musical (cf. Figure 5.1), même si l'interaction entre type de son avec transformation et expérience musicale n'est pas significative.
Une interprétation possible est que l'effet de hauteur '*up*' est en concurrence avec un effet de dissonance qui est interprété négativement (McDermott, Lehr, & Oxenham, 2010) et que seuls les musiciens sont capables de juger cette dissonance comme un effet d'expressivité (ou de s'abstraire de ce caractère émotionnel de la dissonance) ; ceci est à rapprocher des différences culturelles rapportées par McDermott, Schultz, Undurraga, et Godoy (2016).

De façon non-prédite par notre pré-enregistrement, la transformation de hauteur a également un petit effet pour les notes d'activation [F(2, 114)=37.83, p=2.22e-11, η^2_G=0.11], voir Fig. 5.2. Comme pour la valence, cet effet semble conservé pour tous les types de son, même si une interaction marginale entre type de son et transformation [F(6, 342)=2.09, p=.053, η^2_G=0.01] suggère un effet activant de la hauteur montante '*up*' moins grand pour les conditions avec accompagnement musical que pour les conditions de voix parlée et chantée a cappella.

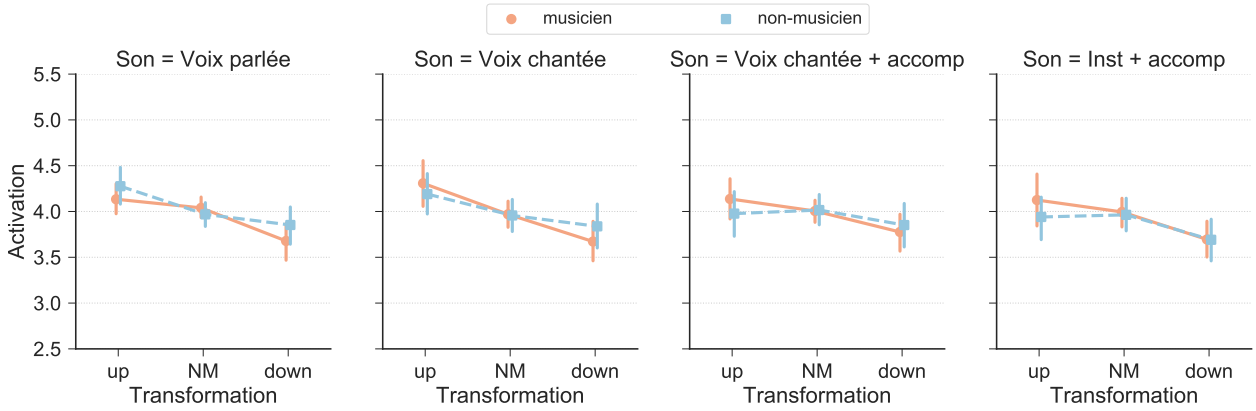
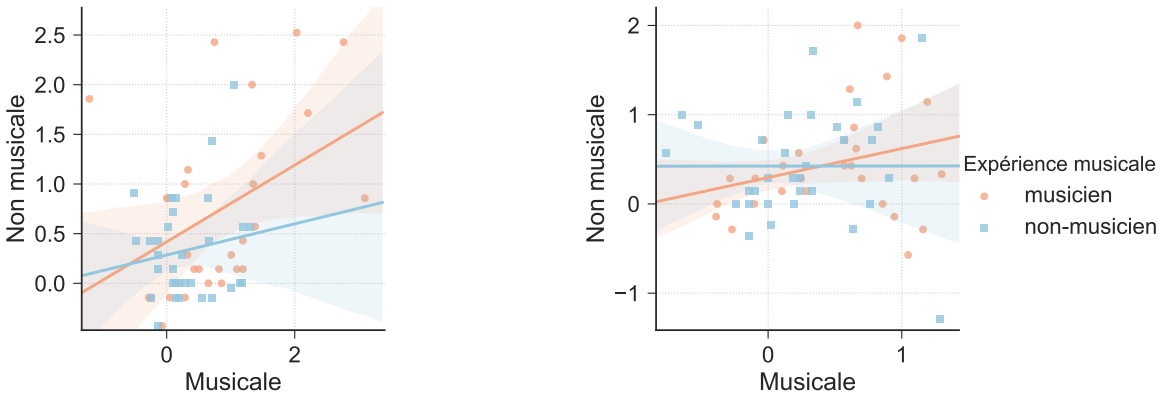


FIGURE 5.2 – Résultats d'activation sur la hauteur pour tous les types de son, séparés par expérience musicale. Intervalles de confiance à 95%.

Un moyen supplémentaire d'examiner si le transfert de l'effet émotionnel de la manipulation de hauteur entre type de sons est valide quelques soient les participants, consiste à visualiser la corrélation de l'effet de la hauteur entre la condition non-musicale (voix parlée) et les conditions musicales (voix chantée, voix chantée avec accompagnement musical et instrument avec accompagnement musical). Pour la valence (fig. 5.3a), cette corrélation est significative r= 0.40, p= 0.001, indiquant que les participants fortement influencés par la hauteur pour la voix parlée le sont également pour la musique. Cependant, cette corrélation est principalement portée par les participants musiciens (r= 0.41, p= 0.025), mais n'est pas significative pour les participants non-musiciens (r= 0.16, p= 0.397). Pour l'activation (fig. 5.3b), la corrélation des notes d'activation entre la condition non-musicale (voix parlée) et les conditions musicales (voix chantée, voix chantée avec accompagnement musical et instrument avec accompagnement musical) n'est significative ni pour les sujets musiciens, r= 0.27, p= 0.158, ni pour les sujets non-musiciens, r= 0.001, p= 0.995.

5.2 Formants

La transformation de formants '*smile*' a un effet principal pour les notes de valence [F(2, 114)=83.96, p=7.61e-16, η^2_G=0.36] (fig. 5.4), les sons '*smile*' étant jugés plus positifs que les sons '*unsmile*' pour tous les types de sons. Comme pour la hauteur, ce résultat valide notre hypothèse pré-enregistrée pour la voix parlée,



(a) Valence (b) Activation

FIGURE 5.3 – Corrélation des notes sur la hauteur entre toutes les conditions musicales (abscisse) et la voix parlée (ordonnée), séparée par expérience musicale.

et confirme que la même manipulation provoque le même type de jugement sur des stimuli musicaux, et en particulier la musique instrumentale.

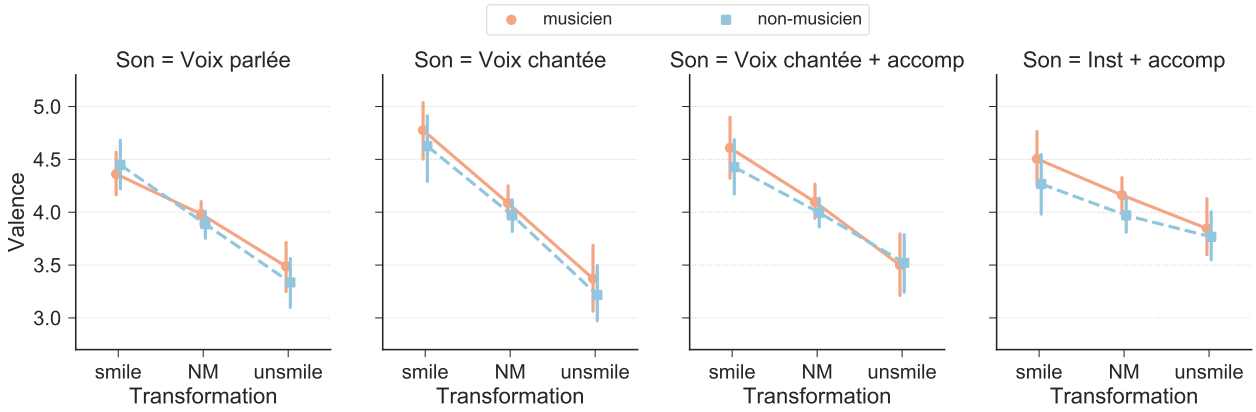


FIGURE 5.4 – Résultats de valence sur les formants pour tous les types de son, séparés par expérience musicale. Intervalles de confiance à 95%.

De façon non-prédite par notre pré-enregistrement, les résultats montrent une interaction significative entre type de son et transformation [F(6, 342)=11.89, p=5.32e-10, η^2_G=0.05]. Même si les deux directions de transformation (‘*smile*’ et ‘*unsmile*’) sont jugées de façon symétrique quelque soit le type de son, la transformation a un effet plus important sur la voix chantée a cappella, et réduit pour le type de son instrument avec accompagnement. Une interprétation possible est qu’il y a une plus grande amplitude du décalage de timbre sur la voix chantée, comme on l’avait pour la hauteur (relation apparente avec la ‘*super-expressive voice theory*’), un masquage spectral (qui réduit l’effet par rapport à la voix a cappella, mais pas au point de le réduire par rapport à la voix parlée) et un amoindrissement de l’effet ‘sourire’ sur un timbre de source non-vocale (mais qui reste significativement supérieur à 0). Cet affaiblissement est soit dû à la technique (mauvaise détection de ‘formants’, intensité de la transformation plus faible) soit à la cognition (incongruité entre le percept de sourire et l’identification de la source comme non-vocale), soit aux deux.

De façon non prédite par notre pré-enregistrement, la transformation formantique a également un effet sur les notes d’activation [F(2, 114)=102.74, p=2.24e-17, η^2_G=0.34], voir fig. 5.5. Même si cet effet est non-prédit, l’effet est présent dans la même direction sur la voix parlée que sur les conditions musicales, ce qui est cohérent avec le transfert vu pour la valence. Pour l’activation, l’interaction entre type de son et transformation est significative [F(6, 342)=19.76, p=2.75e-17, η^2_G=0.10]. Même si l’effet de la transformation sur l’activation est dans la même direction pour tous les types de sons, l’effet activant est plus grand pour les conditions voix parlée et voix chantée a cappella que pour les deux conditions avec accompagnement musical. Une interprétation

possible serait un masquage spectral plus grand quand il y a un accompagnement musical - l'effet semble indépendant du fait que la piste est vocale ou instrumentale.

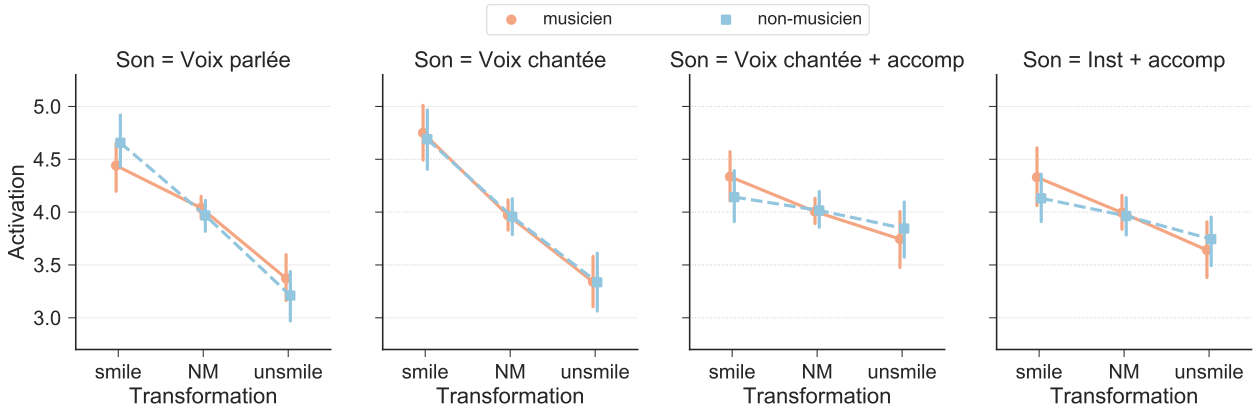
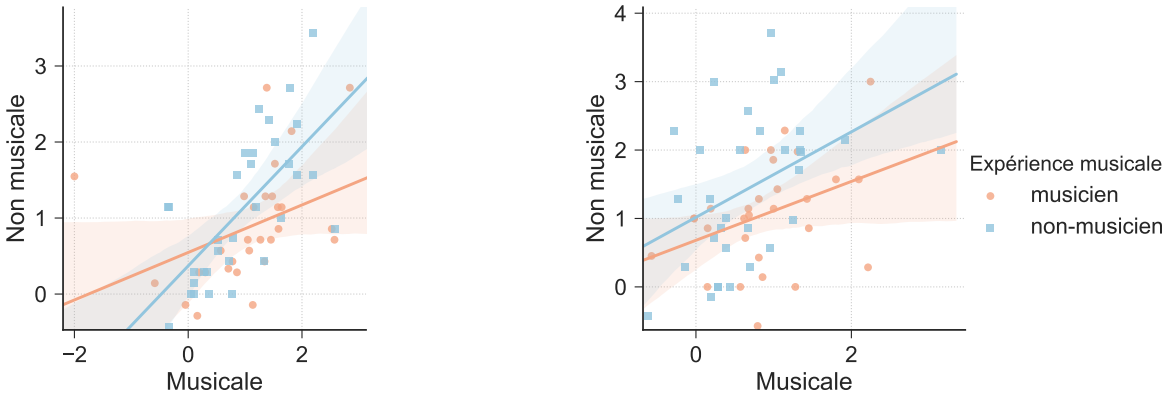


FIGURE 5.5 – Résultats d'activation sur les formants pour tous les types de son, séparés par expérience musicale. Intervalles de confiance à 95%.

Comme précédemment, nous examinons la corrélation de l'effet de manipulation formantique sur les notes de valence entre la condition non-musicale (voix parlée) et les conditions musicales (voix chantée, voix chantée avec accompagnement musical et instrument avec accompagnement musical). Cette corrélation est significative pour les sujets musiciens, r= 0.39, p= 0.034, comme pour les sujets non-musiciens, r= 0.67, p= 2.84e-05 (fig. 5.6a). Cela indique que les participants fortement influencés par le sourire vocal quand ils jugent la voix parlée le sont également quand la même manipulation s'applique à des sons musicaux. De la même façon, la corrélation de l'effet du sourire sur les notes d'activation entre la condition voix parlée et les conditions musicales est marginalement significative pour les sujets musiciens, r= 0.36, p= 0.051, et significative pour les sujets non-musiciens, r= 0.42, p= 0.018 (fig. 5.6b).



(a) Valence

(b) Activation

FIGURE 5.6 – Corrélation des notes sur la transformation des formants entre toutes les conditions musicales (abscisse) et la voix parlée (ordonnée), séparée par expérience musicale.

5.3 Vibrato

La transformation de vibrato a un effet principal pour les notes de valence [F(1, 57)=48.48, p=3.67e-09, η^2_G=0.11], les sons tremblants étant perçus comme plus négatifs que les sons non-modifiés (fig. 5.7). Cet effet est dans la même direction pour tous les types de son. L'interaction entre transformation et type de son est non-significative [F(3, 171)=1.95, p=.12], même s'il y a une tendance à une amplification de l'effet pour la voix chantée a cappella (a cappella : $M \simeq 0.55$ vs voix parlée : $M \simeq 0.2$). Ce résultat confirme notre hypothèse

pré-enregistrée pour la voix parlée, et, comme pour la hauteur et le sourire, confirme que les indices expressifs du vibrato sont traités de la même manière quand ils s'appliquent à des sons musicaux.

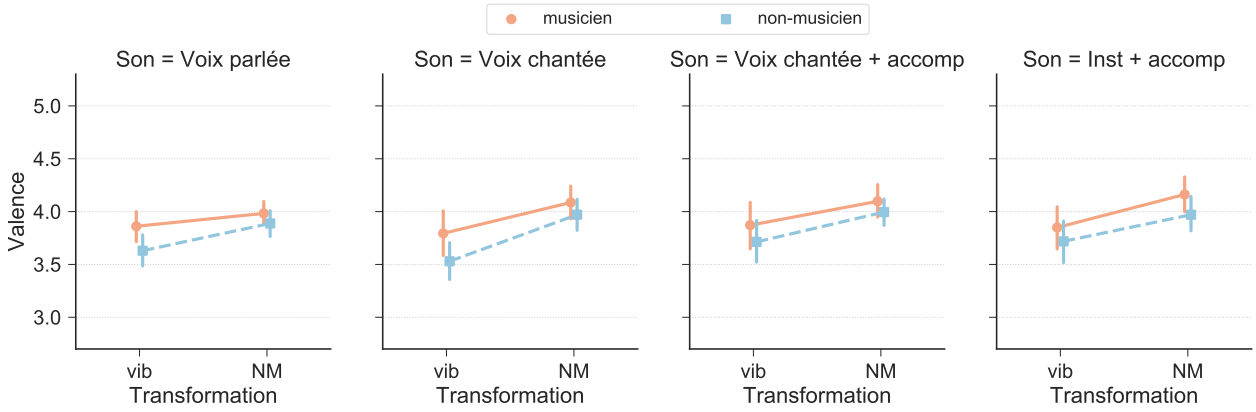


FIGURE 5.7 – Résultats de valence sur le vibrato pour tous les types de son, séparés par expérience musicale. Intervalles de confiance à 95%.

On peut noter que, contrairement à la manipulation de hauteur pour laquelle l'effet était plus important pour les musiciens, il n'y a ici pas d'interaction de l'effet du vibrato avec l'expérience musicale. Ceci est surprenant car la manipulation est elle aussi basée sur la hauteur et l'amplitude du vibrato (50 cents pic à pic) n'est pas supérieure à celle de la hauteur (55 cents). Il est possible que le traitement psychoacoustique du vibrato est différent de celui du décalage fréquentiel (par exemple, un traitement de la modulation).

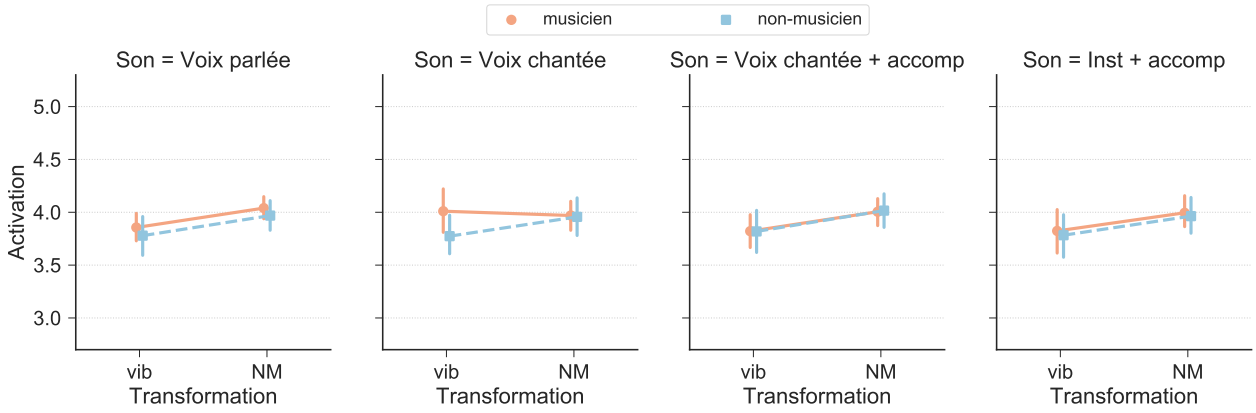
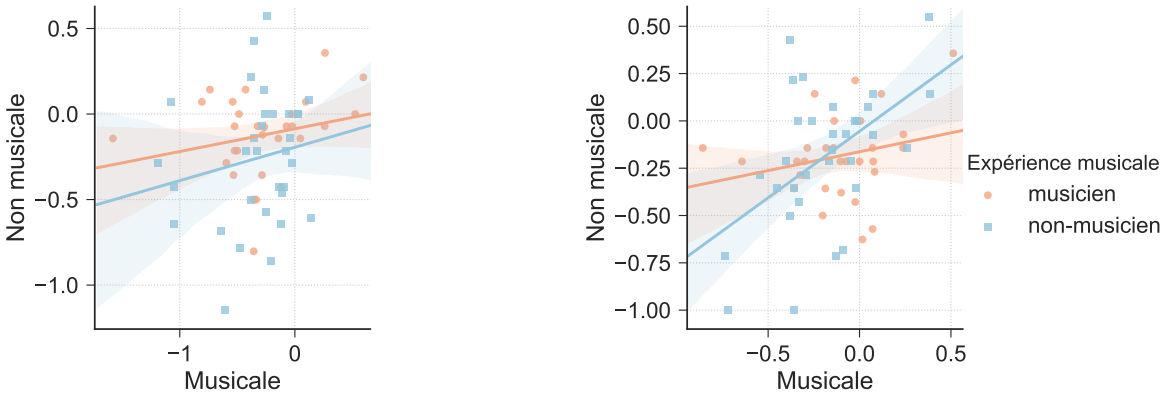


FIGURE 5.8 – Résultats d'activation sur le vibrato pour tous les types de son, séparés par expérience musicale. Intervalles de confiance à 95%.

La transformation de vibrato a également un effet principal pour les notes de activation [F(1, 57)=24.28, p=7.51e-06, η^2_G=0.04], mais cet effet est dans la direction inverse de celle prédite par notre pré-enregistrement : pour la voix parlée comme pour les conditions musicales, le vibrato baisse l'activation (fig. 5.8). Ce résultat est surprenant : dans l'étude de (Rachman et al., 2018), une manipulation '*afraid*' constitué de vibrato et d'une inflexion de hauteur sur le début des phrases avait généré une baisse de valence et une hausse d'activation, ce qui était interprété comme une peur ou une tristesse 'activée' (type anxiété, panique ou pleurs). Cette différence ne semble pas attribuable aux stimuli (pour la condition voix parlée, l'étude de Rachman et al. utilise comme nous des phrases complexes de quelques secondes), à la langue (Rachman et al. testent le français et l'anglais, comme ici) ou à la population de participants, similaire. Il est possible que l'effet de vibrato seul (sans l'inflexion utilisée simultanément par Rachman et al.) ait un effet différent, suggérant un état négatif peu activé (type déprimé, découragé). Même si cet effet était non-prédit, il reste le même sur la voix parlée et les conditions musicales, sans interaction significative avec le type de son, ce qui vient renforcer la notion de transfert de la voix vers la musique.

Contrairement au sourire et, dans le cas des musiciens, à la hauteur, la corrélation entre l'effet du vibrato sur les notes de valence entre la voix parlée et les conditions musicales n'est significative ni pour les sujets

musiciens, r= 0.23, p= 0.222, ni pour les sujets non-musiciens, r= 0.18, p= 0.326 (fig. 5.9a). La corrélation des notes d'activation entre la voix parlée et les conditions musicales n'est pas significative pour les sujets musiciens, r= 0.24, p= 0.211, mais elle est significative pour les sujets non-musiciens, r= 0.52, p= 0.003 (fig. 5.9b).



(a) Valence

(b) Activation

FIGURE 5.9 – Corrélation des notes sur la transformation vibrato entre toutes les conditions musicales (abscisse) et la voix parlée (ordonnée), séparée par expérience musicale.

5.4 Rugosité

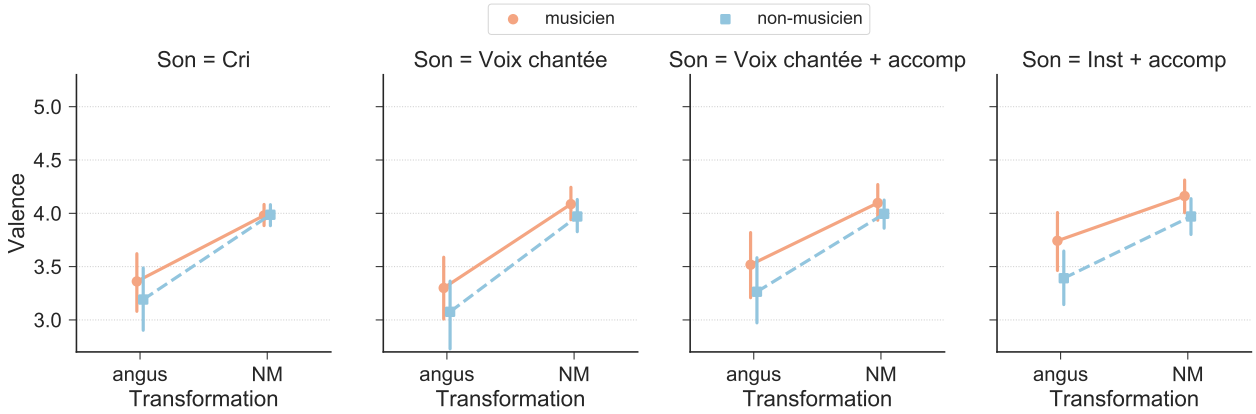


FIGURE 5.10 – Résultats de valence sur la rugosité pour tous les types de son, séparés par expérience musicale. Intervalles de confiance à 95%.

La transformation de rugosité a un effet principal pour les notes de valence [F(1, 56)=63.87, p=7.97e-11, η^2_G=0.26], les sons rugueux étant perçus comme plus négatifs que les sons non-modifiés (fig. 5.10). Comme précédemment, la transformation agit dans la même direction pour les quatre types de sons. Ces résultats valident à la fois notre hypothèse pré-enregistrée pour la voix parlée, mais également l'idée de transfert des indices de rugosité de la voix vers les sons musicaux.

L'interaction entre type de son et transformation est significative [F(3, 168)=4.73, p=6.48e-03, η^2_G=0.01]. Même si la transformation agit dans la même direction pour les 4 types de sons, elle a un effet plus important sur la voix chantée a cappella, et plus réduit pour la condition de piste instrumentale avec accompagnement. La voix a cappella est au même niveau que les cris, ce qui s'explique sans doute par la grande similarité acoustique entre les deux types de sons. On peut également noter un effet principal du type de son sur la valence [F(3, 168)=5.94, p=1.97e-03, η^2_G=0.02], la valence jugée dans les paires de cris ou de voix chantée a cappella étant toujours inférieure à celle jugée dans les paires avec accompagnement musical. Cet effet n'est présent que pour les paires manipulées (et pas dans les paires contrôles NM-NM), donc cela n'est pas un biais

de réponse, mais plutôt un effet de masquage qui amoindrit la perception négative des signes de rugosité dans les paires avec accompagnement musical.

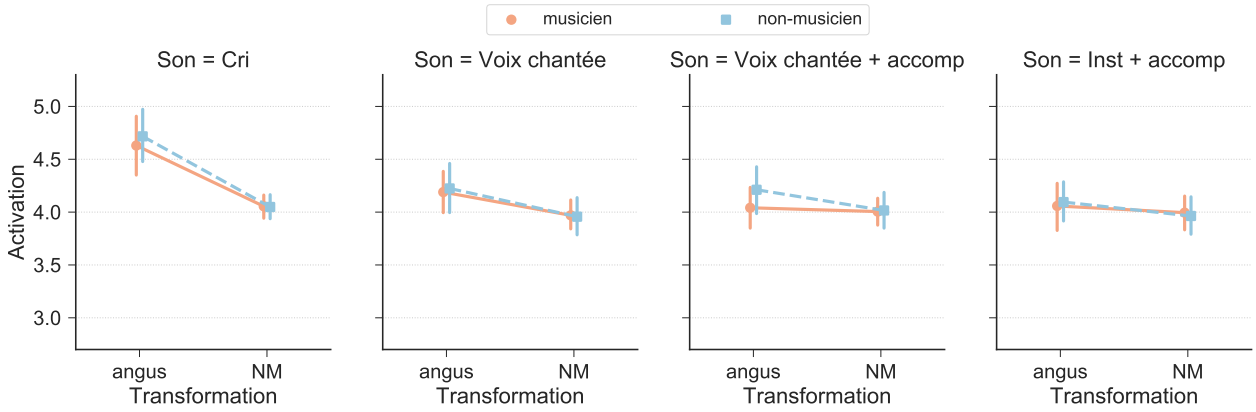
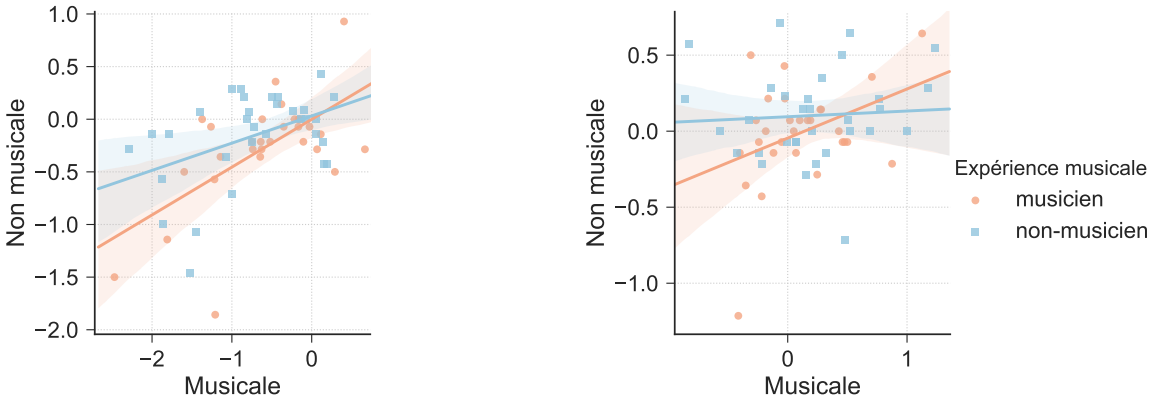


FIGURE 5.11 – Résultats d'activation sur la rugosité pour tous les types de son, séparés par expérience musicale. Intervalles de confiance à 95%.

Comme prédit dans notre pré-enregistrement, la transformation de rugosité a également un effet principal sur les notes d'activation [F(1, 56)=23.52, p=1.02e-05, η^2_G=0.08] (fig. 5.11), qui est vrai pour les 4 types de sons. Cependant, l'effet de la transformation est plus activant pour les cris que pour les autres conditions, en particulier la voix a cappella (interaction type de son / transformation [F(3, 168)=12.60, p=7.16e-07, η^2_G=0.05]). Cet effet semble peu compatible avec l'hypothèse de '*super-expressive voice theory*'. Peut-être s'agit-il d'un effet stylistique de la rugosité comme le '*growl*' (son qui imite un grognement), qui réduit l'impact émotionnel de la rugosité quand le contexte est musical (Liuni, Ponsot, & Aucouturier, s. d.) — il est à noter que cette interaction ne dépend pas de l'expérience musicale.

La corrélation de l'effet de la rugosité sur les notes de valence entre la condition non-musicale (cris) et les conditions musicales (voix chantée, voix chantée avec accompagnement musical et instrument avec accompagnement musical) est significative pour les sujets musiciens, r= 0.62, p= 3.79e-04, et pour les sujets non-musiciens, r= 0.45, p= 0.011 (fig. 5.12a). La corrélation des notes d'activation entre la condition non-musicale (cris) et les conditions musicales est significative pour les sujets musiciens, r= 0.37, p= 0.047, mais ne l'est pas pour les sujets non-musiciens, r= 0.06, p= 0.726 (fig. 5.12b).



(a) Valence

(b) Activation

FIGURE 5.12 – Corrélation des notes sur la transformation de rugosité entre toutes les conditions musicales (abscisse) et les cris (ordonnée), séparée par expérience musicale.

5.5 Remarque méthodologique

Notre protocole expérimental, inspiré de Ma et Thompson (2015), présente des paires de son en variant entre sujets si le jugement porte sur le premier ou le deuxième son d'une paire. Les résultats sont ensuite agglomérés indépendemment de cet ordre de présentation (ex. l'effet de la transformation 'up' est calculé à partir des paires 'up-down' pour ceux des participants devant juger des premiers sons de chaque paire, et des paires 'down-up' pour les participants devant juger des deuxièmes sons de chaque paire). Cette précaution expérimentale, introduite par Ma et Thompson (2015), était-elle utile ?

Cela semble être le cas. Dans les 4 transformations, on observe des effets d'interaction entre le rang jugé par les participants et le type de son (par ex. pour la valence, hauteur : [F(3, 171)=8.28, p=3.57e-05, η^2_G=0.04] ; formants : [F(3, 171)=3.03, p=3.09e-02, η^2_G=0.01] ; vibrato : [F(3, 171)=6.29, p=4.53e-04, η^2_G=3.14e-2] ; rugosité : [F(3, 168)=4.94, p=5.61e-03, η^2_G=0.02)]. Ces interactions n'ont aucun intérêt théorique, car elles sont indépendantes du fait que les stimuli aient été manipulés ou non. Elles correspondent en fait à un biais de réponse positif pour les participants évaluant les sons en deuxième position, spécifique aux sons de voix chantée avec accompagnement (la musique chantée la plus récente parait toujours plus émotionnellement positive, indépendamment de la manipulation appliquée). Cependant, elles montrent qu'il était important de contrebalancer dans quelle position est jugée le son, et en particulier de ne pas toujours juger le deuxième son d'une paire, comme on a souvent tendance à le faire (Rachman et al., 2018), car cela peut amplifier l'effet d'une catégorie de son, voire d'une transformation. Le fait que la musique vocale avec accompagnement évaluée en deuxième position soit toujours plus positive et plus énergique est consistant avec l'effet de la répétition sur les préférences musicales ('*mere exposure effect*' - Hargreaves, 1984). Il est toutefois intéressant que cet effet n'est pas observé pour la voix parlée, ni pour la musique instrumentale.

Chapitre 6

Discussion et perspectives

6.1 Hypothèse générale

Les résultats présentés dans le chapitre précédent confirment non seulement les hypothèses pré-enregistrées sur l'influence des 4 manipulations sur la voix parlée (la hauteur et *smile* augmente la valence, vibrato et rugosité la diminue), mais surtout le fait que ces 4 transformations ont un effet similaire sur les 3 conditions musicales. Ce résultat établit la preuve expérimentale que les indices acoustiques de la parole expressive sont impliqués de façon causale dans la perception des émotions dans la musique vocale et instrumentale, une notion qui avait été jusqu'à présent beaucoup théorisée (Juslin & Västfjäll, 2008), mais difficile à confirmer de façon expérimentale (Juslin & Laukka, 2003 ; Ilie & Thompson, 2006).

Ce résultat est d'autant plus intéressant qu'il examine, d'une part, une assez grande diversité d'indices acoustiques, liés à la hauteur de façon statique - *up* et *down* -, de façon dynamique - *vibrato* - et aux aspects spectraux - *smile* - et dynamiques - rugosité - du timbre ; d'autre part, qu'il établit un continuum de transfert allant de la voix chantée a cappella aux musiques non-vocales. À son extrême, le résultat est peu trivial : les auditeurs semblent bel et bien juger de l'émotion exprimée par une partie de violon selon que l'instrument leur semble légèrement désaccordé vers le haut comme une voix joyeuse, tremblant comme une voix triste, rugueux comme un cri de colère ou même brillant comme une bouche dont on étire les muscles zygomatiques. D'un point de vue théorique, ce résultat renforce un certain nombre de données comportementales et d'imagerie cérébrale montrant que la musique et la langage partagent un grand nombre de ressources cognitives (Patel, 2010 ; Escoffier et al., 2013). Il peut notamment être discuté dans la perspective des origines biologiques de ces deux aptitudes spécifiquement humaines. Le fait que la voix parlée et la musique partage ainsi une partie du même 'code émotionnel' peut indiquer que langage et musiques découlent tous deux d'un même 'ancêtre commun', ce que Darwin a décrit comme un 'protolangage musical' largement consacré à la communication émotionnelle [1]. En particulier, le fait que deux des indices manipulés ici (*smile* et rugosité) soient spécifiquement basés sur la physiologie de l'appareil phonatoire mammalien (articulation vocale pour l'un, saturation laryngée pour l'autre) et qu'ils se transfèrent malgré tout à la musique instrumentale semble devoir suggérer une origine vocale d'au moins une partie de ce qui rend la musique expressive.

6.2 L'hypothèse de la *super-expressive voice*

La théorie de la *super-expressive voice* (Juslin, 2001 ; cf. 2.4) prédit non seulement que les traits acoustiques communs entre la musique et la voix seraient la source de la réponse émotionnelle générée par la musique mais aussi que ces attributs musicaux sont des versions extrêmes de ceux trouvés dans la voix parlée ; ils produisent ainsi des effets émotionnels plus intenses que ceux possibles avec la voix. Examinons cette idée pour les quatre transformations appliquées ici (fig. 6.1) :

1. "Nous verrons, lorsque nous nous occuperons de la sélection sexuelle, que les hommes primitifs, ou plutôt quelque antique ancêtre de l'homme s'est probablement beaucoup servi de sa voix, comme le font encore aujourd'hui certains gibbons, pour émettre de véritables cadences musicales, c'est-à-dire pour chanter. Nous pouvons conclure d'analogies très généralement répandues que cette faculté s'exerçait principalement aux époques où les sexes se recherchent, pour exprimer les diverses émotions de l'amour, de la jalousie, du triomphe, ou pour défier leurs rivaux. Il est donc probable que l'imitation des cris musicaux par des sons articulés ait pu engendrer des mots exprimant diverses émotions complexes." - (Darwin, 1891)

Pour la transformation de hauteur ('up-down'), l'effet est en partie cohérent avec l'hypothèse : à décalage de hauteur identique, l'influence sur les jugements émotionnels est plus grande sur la voix chantée a cappella que sur la voix parlée. Cependant, cet effet semble compensé sur la musique vocale avec accompagnement, et la musique instrumentale, qui ont des amplitudes émotionnelles comparables à celle de la voix parlée. Ceci peut suggérer un phénomène de 'masquage' des différences de hauteur dû à l'ajout d'un accompagnement harmoniquement riche, ou la concurrence pour la transformation '*up*' d'un effet de dissonance, à valence négative.

Pour la transformation de formants, l'effet est à nouveau en partie cohérent avec l'hypothèse : à décalage de formants identique, l'influence sur les jugements émotionnels est plus grande sur la voix chantée a cappella que sur la voix parlée. L'effet est réduit par rapport à la voix chantée pour les musiques vocales avec accompagnement, ce qui suggère à nouveau un phénomène de masquage spectral, et réduit encore davantage pour les pistes instrumentales, ce qui est peut-être une conséquence d'une moindre adaptation de l'effet de sourire à des sources non-vocales.

Pour la transformation vibrato, l'effet est à nouveau cohérent avec l'hypothèse, avec un effet plus important pour la voix a cappella que la voix parlée, et réduit sur les conditions musicales avec accompagnement et instrumentales.

Enfin, pour la rugosité, l'effet est également plus activant sur les conditions musicales que la voix parlée, mais non si on considère comme contrôle les cris plutôt que la voix parlée.

En résumé, le support expérimental pour la *super-expressive voice theory* semble mixte. D'un côté, à intensité de transformation égale, il apparaît que l'influence sur les jugements émotionnels est plus importante pour la voix chantée a cappella que pour la voix parlée. Cela traduit possiblement une tendance de l'auditeur à traiter la voix chantée comme un super-signal émotionnel, peut-être à cause d'une plus grande résolution perceptive (qui fait paraître un même changement acoustique comme plus saillant sur le chant) ou un biais de jugement (qui fait paraître un même changement acoustique comme reflétant une émotion plus forte sur le chant). D'un autre côté, cet effet semble rapidement compensé dès que la texture musicale devient complexe, ou que la voix est remplacée par un instrument, et ne semble pas si flagrant si on compare la musique non pas à la voix parlée, mais aux cris [2].

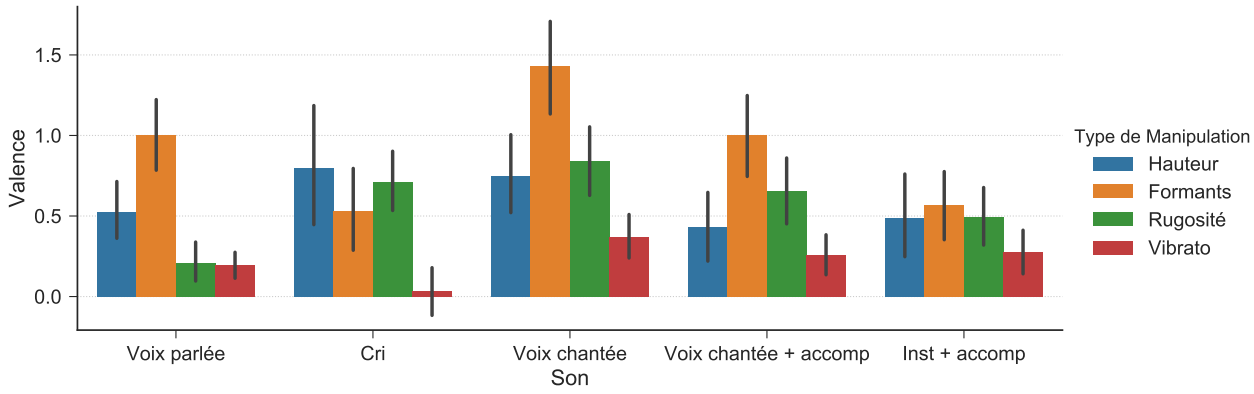


FIGURE 6.1 – Différence des notes moyennes de valence par sujet, entre transformations opposées. $\Delta_{\text{transformation}}$ = (*up*−*down*, *smile*−*unsmile*, NM−*vibrato*, NM−*angus*), pour tous les types de son. Intervalles de confiance à 95%.

6.3 Au delà de la théorie, des subtilités

Même si nos résultats semblent apporter la confirmation générale d'un transfert des indices expressifs vocaux vers la musique, l'analyse précise fournit quelques détails intéressants à discuter :

Premièrement, le fait de décaler la hauteur des signaux musicaux entraîne, par définition, une sorte de dissonance entre les pistes modifiées et l'accompagnement musical. Pour la transformation de hauteur '*up*', cet

2. les résultats concernant les transformations de hauteur, formants et vibrato sur les cris, ainsi que de rugosité sur la voix parlée, visualisés dans la figure 6.1 sont donnés ici pour référence, mais ne sont pas discutés dans le chapitre précédent.

effet de dissonance, à valence négative, peut rentrer en compétition avec l'effet à valence positive prédit par la voix parlée. Cet effet explique peut-être l'asymétrie émotionnelle observée, chez les non-musiciens, entre le gain dû à la voix 'up' et la voix 'down'. De manière générale, le transfert vers la musique d'indices acoustiques d'origine vocale peut créer des effets 'purement musicaux' (par ex. de l'ordre de la consonance, de l'harmonie, du rythme), non émotionnellement neutres et possiblement soumis à l'expertise musicale de l'auditeur, et qui compliquent la simple généralisation des effets vocaux à la musique.

Deuxièmement, nous avons trouvé une différence d'activation dans les conditions musicales pour la transformation de rugosité : sur des cris non-musicaux, la rugosité a un fort effet activant, mais cet effet semble disparaître sur la voix chantée, pourtant très proche du point de vue acoustique, et les autres conditions musicales. Il est possible que, dans un contexte musical, la rugosité ne soit pas systématiquement perçue comme une expression émotionnelle mais plutôt stylistique, à l'instar du 'growl' utilisé dans le chant pop/rock, ou même du timbre enroué de certains chanteurs (par ex. Louis Armstrong, cité dans Sakakibara, Fuks, Imagawa, & Tayama, 2004). Même s'il n'est pas aussi clair (voir les sujets musiciens, en condition voix chantée, dans la figure 5.8), un effet similaire aurait pu être attendu pour la transformation de vibrato, qui indique une émotion négative pour la voix parlée, mais a une signification stylistique forte et non-nécessairement émotionnelle dans un contexte musical, par exemple dans le chant lyrique ou la musique baroque (Donington, 1963). En somme, les émotions signalées par la musique ne sont pas seulement le produit de l'évolution biologique (similaire à celle de la voix parlée), mais aussi d'une évolution culturelle qui la prolonge et peut la contredire.

Finalement, grâce à la séparation des participants en groupes de musiciens et non-musiciens (cf. section 4.2), on a pu constater plusieurs différences de traitement des transformations selon l'expertise de l'auditeur. Pour la transformation de hauteur, les sujets musiciens sur-évaluent l'impact de la transformation de hauteur '*up*' par rapport aux non-musiciens ; ceci peut être dû à une meilleure perception, ou un traitement différent de la dissonance dûe à l'effet. Pour la transformation de vibrato, les musiciens ne perçoivent pas de différence d'activation dû au vibrato sur la voix chantée a cappella. Pour la rugosité, les musiciens ne perçoivent pas de différence d'activation sur la voix chantée avec accompagnement. Même si ces effets n'ont probablement pas d'explication unique, ils renforcent l'impression que la théorie originale du simple transfert de la voix vers la musique est probablement sous-spécifiée : c'est toute la richesse, acoustique et culturelle, du stimulus musical qu'il s'agit de considérer pour expliquer ces effets.

Plus généralement, ces quelques cas particuliers montrent l'intérêt de traduire des théories en modèles computationnels, pour en tester les paramètres de façon explicite (si j'augmente X, qu'advient-il d'Y ?) : par exemple, s'il était séduisant de postuler qu'un décalage de hauteur suffit à créer une musique plus joyeuse (Juslin & Västfjäll, 2008), ce n'est qu'en essayant de construire de tels exemples avec nos méthodes de transformation de signaux que l'on se rend compte que ces décalages peuvent interagir avec l'accompagnement ou le style musical pour produire au contraire des dissonances à valence négative. Ces résultats, parfois contre-intuitifs, peuvent ensuite nous conduire à raffiner la théorie originelle, d'une façon qui aurait été difficilement possible sans avoir construit ces stimuli.

6.4 Amélioration du protocole

En guise de perspective, on propose ici quelques améliorations au protocole expérimental présenté dans la section 4.2, qui pourraient bénéficier la recherche future sur le sujet.

- Création des stimuli spécifiques pour contrôler tous les aspects musicaux (genre, durée, tonalité) et acoustiques dans l'enregistrement (choix des instruments, réponse fréquentielle des microphones). En plus, de cette manière la même personne (avec un même profile acoustique de voix) chanterait, enregistrerait la voix parlée et les cris.
- Implémenter une classification plus rigoureuse par rapport à l'expérience musicale, dès le recrutement, avec des sous-critères qui filtrent des particularités comme l'amusie et/ou l'oreille absolue. Utiliser un système de vérification des capacités musicales d'un sujet, comme le PROMS (Law & Zentner, 2012) ou le Goldsmiths MSI (Müllensiefen, Gingras, Musil, & Stewart, 2014). Ceci peut éliminer le problème des critères auto-déclarés.
- Réduire la durée de l'expérience ou diviser l'expérience en plusieurs sessions pourrait aider à réduire la fatigue cognitive des participants et les possibles problèmes associés, notamment une aversion à la tâche ou une augmentation d'un humeur négatif (Lorist et al., 2000). L'inconvénient avec les sessions divisées, est que si un participant manque une session, ses données seront inutilisables.

— Introduire des méthodes multi-approche de mesures cognitives continues ou en temps réel (Schubert, 2007 ; Nagel, Kopiez, Grewe, & Altenmüller, 2007). Ceci favoriserait une prise en compte du temps de réponse, permettrait de caractériser les réponses dans l'espace (cf. figure 2.1) et mesurer la valence et activation en même temps.
— Inclure d'autres systèmes musicaux qui n'appartiennent pas à la musique populaire occidentale (gammes hexatoniques ou pentatoniques) ou travailler le décalage fréquentiel avec des intervalles musicaux différents, par exemple un intervalle plus large de 300 cents vers le bas (la tierce mineure utilisée par Curtis & Bharucha, 2010).
— Enfin, on pourrait utiliser d'autre méthodes pour mesurer l'affect, soit à partir de données psychophysiologiques comme les battements du coeur, les frissons ou l'activité musculaire, soit à partir de techniques d'imagerie cérébrale notamment la neuroimagerie fonctionnelle qui permettrait de caractériser les zones activées par les processus cognitifs de l'expérience.

Annexe A

Tableaux et Figures complémentaires

Item	Caractéristiques			Quantité
	Marque	Modèle	Autres	
Ordinateur	Dell	N/A	Système d'exploitation : Windows 10 Architecture : 64 bits Processeur : Intel i5 ∼ 3000MHz Mémoire RAM : 8Go	10
Casque audio	Beyerdynamic	770 PRO	80 Ω - 250 Ω	10
Carte son	RME	Fireface 400	N/A	1
Microphone (violon)	DPA	4061	Directivité : Omnidirectionnel Réponse en fréquence : 20Hz-20kHz Sensibilité : 6 mV/Pa	1
Microphone (voix)	DPA	4066	Directivité : Omnidirectionnel Réponse en fréquence : 20Hz-20kHz Sensibilité : 6 mV/Pa	1

Table A.1 – Matériel utilisé pour la conception et passation de l'expérience

N°	Artiste	Titre	Paroles
1	Enda Reilly	An Nasc Nua	Bí ann go fiáin
2	Triviul	Angelsaint	I guess I'm a grave digger, save your cinder. Nothing healthy to offer you
3	The Long Wait	Back Home To Blue	I'm going far away
4	*Inconnu*	Wallabout	We threw the key away as well as the good life
5	*Inconnu*	*Inconnu*	S'il est vrai que c'est ainsi, on me dit, merci
6	I Am Cassettes	Believe	The hands on my wrists, and my stubbly lips are here
7	Red Hot Chili Peppers	Under the bridge	I never worry, now that is a lie
8	BKS	Too much	Stinging from the lonliest touch
9	Spektakulatius	Our Love Is Here To Stay	Our love is here to stay
10	James May	Hold on you	And it's time to make your escape
11	Sound on Sound' (demo)	Mystery	Our mystery was meant to be complete
12	Nerve 9	Pray For The Rain	Pray for the rain to wash it away
13	Speak Softly	Broken Man	Try to find my way
14	Radiohead	Nude	There'll be something missing

Table A.2 – Caractéristiques des stimuli

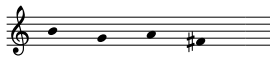
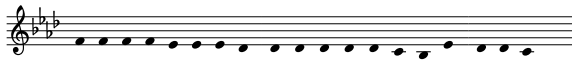
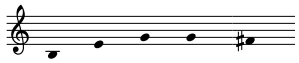
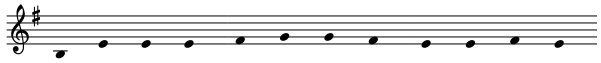
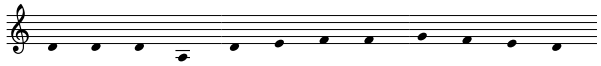
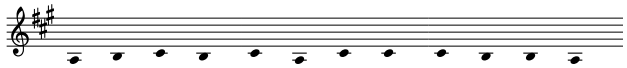
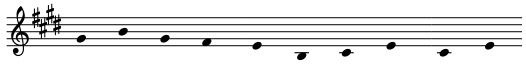
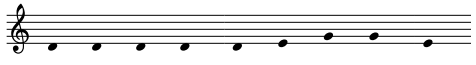
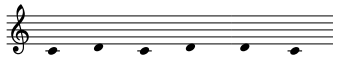
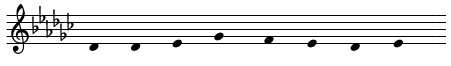
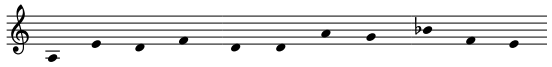
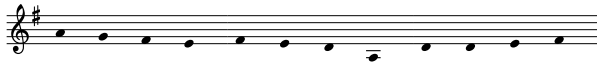
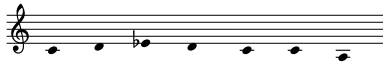
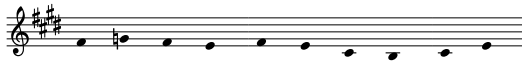
N°	Partition
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	

TABLE A.3 – Représentation des morceaux du corpus sous notation musicale, rapprochées à la gamme tempérée. Partitions en rythme libre

N°	Genre	Lange	Licence	D_M [s]	D_V [s]
1	Masculin	Irlandais	Ouverte	4.64	1.51
2	Masculin	Anglais	Ouverte	8.05	4.77
3	Féminin	Anglais	Ouverte	5.44	1.87
4	Féminin	Anglais	Unconnu	8.25	3.36
5	Féminin	Français	Unconnu	5.76	3.03
6	Masculin	Anglais	Ouverte	7.24	4.32
7	Masculin	Anglais	Fermée	7.86	3.27
8	Masculin	Anglais	Ouverte	3.81	2.74
9	Féminin	Anglais	Ouverte	6.66	2.04
10	Masculin	Anglais	Ouverte	6.75	3.01
11	Féminin	Anglais	Ouverte	8.13	2.76
12	Féminin	Anglais	Ouverte	7.47	2.73
13	Féminin	Anglais	Ouverte	4.23	1.73
14	Masculin	Anglais	Fermée	8.11	1.77

TABLE A.4 – Caractéristiques techniques de stimuli. D_M : Durée moyenne - musique, D_V : Durée moyenne - voix parlée

Nº	Plage F0 [Hz]			
	A cappella	**Voix parlée**	**Violon**	**Cris**
1	175 - 255	50 - 150	355 - 510*	447 - 477
2	240 - 370	50 - 150	220 - 360	566 - 596
3	180 - 410	140 - 220	180 - 410	717 - 752
4	270 - 413	140 - 220	235 - 410	449 - 479
5	255 - 407	150 - 360	265 - 405	569 - 599
6	190 - 300	50 - 150	210 - 290	722 - 755
7	100 - 295	50 - 150	205 - 590*	198 - 212
8	270 - 410	50 - 150	560 - 810*	250 - 270
9	216 - 320	140 - 220	250 - 310	315 - 338
10	122 - 190	50 - 150	265 - 385*	200 - 215
11	200 - 485	140 - 220	205 - 490	252 - 272
12	190 - 470	140 - 220	205 - 470	318 - 342
13	195 - 335	140 - 220	210 - 310	—
14	238 - 403	50 - 150	235 - 400	—

TABLE A.5 – Intervalle approximatif de F0 (Fréquence fondamentale estimée) pour les transformations de hauteur ('*up*', '*down*' et '*vibrato*') de chaque morceau par type de son.
* Morceaux octaviés pour l'enregistrement de violon.

Nº	Nom original du fichier	Phonème	Genre	Hauteur	Méthode
1	`F1a.pitch1.orig_transp.neutral.rmsNorm.wav`	[a]	Féminin	1	—
2	`F1a.pitch2.orig_transp.neutral.rmsNorm.wav`	[a]	Féminin	2	—
3	`F1a.pitch3.orig_transp.neutral.rmsNorm.wav`	[a]	Féminin	3	—
4	`F1i.pitch1.orig_transp.neutral.rmsNorm.wav`	[i]	Féminin	1	—
5	`F1i.pitch2.orig_transp.neutral.rmsNorm.wav`	[i]	Féminin	2	—
6	`F1i.pitch3.orig_transp.neutral.rmsNorm.wav`	[i]	Féminin	3	—
7	`M1a.pitch1.pitch2_transp_PaN_-400.neutral.rmsNorm.wav`	[a]	Masculin	1	PaN
8	`M1a.pitch2.medium_clean_sinModel.neutral.rmsNorm.wav`	[a]	Masculin	2	sinModel
9	`M1a.pitch3.pitch2_transp_PaN_+400.neutral.rmsNorm.wav`	[a]	Masculin	3	PaN
10	`M2a.pitch1.pitch2_transp_PaN_-400.neutral.rmsNorm.wav`	[i]	Masculin	1	PaN
11	`M2a.pitch2.medium_clean_sineModel.neutral.rmsNorm.wav`	[i]	Masculin	2	sinModel
12	`M2a.pitch3.pitch2_transp_PaN_+400.neutral.rmsNorm.wav`	[i]	Masculin	3	PaN

TABLE A.6 – Caractéristiques des cris. Le champ 'Méthode' fait référence aux algorithmes de synthèse utilisés par Liuni et al. pour générer les stimuli.

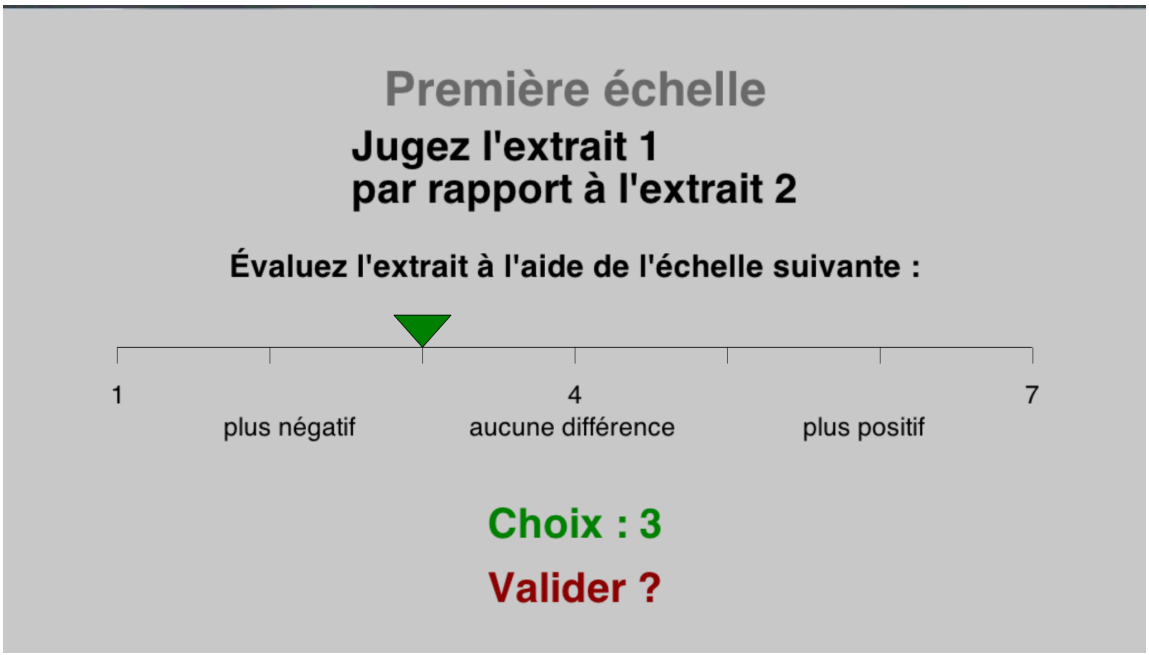


FIGURE A.1 – Capture d'écran de mesure de l'échelle de valence utilisé pendant l'expérience. L'écran affichait le numéro choisit par le participant, qui pouvait changer la note avec le clavier numérique, avant de valider.

Annexe B

Outils statistiques

On décrit par la suite les outils statistiques utilisés pour analyser les données expérimentales.

B.1 ANOVA

On peut décrire le modèle ANOVA comme un test statistique qui permet d'évaluer plus de deux groupes en même temps et éviter l'erreur *family-wise error*[1].

Les assomptions de base pour l'ANOVA sont : les populations suivent des distributions normales, elles ont une variance égale (homoscédasticité) et les observations sont indépendantes (Dodge, 2008). L'hypothèse nulle (H_0) est que toutes les moyennes du groupe sont égales. Si l'on rejette l'hypothèse nulle, l'hypothèse alternative (H_1) nous indique qu'au moins une moyenne est différente des autres, mais cela ne peut pas nous dire quelle moyenne est différente de laquelle (Lynch, 2013).

B.1.1 Modèle de régression

On peut regarder l'ANOVA comme la résolution d'un modèle mathématique de régression linéaire (Maxwell & Delaney, 2003) que s'écrit :

$$y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} + \epsilon_i \tag{B.1}$$

où y_i est l'observation de l'individu y sur la variable dépendante i, les X donnent information du niveau du sujet i sur les facteurs que l'on contrôle, chaque β indique la relation entre le facteur X et la variable dépendante, et ϵ est l'erreur par sujet i.

Une ANOVA est un cas spécial de ce modèle et peut aussi s'écrire (à un seul facteur) de manière simplifiée comme :

$$y_{i,j} = \mu + \gamma_i + \epsilon_{i,j} \tag{B.2}$$

où μ est la moyenne générale commune à toutes les conditions, γ_i est l'effet de la condition i sur l'observation et $\epsilon_{i,j}$ l'erreur expérimentale par observation $y_{i,j}$.

L'ANOVA nous permet de comparer les moyennes de groupes différents en utilisant la relation entre deux estimations de variance σ^2 pour chaque groupe. La première utilise la variance expliquée par le modèle (variabilité inter-classe) et la deuxième emploi la variance non expliquée par le modèle (variabilité intra-classe). La variance totale dans le modèle est calculée (à un facteur multiplicatif près) comme la différence entre la valeur observée et la moyenne globale, élevées au carrée, puis, additionnées. Cette somme s'appelle Somme des carrés des écarts SS_T :

1. Probabilité de trouver un effet alors qu'il n'existe pas car on réalise plusieurs fois le même test sur une même base de données.

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{\bar{y}})^2 \tag{B.3}$$

avec la moyenne globale de l'échantillon notée $\bar{\bar{y}}$ et la moyenne par groupe notée $\bar{y}$.

Si on écrit la somme des carrés du modèle, entre groupes (inter-classe ou *between*) :

$$SS_B = \sum_{g=1}^{G} n_g (\bar{y}_g - \bar{\bar{y}})^2 \tag{B.4}$$

où G est le nombre de groupes, n_g est le nombre de personnes dans le groupe g, $y_{i,g}$ est la valeur de y pour la $i$ème personne dans le groupe g.

On peut aussi écrire la somme des carrés de l'erreur, à l'intérieur des groupes (intra-classe ou *within*) comme :

$$SS_W = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{i,g} - \bar{y_g})^2 \tag{B.5}$$

Alors, on peut voir que l'équation B.3 peut s'écrire aussi comme :

$$SS_T = SS_B + SS_W \tag{B.6}$$

Ceci nous permet d'obtenir la relation entre la variance systématique par rapport à la variance non-systématique, avec le ratio F_{test} :

$$F_{\text{test}} = \frac{\frac{SS_B}{DDL_B}}{\frac{SS_W}{DDL_W}} \tag{B.7}$$

Les degrés de liberté DDL (nombre de coefficients moins nombre de contraintes) vont définir la forme de la distribution et permettent d'estimer les paramètres de la population à partir de l'échantillon. Pour le cas d'une ANOVA simple, on a $DDL_B = G - 1$ et $DDL_W = n - G$.

L'équation B.7 nous permet d'évaluer si la relation entre les variances est significative statistiquement, $p < 0.05$. La valeur p (souvent fixée à 5%) nous permet de définir un seuil de probabilité à partir duquel on peut trouver un effet alors qu'il n'y en a pas.

Si $F_{test} \neq 0$, alors on sait qu'il y a un effet mais on ne peut pas prédire son sens.

B.2 Taille d'effet

Le fait qu'un résultat soit statistiquement significatif ne veut pas nécessairement dire qu'il mesure un phénomène remarquable. L'importance des résultats est mesurée par la taille d'effet, qui est un type de calcul qui permet objectivement de connaître la magnitude de l'effet trouvé en fonction de la taille de la population étudiée.

$$\text{test} = \text{taille d'effet} \times \text{taille d'étude} \tag{B.8}$$

Il y a différentes méthodes pour calculer la taille d'effet. On va décrire brièvement les plus utilisées mais on va se concentrer sur celle appliquée dans l'analyse (chapitre 5).

B.2.1 *d* de Cohen

Il s'agit d'une différence entre les observations, divisée par leurs écart types. Le d de Cohen engendre une famille d'estimateurs construits avec le même principe. On le calcule grâce à la formule :

$$d = \frac{\mu_1 - \mu_2}{\sigma_\varepsilon} \tag{B.9}$$

où σ_ε est l'écart type commun intra-classe. Cependant, on travaille souvent avec une estimation de d :

$$d = \frac{\bar{y}_1 - \bar{y}_2}{S} \tag{B.10}$$

avec $S^2 = SS_W/DDL_W$. L'interprétation courante, suggérée par Cohen, pour différentes tailles d'effet d, est : 0.2 - 'faible', 0.5 - 'moyen' et 0.8 - 'fort' (Maxwell & Delaney, 2003).

B.2.2 *r* de Pearson

Appelé aussi coefficient de corrélation, il est utilisé pour décrire la relation de dépendance entre deux variables quantitatives aléatoires a et b ; il est calculé à partir de la multiplication de leurs covariances divisé par la multiplication de leurs écart types. Le r de Pearson engendre aussi sa propre famille d'estimateurs qui suivent l'association entre les variables. On le calcule à partir de :

$$r = \frac{\mathrm{cov}(a,b)}{\sigma_a \sigma_b} = \frac{\sum_i (a_i - \bar{a})(b_i - \bar{b})}{(N-1)\sigma_a \sigma_b} \tag{B.11}$$

Cet estimateur est défini entre $-1 \leq r \leq 1$, où $r \simeq 0$ signifie une absence de corrélation, $r \simeq -1$ implique une corrélation inverse et $r \simeq 1$ suggère une corrélation linéaire forte.

B.2.3 *eta-squared* η^2

Il s'agit d'un membre de la famille de taille d'effet de r, qui peut être utilisé pour plusieurs groupes observés. Cet indicateur mesures la proportion de la variance associée à l'effet par rapport à la variance totale. C'est la base du calcul de taille d'effet utilisé par la fonction `ezanova` du logiciel `R`, utilisé pendant l'expérience du stage.

$$\eta^2 = \frac{SS_{\text{effet}}}{SS_T} \tag{B.12}$$

Bien que *eta-squared* soit efficace pour comparer les tailles d'effet dans une même étude, la manipulation de variables additionnelles, qui augmente la variance totale SS_T empêche la compatibilité de cet estimateur. On utilise pour cela un estimateur partial, qui prend en compte la variance de l'erreur SS_{erreur} :

$$\eta_p^2 = \frac{SS_{\text{effet}}}{SS_{\text{effet}} + SS_{\text{erreur}}} \tag{B.13}$$

Generalized eta-squared η_G^2

Par ailleurs, pour adresser la même problématique décrite au-dessus, Olejnik et Algina (2003) a proposé une mesure généralisé de l'effet η^2, maintenant comparable à travers de plusieurs types de conceptions expérimentales. D'après Bakeman (2005), cet estimateur s'écrit :

$$\eta_G^2 = \frac{SS_{\text{effet}}}{\delta \times SS_{\text{effet}} + \sum_{\text{mesure}} SS_{\text{mesure}}} \tag{B.14}$$

avec $\delta = 1$ si l'effet d'intérêt est un facteur manipulé, et $\delta = 0$ sinon. La fonction `ezanova` du logiciel `R` utilise l'équation B.14 pour calculer la taille d'effet.

B.2.4 *omega-squared* ω^2

C'est une mesure proposé pour corriger le biais de η^2, qui estime la variance provenant d'un échantillon et non de la population entière. On calcule cet effet avec la formule :

$$\omega^2 = \begin{cases} \dfrac{DDL_{\text{effet}} \times (MS_{\text{effet}} - MS_{\text{erreur}})}{SS_{\text{total}} + MS_{\text{effet}}} & \text{si inter-classe;} \\ \dfrac{DDL_{\text{effet}} \times (MS_{\text{effet}} - MS_{\text{erreur}})}{SS_{\text{total}} + MS_{\text{sujets}}} & \text{si intra-classe.} \end{cases} \tag{B.15}$$

Comme précédemment, on peut augmenter la comparabilité entre études, en calculant la version partielle de cet estimateur, ω_p^2 comme :

$$\omega_p^2 = \frac{DDL_{\text{effet}} \times (MS_{\text{effet}} - MS_{\text{erreur}})}{DDL_{\text{effet}} \times MS_{\text{effet}} + (N - DDL_{\text{effet}}) \times MS_{\text{erreur}}} \tag{B.16}$$

avec MS_{effet}, MS_{erreur} et MS_{sujets} sont les moyennes des carrés (*mean squared*) de l'effet, de l'erreur et des sujets, respectivement.

Il faut remarquer qu'à cause de la complexité de cette mesure, les chercheurs sont encouragés à utiliser plutôt η_G^2, qui à une différence de biais négligeable par rapport à ω^2 (Lakens, 2013).

Références

Arias, P., Soladie, C., Bouafif, O., Robel, A., Seguier, R., & Aucouturier, J.-J. (2018). Realistic transformation of facial and vocal smiles in real-time audiovisual streams. *IEEE Transactions on Affective Computing*.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior research methods*, *37*(3), 379–384.

Barrett, L. F., Lewis, M., & Haviland-Jones, J. M. (2016). *Handbook of emotions*. Guilford Publications.

Bowling, D. L., Purves, D., & Gill, K. Z. (2017). Vocal similarity predicts the relative attraction of musical chords. *Proceedings of the National Academy of Sciences*, 201713206.

Bryant, G. A., & Barrett, H. C. (2008). Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture*, *8*(1), 135–148.

Curtis, M. E., & Bharucha, J. J. (2010). The minor third communicates sadness in speech, mirroring its use in music. *Emotion*, *10*(3), 335.

Darwin, C. (1872). *The expression of the emotions in man and animals* (1st edition éd.). John Murray.

Darwin, C. (1874). *The descent of man and selection in relation to sex* (2nd éd.). Murray.

Darwin, C. (1891). La descendance de l'homme et la sélection sexuelle (E. Barbier, Trad.). *Reinwald, Paris*. (d'après la seconde édition anglaise revue et augmenté par l'auteur (1874))

Davis, M. (1994). The role of the amygdala in emotional learning. In *International review of neurobiology* (Vol. 36, pp. 225–266). Elsevier.

Dodge, Y. (2008). *The concise encyclopedia of statistics*. Springer Science & Business Media.

Donington, R. (1963). The interpretation of early music. *New York*.

Ekkekakis, P. (2013). *The measurement of affect, mood, and emotion : A guide for health-behavioral research*. Cambridge University Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169–200.

Escoffier, N., Zhong, J., Schirmer, A., & Qiu, A. (2013). Emotional expressions in voice and music : same code, same effect ? *Human Brain Mapping*, *34*(8), 1796–1810.

Fan, Y.-T., & Cheng, Y. (2014). Atypical mismatch negativity in response to emotional voices in people with autism spectrum conditions. *PLoS One*, *9*(7), e102471.

Gentilucci, M., Ardaillon, L., & Liuni, M. (2018). Vocal distortion and real-time processing of roughness. In *Proc. international computer music conference (icmc), at daegu, korea.*

Hargreaves, D. J. (1984). The effects of repetition on liking for music. *Journal of research in Music Education*, *32*(1), 35–47.

Holt, F. (2007). *Genre in popular music*. University of Chicago Press.

Ilie, G., & Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception : An Interdisciplinary Journal*, *23*(4), 319–330.

James, W. (1884). What is an emotion ? *Mind*, *9*(34), 188–205.

Juslin, P. N. (2001). Communicating emotion in music performance : A review and a theoretical framework. In *Music and emotion : Theory and research* (p. 309-337). Oxford University Press.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance : Different channels, same code ? *Psychological bulletin*, *129*(5), 770.

Juslin, P. N., & Sloboda, J. (2011). *Handbook of music and emotion : Theory, research, applications*. Oxford University Press.

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music : The need to consider underlying mechanisms. *Behavioral and brain sciences*, *31*(5), 559–575.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science : a practical primer for t-tests and anovas. *Frontiers in psychology*, *4*, 863.

Laukka, P., Elfenbein, H. A., Söder, N., Nordström, H., Althoff, J., Iraki, F. K., ... Thingujam, N. S. (2013). Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. *Frontiers in Psychology*, *4*, 353.

Laukka, P., Juslin, P., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition & Emotion*, *19*(5), 633–653.

Law, L. N., & Zentner, M. (2012). Assessing musical abilities objectively : Construction and validation of the profile of music perception skills. *PloS one*, *7*(12), e52508.

Lewis, M. E., Haviland-Jones, J. M., & Barrett, L. F. E. (2008). *Handbook of emotions*. Guilford Press.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Liuni, M., Ardaillon, L., Lou, S., Vasa, L., & Aucouturier, J.-J. (2018). *Angus : a real-time acoustic transformation to simulate vocal arousal*. (en préparation)

Liuni, M., Ponsot, E., & Aucouturier, J. (s. d.). Vocal expression. In *Book of abstracts* (Vol. 1, p. 94).

Lorist, M. M., Klein, M., Nieuwenhuis, S., De Jong, R., Mulder, G., & Meijman, T. F. (2000). Mental fatigue and task control : planning and preparation. *Psychophysiology*, *37*(5), 614–625.

Lynch, S. M. (2013). Using statistics in social research. *New York : Springer), doi*, *10*, 978–1.

Ma, W., & Thompson, W. F. (2015). Human emotions track changes in the acoustic environment. *Proceedings of the National Academy of Sciences*, *112*(47), 14563–14568.

Magne, C., Schön, D., & Besson, M. (2006). Musician children detect pitch violations in both music and language better than nonmusician children : behavioral and electrophysiological approaches. *Journal of cognitive neuroscience*, *18*(2), 199–211.

Maxwell, S. E., & Delaney, H. D. (2003). *Designing experiments and analyzing data : A model comparison perspective*. Routledge.

McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2010). Individual differences reveal the basis of consonance. *Current Biology*, *20*(11), 1035–1041.

McDermott, J. H., Schultz, A. F., Undurraga, E. A., & Godoy, R. A. (2016). Indifference to dissonance in native amazonians reveals cultural variation in music perception. *Nature*, *535*(7613), 547.

McPherson, M. J., & McDermott, J. H. (2018). Diversity in pitch perception revealed by task dependence. *Nature Human Behaviour*, *2*(1), 52.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians : an index for assessing musical sophistication in the general population. *PloS one*, *9*(2), e89642.

Nagel, F., Kopiez, R., Grewe, O., & Altenmüller, E. (2007). Emujoy : Software for continuous measurement of perceived emotions in music. *Behavior Research Methods*, *39*(2), 283–290.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics : measures of effect size for some common research designs. *Psychological methods*, *8*(4), 434.

Parizet, E. (2006). Perception acoustique et qualité sonore. *Techniques de l'ingénieur. Mesures et contrôle*.

Patel, A. D. (2010). *Music, language, and the brain*. Oxford university press.

Plack, C. J. (2010). Musical consonance : The importance of harmonicity. *Current Biology*, *20*(11), R476–R478.

Puckette, M. (2007). *The theory and technique of electronic music*. World Scientific Publishing Company.

Putman, D. A. (1987). Why instrumental music has no shame. *The British Journal of Aesthetics*, *27*(1), 55-61. Consulté sur http://dx.doi.org/10.1093/bjaesthetics/27.1.55 doi: 10.1093/bjaesthetics/27.1.55

Rachman, L., Liuni, M., Arias, P., Lind, A., Johansson, P., Hall, L., ... Aucouturier, J.-J. (2018). David : An open-source platform for real-time transformation of infra-segmental emotional cues in running speech. *Behavior research methods*, *50*(1), 323–343.

Roebel, A. (2010). Shape-invariant speech transformation with the phase vocoder. In *Eleventh annual conference of the international speech communication association*.

Ross, D., Choi, J., & Purves, D. (2007). Musical intervals in speech. *Proceedings of the National Academy of Sciences*, *104*(23), 9852–9857.

Rossing, T. D. (2007). *Springer handbook of acoustics*. Springer.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion : dissecting the elephant. *Journal of personality and social psychology*, *76*(5), 805.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect grid : a single-item scale of pleasure and arousal. *Journal of personality and social psychology*, *57*(3), 493.

Russo, F. A., & Thompson, W. F. (2005). The subjective size of melodic intervals over a two-octave range. *Psychonomic bulletin & review*, *12*(6), 1068–1075.

Sakakibara, K.-I., Fuks, L., Imagawa, H., & Tayama, N. (2004). Growl voice in ethnic and pop styles. *Leonardo*, *1*, 2.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, *107*(6), 2408–2412.

Scherer, K. R. (1986). Vocal affect expression : A review and a model for future research. *Psychological bulletin*, *99*(2), 143.

Schön, D., Gordon, R. L., & Besson, M. (2005). Musical and linguistic processing in song perception. *Annals of the New York Academy of Sciences*, *1060*(1), 71–81.

Schubert, E. (2007). Real time cognitive response recording. In *The inaugural international conference on music communication science* (pp. 139–142).

Stel, M., & van Knippenberg, A. (2008). The role of facial mimicry in the recognition of affect. *Psychological Science*, *19*(10), 984.

Taschereau-Dumouchel, V., Cortese, A., Chiba, T., Knotts, J., Kawato, M., & Lau, H. (2018). Towards an unconscious neural reinforcement intervention for common fears. *Proceedings of the National Academy of Sciences*, *115*(13), 3470–3475.

Villavicencio, F., Robel, A., & Rodet, X. (2006). Improving lpc spectral envelope extraction of voiced speech by true-envelope estimation. In *Acoustics, speech and signal processing, 2006. icassp 2006 proceedings. 2006 ieee international conference on* (Vol. 1, pp. I–I).

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological bulletin*, *98*(2), 219.

Zwicker, E., & Fastl, H. (2013). *Psychoacoustics : Facts and models* (Vol. 22). Springer Science & Business Media.