



MÉMOIRE D'HABILITATION À DIRIGER DES RECHERCHES DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Faculté d'Ingénierie (UFR 919)

Présenté par

Jean-Julien AUCOUTURIER

Chargé de recherche CNRS
STMS UMR9912 (IRCAM/CNRS/UPMC), Paris

Sujet du mémoire :

**L'apport des sciences et technologies du son
à la recherche en sciences cognitives**

Habilitation soutenue le 28 novembre 2017

devant le jury composé de :

M. Michel BEAUDOIN-LAFON	LRI, Université Paris-Sud	Examinateur
Mme. Isabelle BLOCH	LTCI, Télécom ParisTech, Université Paris-Saclay	Rapportrice
Mme. Nathalie GEORGE	ICM, Université Pierre et Marie Curie	Examinateuse
M. Didier GRANDJEAN	Département de Psychologie, Université de Genève	Rapporteur
M. Jean-Claude MARTIN	LIMSI, Université Paris-Sud	Rapporteur
M. François PACHET	Spotify Creator Technology Research Lab	Examinateur

Table des matières

<i>Introduction</i>	7
<i>Reconnaissance</i>	11
<i>Simulation</i>	17
<i>Analyse</i>	23
<i>Synthèse</i>	29
<i>Interrogations</i>	37

“There is nothing so practical as a good theory.”

(Lewin, 1951)

“There’s nothing so theoretical as a good method.”

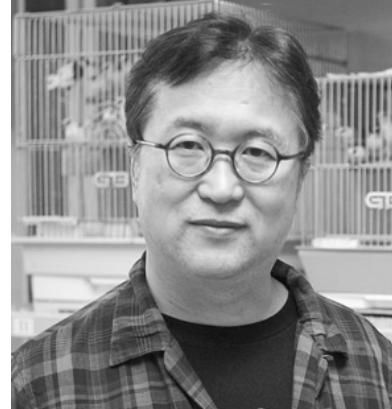
(Greenwald, 2012)

Introduction

LES TRAVAUX DE RECHERCHE RÉSUMÉS DANS CE MÉMOIRE couvrent une période d'une quinzaine d'années (2001-2017), allant de ma thèse effectuée au laboratoire Sony CSL Paris sous la direction de François Pachet et Jean-Pierre Briot à mes travaux actuels en tant que chargé de recherche CNRS au laboratoire Sciences et Technologies de la Musique et du Son (STMS UMR 9912), où je travaille depuis mon recrutement CNRS en Octobre 2012.

Ces travaux s'inscrivent dans deux domaines scientifiques : l'informatique (traitement du signal audio et reconnaissance de formes principalement) et les sciences cognitives (psychologie et neurosciences affectives et sociales). Le premier de ces domaines correspond à ma formation initiale (ingénieur signal, thèse d'informatique), le second à une évolution initiée lors de mes années postdoctorales au Japon (2006-2011) dans les laboratoires de Takashi Ikegami, Katsumi Watanabe et de Kazuo Okanoya. Même si mes travaux actuels dans le cadre du projet ERC Starting Grant CREAM (*Cracking the emotional code of music*, 2014-2019) combinent des apports de ces deux disciplines, le ratio entre ces approches a largement varié au cours de ma carrière : quasiment purement informatique de 2001 à 2008, puis purement expérimentale de 2008 à 2014, et enfin - et ce sera in fine le propos de ce mémoire - intégré de façon constructive aujourd'hui.

Malgré une différence de surface entre ces deux types de contribution (Deux titres d'articles pour en témoigner - 2007 : *A scale-free distribution of false positives for a large class of audio similarity measures*; 2013 : *Music improves verbal memory encoding while decreasing prefrontal cortex activity : a fNIRS study*), tous les travaux résumés dans ce mémoire sont l'expression d'un même domaine d'intérêt : celui de la cognition du son, et en particulier des signaux communicatifs que sont la voix et la musique. Ce sont des questions de cet ordre qui m'ont séduit en 2001 dans la jeune communauté informatique de *Music Information Retrieval* (qui se promettait, par exemple, de trouver automatiquement des morceaux de musique qui "sonnent" comme les Beatles ou les Rolling Stones), puis en 2006 dans l'approche



Trois mentors japonais. Takashi Ikegami (Université de Tōkyō) est spécialiste de simulation de systèmes dynamiques non-linéaires et étudie les processus d'émergence de comportements biologiques. Katsumi Watanabe (Université Waseda) est psychophysiologiste, spécialiste de cognition visuelle. Kazuo Okanoya (RIKEN Brain Science Institute, Université de Tōkyō) est neuroscientifique et étudie les comportements de communication chez l'oiseau et le rongeur.

de simulation *Artificial Life* appliquée aux comportements humains (pour expliquer, par exemple, l'émergence des comportements de turn-taking dans le langage ou la diversité culturelle des systèmes tonaux musicaux), et enfin dans le “mélange” que je pratique aujourd’hui entre algorithmes de synthèse sonore et expérimentation en psychologie/neurosciences cognitives, et qui pose la question du traitement émotionnel par le cerveau des indices transmis par la musique et la voix.

Ce qui a varié, au cours de ces années, n'est donc pas l'objet mais la méthode. Tous ces travaux procèdent en effet d'une seule quête personnelle, qu'on pourrait qualifier d'épistémologique : comprendre/découvrir comment faire de la preuve scientifique (dans le domaine des sciences cognitives de la parole et de la musique) avec les outils de l'analyse et synthèse de signaux sonores. Ma formation initiale est celle d'un informaticien, et non d'un expérimentateur en sciences du vivant¹. Cette interaction n'est donc pas allée de soi : de mes premiers articles dans des conférences d'informatique qui proclamaient, à la va-vite dans un paragraphe de conclusion, le besoin d'aller vers des modèles “plus plausibles cognitivement”² (sans trop savoir ce que ceux-ci pourraient être), à la publication cette année d'un article sur la cognition sociale des interactions musicales dans la revue *Cognition*³, il a fallu des apprentissages variés : ANOVAs et modèles à effets mixtes ont progressivement remplacé dans mon travail les *R-precision* et *F-measure* du machine learning ; le temps passé en cabine à poser des électrodes d'électroencéphalographie (EEG) et des optodes de spectroscopie fonctionnelle proche-infrarouge (fNIRS), celui passé précédemment à déboguer du code Matlab ; et les livres de neuro-anatomie ont succédé à ceux de reconnaissance de formes. Mais bien plus que de nouvelles techniques, c'est surtout un nouveau langage, et avec lui une nouvelle façon de penser, qu'il m'a fallu apprendre - ceux de la communauté des sciences expérimentales. Au fil des manuscrits rejetés, par l'éditeur, puis progressivement après relecture, j'ai appris à comprendre ce qui, pour les sciences cognitives, constituait un problème bien posé, puis même éventuellement un problème intéressant. J'ai abordé cet apprentissage en informaticien, comme on étudierait des *design patterns* ou le style d'un nouveau langage de programmation (on n'écrit pas du Python en 2017 comme on écrivait du Java en 2002). C'est sur ce parcours vers une compréhension plus fine des échanges possibles entre informatique et sciences expérimentales que ce mémoire se propose de faire le point, moins à la façon d'une conclusion définitive que d'un *commit -m*, un point de sauvegarde temporaire, avec l'espoir que le lecteur y trouvera matière à éclairer sa propre pratique.

1. J'alterne ici plusieurs qualificatifs pour décrire la même chose : la psychologie cognitive, la psychoacoustique, les neurosciences cognitives sont ce que j'englobe sous le terme “sciences cognitives”, et j'entends en cela une science expérimentale (= qui collecte des données sur des participants humains) que j'inscris, d'un point de vue à la fois méthodologique et théorique, dans les sciences du vivant. Cette classification reprend peu ou prou l'organisation en section du CNRS, où le champ d'étude du “cerveau, cognition, comportement” (l'actuelle section 26) est rattaché à l'Institut des Sciences Biologiques (INSB). L'Institut des Sciences de l'Information (INS2I), dont dépend le champ d'étude du traitement du signal sonore et de la reconnaissance de formes (section 7), reconnaît quant à lui, parmi ses multiples interfaces avec les autres disciplines, l'interaction avec ces “sciences du vivant” - c'est dans ce cadre que s'inscrivent les travaux décrits ici.

2. Aucouturier, J.-J. and Pachet, F. (2002) *Music Similarity Measures : What's the Use ?*. Proceedings of the International Symposium on Music Information Retrieval (ISMIR), Paris, France
 3. Aucouturier, JJ. & Canonne, C. (2017) *Musical friends and foes : the social cognition of affiliation and control in musical interactions*. Cognition, vol. 161, 94-108.

Avec le recul de cette période, je garde une profonde gratitude pour tous ceux qui ont pris le temps de *m'expliquer*, à différentes étapes du chemin : un éditeur de la revue *Cognitive Science* en 2007 qui, à un set de *reviews* désastreuses, a pris sur lui d'ajouter une *decision letter* de 2 pages pour me traduire ce rejet en termes qui me seraient compréhensibles, et pointer avec bienveillance ce qui, pour le lectorat de la revue, faisait défaut dans mon raisonnement ; un directeur de labo qui, sans doute un peu exaspéré lui-même, m'a assis sur une chaise et a épingle dans mon vocabulaire tous les automatismes d'ingénieur qui ne passaient pas auprès d'un public de psychologues⁴ ; un collègue postdoctorant en psychologie expérimentale qui a eu la patience de collaborer avec moi pour une expérience qui, de notre première rencontre à Tōkyō en 2008 à sa publication l'année dernière, allait prendre 8 ans ; le directeur de labo japonais enfin qui, le premier, a accepté de m'intégrer à son institut de neurosciences alors que j'avais tout à apprendre ; je ne nomme personne et la liste est certainement incomplète, mais toute l'interdisciplinarité dont je peux éventuellement me targuer aujourd'hui, je la dois à leur confiance et leur ouverture d'esprit⁵.

Ce mémoire est organisé de façon thématique et chronologique, car ces deux façons se recoupent commodément dans mon cas. Pour explorer l'apport des sciences et technologies du son à la recherche en sciences cognitives, j'ai d'abord tenté d'utiliser les outils de la reconnaissance de formes typique de la communauté de *Music Information Retrieval* (chap. 1 - **Reconnaissance**), puis ceux de la simulation computationnelle et multi-agent inspirés de la communauté *Artificial Life* (chap. 2 - **Simulation**). J'ai ensuite exploré les outils de l'analyse acoustique, comme ce qui peut être fait en biolinguistique (chap.3 - **Analyse**), et finalement, dans une tradition chère à l'IRCAM qui héberge mes travaux depuis 2012, ceux de la synthèse et de la transformation de signaux (chap. 4 - **Synthèse**). Je conclus ces 4 chapitres par une série de questions que tout cela me pose pour la suite, et que je vous soumets au moins autant qu'à moi-même.

4. **A.O.** “[...] If I can make a small request, and I'm saying this in part jokingly but not solely, it would help indeed if you guys could at least stop using the word *semantics*”
M.C. “Wait... What?”

A.O. “*Semantics* - as in a semantic model of genre classification, *mixing acoustics with semantics, semantic gap*. Just, what do you mean by this?”

M.C. “Well, I suppose we take it as the *high-level* meaning of music, like saying *rock* is semantically related to youth and rebellion, electric guitars, all that linguistic and social knowledge around music. All which is where perception stops and, err..., cognition kicks in ? Activating the semantic networks of musical concepts, err...”

A.O. “See : that. We hate it where you do that.”

Extrait de Aucouturier, J.-J. and Bigand, E. *Mel Cepstrum & Ann Ova : The Difficult dialog between Music Information Retrieval and Cognitive Psychology*, Proceedings of the 2012 International Conference on Music Information Retrieval, 2012.

5. À cette liste, je me dois bien sûr d'ajouter les professeur.e.s Isabelle Bloch, Didier Grandjean et Jean-Claude Martin, qui m'ont fait l'honneur de rapporter ce mémoire au cours de l'été 2017 ; les professeur.e.s Michel Beaudoin-Lafon, Nathalie George et François Pachet, pour avoir accepté de se joindre à mon jury en novembre 2017 ; et tous ceux qui m'ont fait l'amitié de relire et commenter les versions précédentes de ce manuscrit : Gérard Assayag, Isa Aucouturier, Frédéric Bevilacqua, Clément Canonne, Nicolas Obin et Patrick Susini. Merci à tous.

Reconnaissance

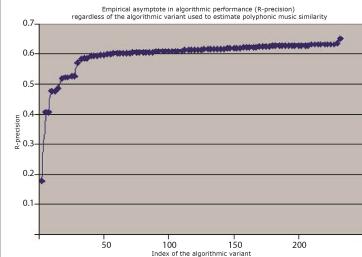
MES PREMIERS TRAVAUX EN INFORMATIQUE prennent racine dans l'émergence, dans les années 2000–2001, de la jeune discipline de Fouille d'Information Musicale (*Music Information Retrieval*, MIR). Dans une série d'articles publiés pendant mon master, avec Mark Sandler (King's College London), puis mon doctorat, avec François Pachet (SONY CSL) et Jean-Pierre Briot (LIP6, UPMC), nous avons établi un nouveau paradigme computationnel pour calculer des similarités acoustiques entre signaux musicaux, où chaque morceau est représenté par la distribution statistique de ses caractéristiques spectrales instantanées, et les distributions comparées deux à deux par des mesures de probabilités. Cette architecture, que nous avons baptisée *bag-of-frames* (par analogie avec le modèle *bag-of-words* en recherche d'information qui représente un texte par le set non structuré de ses mots sans préserver leur position d'occurrence), s'instancie par exemple en prenant comme caractéristiques les premiers coefficients mel-cepstraux (MFCCs) de chaque trame de 50ms, en apprenant leur distribution statistique par un modèle de mélange de gaussiennes (GMM) dont les paramètres sont évalués pour chaque morceau par maximum likelihood, et en comparant les GMMs deux à deux avec une approximation par Monte-Carlo de la divergence de Kullback-Leibler⁶.

Dans le contexte de la jeune communauté MIR, ces travaux sur la similarité timbrale ont fait tâche d'huile : le problème tel que nous l'avons identifié est devenu une des tâches centrales de la discipline, et est inclus dans la campagne annuelle d'évaluation MIREX depuis 2006 (année où notre algorithme a également obtenu la première place dans cette compétition). En 2009, notre article “Music Similarity Measures : What’s the use ?” était le deuxième le plus cité dans la communauté (source : Lee, Jones & Downie , 2009).

La fin de mon doctorat et le début de mes travaux postdoctoraux ont été marqués par plusieurs découvertes expérimentales qui ont remis en question l'utilisation faite jusqu'alors en MIR de ces algorithmes bag-of-frames. D'une part, notre étude JNRSAS 2004 (avec François Pachet) a mis en évidence l'existence d'un “plafond de verre”

6. Aucouturier, JJ., Pachet, F. & Sandler, M. (2005) *The Way It Sounds : Timbre Models For Analysis and Retrieval of Polyphonic Music Signals*. IEEE Transactions of Multimedia, 7(6) :1028-1035

Encadré 1 : Le plafond de verre



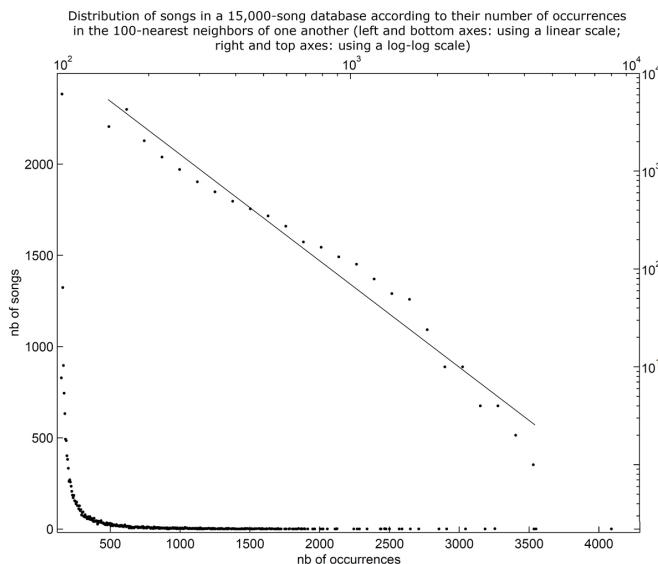
Les premières années de recherche en *Music Information Retrieval* avaient formé l'intuition que le jugement humain de similarité entre 2 morceaux de musique polyphonique était optimalement approximable par des modèles *bag-of-frame* des signaux. Notre étude JNRSAS'04 testa cette hypothèse en créant une des premières bases de test standardisée du domaine, et en y testant plus de 200 variétés algorithmiques du même paradigme computationnel - mettant en évidence une asymptote empirique à 65% de R-précision qui invalida l'intuition de la communauté.

Aucouturier, J.-J. & Pachet F. *Improving Timbre Similarity : How high is the sky ?*. Journal of Negative Results in Speech and Audio Sciences, 1(1), 2004.

(*glass ceiling*) dans la précision des ces algorithmes (Encadré 1). Le terme “glass ceiling” est depuis entré dans le vocabulaire courant de la discipline⁷.

D'autre part, en 2008, notre article *Pattern Recognition* (avec François Pachet) a montré que les erreurs de Type-I des algorithmes de similarité sont distribuées comme une loi de puissance : un petit nombre d'instances (que nous avons appelés des *hubs*) concentre la majorité des erreurs (Encadré 2). Suite à notre mise en évidence, le “problème des hubs” a dépassé la seule communauté MIR et est devenu une problématique générale en *machine learning*.⁸

Encadré 2 : Le problème des hubs



Le fait, pour la première fois, de pouvoir manipuler de gros volumes de données (à l'époque, plusieurs dizaines de milliers de mp3s) nous a permis de mettre en évidence une distribution atypique des erreurs de type I associées aux modèles de similarité timbrale MIR. Un très grand nombre de morceaux, que nous avons baptisés *hubs*, sont évalués à tort parmi les 100 plus proches voisins d'un très grand nombre d'autres morceaux. Notre étude *Pattern Recognition* 2007 sur le sujet a diagnostiqué le problème : nous avons par exemple montré que les *hubs* sont distribués selon une loi à invariance d'échelle et qu'ils sont une propriété structurelle des algorithmes, et non des chansons elles-mêmes.

Aucouturier, J.-J. & Pachet, F. *A scale-free distribution of false positives for a large class of audio similarity measures*. *Pattern Recognition*, vol. 41(1), pp. 272-284, 2007.

7. voir par , *Looking through the 'glass ceiling' : a conceptual framework for the problems of spectral similarity*, Nanopoulos et al., 2010

8. voir par exemple Radovanovic et al. *Hubs in space : popular nearest neighbors in high-dimensional data*, *Journal of Machine Learning Research*, 2010 ; Tomašev, N. & Buza, K. *Hubness-aware kNN classification of high-dimensional data in presence of label noise*, *Neurocomputing*, 2015

Quelques autres résultats de cette période :

Aucouturier, JJ., Defreville, B. & Pachet, F. (2007) *The bag-of-frame approach to audio pattern recognition : A sufficient model for urban soundscapes but not for polyphonic music*. *Journal of the Acoustical Society of America*, 122(2) :881-91.

Aucouturier, JJ. & Pachet, F. (2007) *The influence of polyphony on the dynamical modelling of musical timbre*. *Pattern Recognition Letters*, vol. 28 (5), pp.654-661.

Aucouturier, JJ., Pachet, F. & Sandler, M. (2005) *The Way It Sounds : Timbre Models For Analysis and Retrieval of Polyphonic Music Signals*. *IEEE Transactions of Multimedia*, 7(6) :1028-1035.

Aucouturier, JJ. & Pachet, F. (2003) *Representing Musical Genre : A State of the Art*. *Journal of New Music Research*, 32(1).

Ces premiers résultats MIR, semblant imiter convenablement certains jugements humains sur la musique (le genre, les émotions, la similarité timbrale, etc.), m'avaient donné l'intuition qu'on devrait pouvoir se servir de ces outils pour comprendre quelque chose sur la cognition humaine. Cependant, mes premiers essais pour construire des arguments cognitifs sur la base de tels résultats computationnels se sont avérés difficiles. J'avais, par exemple, soumis à une revue de sciences cognitives un résultat qui me semblait pertinent pour cette communauté⁹ : nous avions étudié un corpus assez important de 10,000 chansons, chacune annotée selon plus de 800 *tags* musicaux (est-ce du *rock*, de la *pop*, y a-t-il de la guitare, est-ce une bonne chanson pour conduire sur l'autoroute, pour un repas aux chandelles, etc. ?), et avions soumis chacune de ces annotations à des algorithmes de reconnaissance MIR de l'état de l'art. Nous avions constaté que, malgré cette technicité computationnelle, la grande majorité de ces tags n'étaient pas reconnus au-delà du hasard. Cela était vrai à la fois pour les annotations qui semblaient effectivement très haut-niveau et contextuelles (*old-fashioned*, *danceable*, *makes me want to drive a car fast*) mais aussi, de manière surprenante, pour des catégories comme le genre, l'émotion ou même les instruments, que l'intuition place pourtant plus proches du signal.

Notre argument était le suivant : vu que nous avions utilisé des optimisations sophistiquées provenant de l'effort collectif de plus de dix ans de recherche dans la communauté MIR, ce constat d'échec avait probablement une signification cognitive : il y avait sans doute beaucoup moins d'information que l'on croyait dans le “son” des musiques pour expliquer nos jugements cognitifs quotidiens. A quel point ces jugements étaient-ils conditionnés par la perception auditive, et à quel point étaient-ils au contraire construits extrinsèquement, à partir de stimuli ambigus ou même arbitraires (*How much audition involved in everyday categorization of music ?* demandions-nous dans le titre du manuscrit).

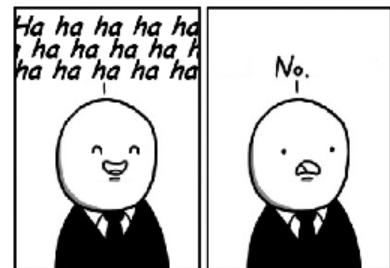
Le manuscrit fut envoyé en relecture, mais rejeté unanimement : “comment pouvez-vous exclure, disaient les commentaires, que cet échec à catégoriser la musique n'est pas simplement une insuffisance de l'algorithme ? Comme vous ne vous basez pas sur des jugements humains, mais sur une simulation algorithmique de jugements humains, vos résultats disent possiblement quelque chose sur les propriétés des algorithmes mais pas sur les propriétés de la cognition humaine. Vous devriez soumettre ce travail à une revue d'ingénierie”. Ces derniers mots étaient particulièrement révélateurs : ce n'était pas que notre travail était évalué comme étant de la mauvaise psychologie cognitive (ce qu'il était probablement, pour être honnête), mais comme n'en étant pas du tout. Il n'y avait pas de bug, ça ne compilait même pas.

9. Ce résultat a finalement été publié dans :

Aucouturier, J.-J. (2009) *Sounds like Teen Spirit : Computational Insights into the Grounding of Everyday Musical Terms*. In (Minett, J. and Wang, W. eds.) *Language, Evolution and the Brain, Frontiers in Linguistics Series*, Taipei : Academia Sinica Press

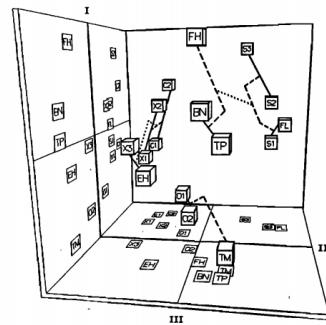
“It seems to me that a more reasonable title and framing would go something like this. Title : ‘How much of everyday music categorization can be captured by conventional measures of acoustic similarity between songs ?’ Framing : Music information retrieval is a thriving field, but it is largely based on simple methods for measuring the acoustic similarity of two songs. These similarity measures capture a range of intrinsic sound properties roughly corresponding to musical ‘timbre’, which can be computed from a brief 200 msec snapshot of a song, but they ignore other aspects of a song that require temporally extended observations, such as its pitch, harmony or rhythm. Here we ask how well these conventional metrics of timbral similarity perform at capturing a wide range of everyday musical categories from nonexpert listeners. We show that they do not perform very well for most kinds of categorization. The conclusion is either that music information retrieval needs to develop richer measures of acoustic similarity, or that everyday musical categorization is not based heavily on acoustic similarity, or perhaps both. Future research on better measures of sound similarity will be necessary to tease apart these alternatives.”

- un reviewer anonyme de l'article.



Source : @thirdreviewer

Avec le recul, je vois deux problèmes avec ce type d'argument et qui se retrouvent encore aujourd'hui dans beaucoup d'essais de convergence entre les techniques de reconnaissance automatique et les sciences cognitives. Premièrement, les résultats MIR sont souvent interprétés selon une métaphore psychoacoustique qui n'est, en fait, pas adaptée aux objets étudiés. En première approximation, les deux approches se ressemblent pourtant beaucoup : d'un côté, les études psychoacoustiques traditionnelles, par exemple du timbre musical¹⁰, collectent des jugements de similarité moyennés sur un échantillon de participants, puis sélectionnent des caractéristiques acoustiques qui expliquent la plus grande partie de la variance des données - montrant, par exemple, que le centre de gravité spectral d'un extrait corrèle à 93% avec la première dimension de l'espace reconstruit par *multi-dimensional scaling* (Grey, 1977). MIR, d'un autre côté, montre que des caractéristiques comme Liu & Zhang's *average autocorrelation peak*¹¹ ou Alluri & Toivainen's *6th band spectral flux*¹² permettent de prédire ou de classifier les émotions d'un morceau de musique avec 95% de précision - ce deuxième procédé semble donc similaire, voire même supérieur car il utilise des caractéristiques et des algorithmes de mapping plus sophistiqués (par exemple en se plaçant dans un manifold optimal par SVM, plutôt que dans un espace 2D par MDS ou PCA) et plus de données (la psychoacoustique considère rarement plus de 100 stimuli, alors qu'une étude typique en MIR en utilise des dizaines de milliers). Il peut donc sembler incompréhensible, voire même franchement rétrograde, que la communauté des sciences cognitives accepte comme "psychologiquement valide" le résultat du premier processus (le centre de gravité spectral pour le timbre) tout en ignorant ou rejetant les résultats du second. En vérité, elle a de bonnes raisons de les rejeter : la méthodologie psychoacoustique est conçue pour étudier des *percepts*, c'est-à-dire les gestalts psychologiques instantanés qui correspondent aux quelques caractéristiques physiques qui définissent un objet sonore indépendamment de l'auditeur et de sa culture. Un son musical donné "a" un pitch, une durée, une intensité et un timbre, et ces sensations peuvent être étudiées par la psychoacoustique. Ce même son, cependant, "n'a pas" un genre musical et une émotion - ces caractéristiques sont des jugements, construits cognitivement, et dont la valeur pourrait changer (un même morceau pourrait être décrit comme de la *pop*, plutôt que du *rock*) sans changer la définition physique du son. Même s'il existe un débat en psychologie cognitive concernant les processus qui peuvent vraiment être décrits comme de la perception et ceux relevant de processus cognitifs décisionnels de plus haut-niveau¹³, la plupart des chercheurs du domaine tombent d'accord pour reconnaître une différence fondamentale entre



La solution MDS 3D historique de J.M. Grey, calculée pour 16 timbres instrumentaux comparés par 35 participants (Grey, 1977).

10. Voir par exemple Grey, J. M. (1977). *Multidimensional perceptual scaling of musical timbres*. the Journal of the Acoustical Society of America, 61(5), 1270-1277 ; McAdams, S., Winsberg, S., Donnadieu, S., Soete, G., & Krimphoff, J. (1995). *Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes*. Psychological research, 58(3), 177-192.
11. Liu, D., Zhang, H.J. (2006) *Automatic mood detection and tracking of music audio signal*. IEEE Transactions on Speech and Audio processing 14(1), 5-18
12. Alluri, V., Toivainen, P. (2010) *Exploring perceptual and acoustic correlates of polyphonic timbre*. Music Perception 27(3), 223-241
13. Voir par exemple : Firestone, C. & Scholl, B. J. (2016). *Cognition does not affect perception : Evaluating the evidence for 'top-down' effects*. Behavioral and brain sciences, 39.

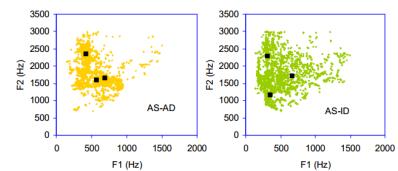
le pitch et le genre¹⁴. Selon eux, les résultats MIR appliquent donc une métaphore psychoacoustique (chercher des corrélats acoustiques) à des comportements (l'évaluation d'un genre, d'une émotion) auxquels elle ne s'applique pas. Une façon plus raisonnée de décrire les résultats des algorithmes MIR de classification de genre ou d'émotion serait en fait de reconnaître que ceux-ci ne modélisent pas ce que sont les chansons "jazz" ou "tristes", mais plutôt les choses qui "sonnent comme du jazz" ou "comme une chanson triste". Il en découle que les bons résultats obtenus par ces algorithmes ont beaucoup moins à dire sur le fonctionnement cognitif humain que sur une certaine réalité sociologique : la musique est une activité humaine qui obéit à des conventions culturelles relativement stables¹⁵ et la plupart des chansons "tristes" se ressemblent donc beaucoup (timbres sombres, mélodies graves, harmonies mineures, etc.). Ces caractéristiques ne rendent pas une chanson nécessairement triste, et on peut trouver de nombreux contre-exemples¹⁶. Mais comme ces exemples sont proportionnellement rares dans les corpus d'apprentissage, les algorithmes MIR peuvent afficher des performances de 90% sans, en fait, parvenir à modéliser quoi que ce soit sur la façon dont un jugement de genre ou d'émotion est construit cognitivement.

Deuxièmement, il me semble que l'on se trompe sur le type de preuve scientifique que peuvent établir les algorithmes de reconnaissance audio de la communauté MIR. Plutôt que de les prendre à tort pour des modèles cognitifs ou perceptifs d'un jugement humain, les algorithmes MIR gagneraient à être considérés comme des mesures physiques de l'information disponible au traitement humain dans un stimulus pour une tâche donnée. Cette mesure peut ensuite être utilisée pour construire des "preuves de faisabilité" dans une variété de situations scientifiquement importantes. Par exemple, dans le domaine du développement, de Boer & Kuhl¹⁷ ont montré que des algorithmes de reconnaissance de parole à base de GMMs obtiennent de meilleurs résultats de reconnaissance de phonèmes quand ils sont entraînés et testés sur du *motherese* (*infant-directed speech*, "mamanais" en français) que sur du langage adulte. Ce résultat apporte un fondement à l'argument selon lequel la fonction développementale du *motherese* est de servir d'amorçage à l'apprentissage du langage par l'enfant. Dans cet argument computationnel, l'algorithme n'est pas présenté comme un modèle cognitif : les auteurs ne prétendent pas que le cerveau du bébé implémente un GMM. Ce résultat donne seulement une preuve de faisabilité, cad que d'un strict point de vue physique, l'information est disponible dans le signal du *motherese* pour permettre un traitement linguistique favorisé par rapport à la prosodie adulte. Un argument similaire est fait par Frédéric Kaplan et ses collègues, qui montrent que des algorithmes de reconnaissance de formes au-

14. "I think to be published in a cognitive science journal, the authors would need to focus on a particular set of descriptors that are of cognitive interest. Musical genres are important for commercial application, and of interest from the standpoint of musicology, but I'm not sure they are of cognitive interest" - un reviewer anonyme

15. Serra, J., Corral, A., Boguna, M., Haro, M., Arcos, J.L. (2012) *Measuring the evolution of contemporary western popular music*. Scientific Reports 2.

16. Le plus radical est sans doute celui, chez les tziganes de Transylvanie, d'une même chanson utilisée successivement aux mariages... et aux enterrements ! - Bonini, F. (2009) *All the pain and joy of the world in a single melody : A transylvanian case study on musical emotion*. Music Perception 26(3), 257-261



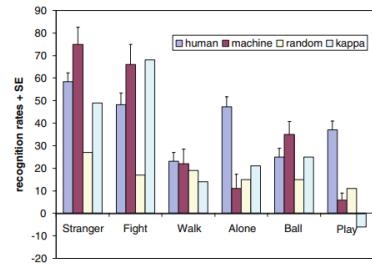
Deux modèles de phonèmes appris sur la prosodie adulte (gauche) et le motherese (droite). Les catégories apprises (dont on voit les centres des gaussiennes dans l'espace formantique) sont plus discriminantes et plus performantes sur le motherese. (de Boer & Kuhl, 2003)

17. De Boer, B., Kuhl, P. (2003) *Investigating the role of infant-directed speech with a computer model*. Acoustics Research Letters Online 4(4), 129-134

dio peuvent catégoriser des aboiements de chiens selon leur contexte d'enregistrement (inquiétude, jeu, etc.) et établissent ainsi que les vocalisations de chiens contiennent de l'information contextuelle¹⁸ (que cette information soit utilisée ou non par l'auditeur est une autre question).

La différence entre ce type de preuve et l'argument rejeté de notre article ci-dessus est subtile mais importante. Ce dernier essayait de faire une preuve d'hypothèse négative : "si les modèles MIR basés sur un certain type de caractéristiques sonores ne peuvent pas facilement classifier ceci ou cela, alors cela indique que, contrairement à notre intuition, le traitement humain ne peut pas être basé sur ces caractéristiques". L'erreur dans ce raisonnement tient au fait qu'il est impossible de prouver que la cognition humaine n'a absolument aucun moyen d'exploiter ce type de caractéristiques d'une façon qui soit inaccessible aux algorithmes. Une conclusion tout aussi valide (et sans doute plus probable) tirée de la même observation est, beaucoup plus trivialement, que les modèles MIR testés ne traitent pas la musique de la même façon que les humains. Les arguments utilisés par de Boer et Kaplan partent du point de vue opposé : ils cherchent à montrer que, contrairement à notre intuition, un certain ensemble de caractéristiques ou processus sont suffisants (plutôt que : non suffisants) pour permettre la tâche. Il s'agit en effet d'une question expérimentale ouverte que de savoir si un aboiement de chien ou un cri d'enfant contient suffisamment d'information acoustique pour communiquer son contexte ; après tout, le contexte peut être décodé à partir d'autres indices qui co-occurrent avec le son, comme la posture ou la situation, et il est difficile de contrôler que ces indices-là ne sont pas traités par un participant humain, même dans un contexte expérimental. A l'inverse, si on restreint les participants à n'écouter que des stimuli acoustiques et qu'ils n'arrivent pas à faire la tâche, on peut toujours opposer que celle-ci n'est pas écologique, et qu'avec suffisamment d'apprentissage, ils y parviendraient peut-être. La seule façon "propre" de tester cette hypothèse de faisabilité est de construire un classificateur naïf capable d'exploiter l'information acoustique de façon optimale - ce qui est précisément ce que les algorithmes MIR permettent de faire. La psychologie contemporaine est friande de telles explications parcimonieuses¹⁹, et ce type de preuve peut être appliquée à une multitude de problèmes importants en cognition du son en particulier : comment les auditeurs d'une culture peuvent-ils décoder les émotions exprimées dans la langue ou les émotions d'autres cultures ; comment les non-musiciens détectent des déviations dans des séquences d'accord alors qu'ils n'ont pas de connaissance formelle sur l'harmonie ; comment se forme-t-on une représentation cohérente d'un genre musical sur la base de seulement deux ou trois exemples, etc.

18. Molnar, Kaplan, Roy, Pachet, Pongracz, Doka & Miklosi (2008) *Classification of dog barks : a machine learning approach*. Animal Cognition 11(3), 389-400



Performance de reconnaissance du contexte de 6000 aboiements de chiens par un classifieur Bayésien naïf (performance au hasard : 16.7% ; Molnar et al, 2008)

19. Voir par exemple Gigerenzer, G., Todd, P.M. (1999) *Simple heuristics that make us smart*. New York : Oxford University Press ; Teglas, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J.B., Bonatti, L.L. (2011) *Pure reasoning in 12-month-old infants as probabilistic inference*. Science 332, 1054-1059

Simulation

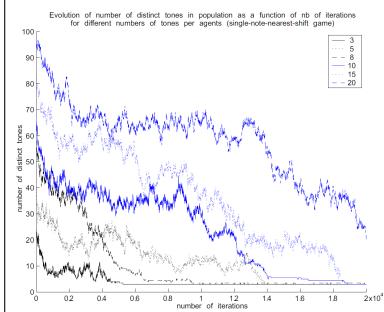
LES TRAVAUX DÉCRITS PRÉCÉDEMMENT sur la fouille d'information musicale visaient des comportements d'écoute musicale. La fin de mon doctorat (impressionné par le développement à la même époque du premier système “Continuator” de François Pachet²⁰) et mes premières expériences post-doctorales, d'abord au département de musique de la *School of Oriental and African Study* (SOAS) de Londres avec l'ethnomusicologue Keith Howard, puis avec le physicien Takashi Ikegami à l'Université de Tōkyō, m'ont ensuite amené à m'intéresser à la simulation des comportements de *production* sonore, à travers le prisme des systèmes dynamiques.

Un premier thème de recherche concerne la notion d'émergence et d'auto-organisation. En collaboration avec les musicologues de SOAS, nous avons proposé un nouveau mécanisme pour expliquer la diversité des systèmes tonaux rencontrés dans les cultures musicales humaines. En effet, bien qu'une grande partie de la musique occidentale soit basée sur le même système de 12 notes dit de tempérament égal (do, do#, ré, ré#....), de nombreux autres systèmes existent, qui varient selon leur nombre de classes de notes (jusqu'à 22 dans la musique indienne) et leur arrangement dans le continuum de fréquence (par exemple, l'accord sléndro de l'orchestre de Gamelan javanais, qui possède 5 notes par octave). A la manière des “jeux de langage” déjà utilisés en linguistique évolutionnelle, nous avons implémenté un système social artificiel, où des agents interagissent les uns avec les autres en échangeant quelques notes qu'ils essaient d'accorder en optimisant leur consonance. Nous avons montré que, par la seule dynamique des interactions locales, et sans nécessiter aucune optimisation mathématique centralisée, des systèmes tonaux peuvent émerger présentant les mêmes caractéristiques et la même diversité que les systèmes naturels (Encadré 3).

Un second thème de recherche concerne la notion de compromis entre autonomie et réactivité. Dans un premier système, conçu avec François Pachet, nous avons utilisé le formalisme informatique de la satisfaction de contraintes pour construire un agent musical capable de

20. Pachet, F. (2003). *The continuator : Musical interaction with style*. Journal of New Music Research, 32(3), 333-341.

Encadré 3 : Auto-organisation des systèmes tonaux

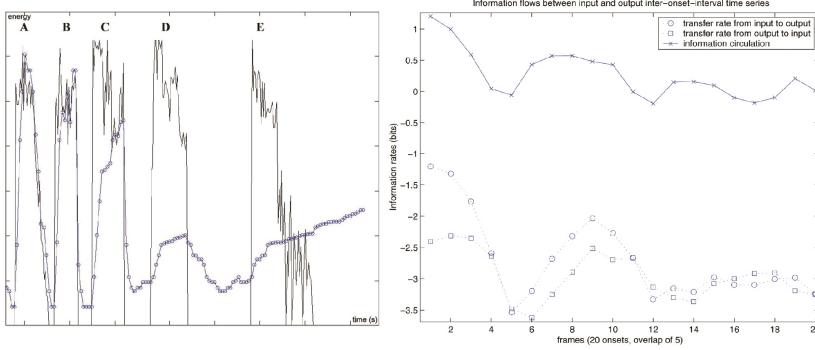


Quand des agents artificiels accordent leurs instruments de musique deux à deux de façon à optimiser leur consonance, la population, quelque soit sa taille, converge vers un système composé de 3 notes de musique : octave, tierce et quarte.

Aucouturier, J.J. (2008) *The hypothesis of self-organization for musical tuning systems*. Leonardo Music Journal, vol. 18, pp. 63—69.

concaténer de courts extraits musicaux de façon à interagir en temps-réel avec un musicien humain²¹. Chaque extrait successif est sélectionné pour satisfaire à des contraintes de continuité et de réactivité, qui sont pondérées dynamiquement de façon à contrôler le degré d'autonomie du système. Dans un second système, développé à l'Université de Tōkyō, nous avons utilisé une architecture de réseau de neurones à décharge pour contrôler les mouvements du robot musical Miuro²². En connectant le réseau en entrée à un train d'impulsions correspondant aux pulsations de la musique, la sortie du réseau adopte un comportement d'itinérance chaotique (*chaotic itinerancy*), qui présente une alternance spontanée entre des phases de fort et faible couplage entre la sortie (mouvement) et l'entrée (musique). Entre d'autres termes, ce système lui aussi oscille entre réactivité et autonomie. Takashi Ikegami a montré depuis que ces deux approches sont réconciliables dans le cadre mathématique d'un oscillateur forcé de Van der Pol (Encadré 4).

Encadré 4 : Autonomie et réactivité des interactions musicales

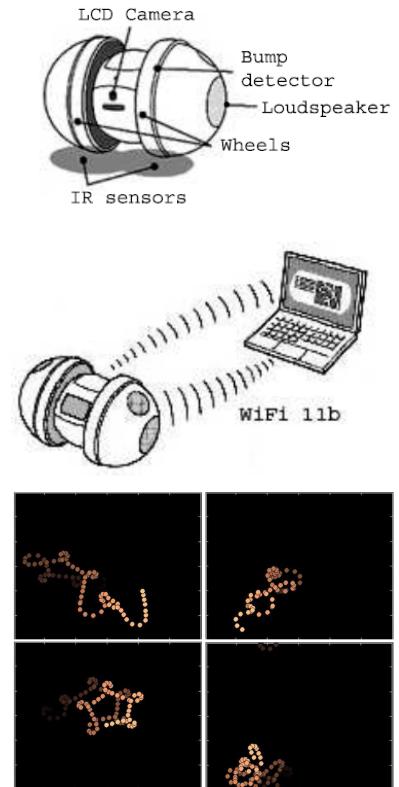


Deux mécanismes différents arrivent au même compromis dynamique entre autonomie et réactivité dans un système interactif. À gauche : un système de satisfaction de contraintes, qui cherche à tout instant à optimiser des contraintes de réaction à l'environnement, et des contraintes de continuité interne. Selon la pondération entre les 2 types de contraintes (A,B,C,D ou E), le système est entièrement réactif (A : le système - courbe bleue - suit l'environnement - courbe noire), soit entièrement autonome (E), soit dans un état intermédiaire. À droite : un robot contrôlé par un réseau de neurones à itinérance chaotique, qui oscille spontanément entre des phases de fort couplage à l'environnement (haute circulation d'information) et des phases de fort découplage (basse circulation d'information).

Aucouturier, J.-J. & Ikegami, T. (2009) *The Illusion of Agency : Two Engineering Approaches to Compromise Reactivity and Autonomy in an Artificial System*. *Adaptive Behavior*, vol.17(5).

21. Aucouturier, JJ. & Pachet, F. (2006) *Jamming with plunderphonics : Interactive contatenative synthesis of music*. *Journal of New Music Research*, vol. 35(1), pp. 35-50.

22. Aucouturier, JJ., Ogai, Y. & Ikegami, T. (2008) *Using Chaos to Trade Synchronization and Autonomy in a Dancing Robot*. *IEEE Intelligent Systems*, 23(2).



La plateforme robotique musicale Miuro (ZMP Inc.) utilisée pour les simulations d'itinérance chaotique, et quelques exemples de trajectoires obtenues en couplant le mouvement du robot au rythme de la musique.

Cette approche de simulation computationnelle des comportements comme des systèmes dynamiques non-linéaires (correspondant, entre autres, à la communauté scientifique d'*Artificial Life*) semble avoir beaucoup à offrir pour caractériser des mécanismes cognitifs. D'une part, là où la fouille de données musicales réduit typiquement le comportement humain à un jugement isolé par chanson (une annotation utilisée pour l'apprentissage), ce type de travaux amène au contraire à considérer des comportements riches, échantillonnes au cours du temps, avec l'intuition que c'est justement dans leur dynamique microscopique (instantané, locale, individuelle) que l'aspect macroscopique peut trouver une explication (long-terme, globale, collective). Ce type de mesure du comportement comme série temporelle (par exemple, physiologique ou acoustique) reste depuis ces premiers travaux une constante dans mon travail.

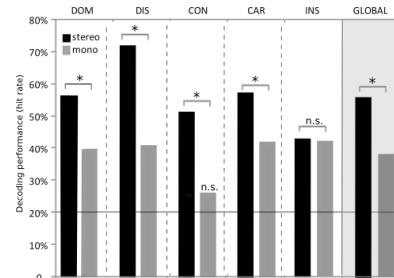
D'autre part, la notion de système émergent à partir de lois d'interaction locale est un formalisme particulièrement riche pour l'étude des comportements communicatifs comme la voix ou la musique. Il est à la base des réflexions actuelles dans la communauté, à un niveau comportemental, sur la cognition sociale des interactions²³, et à un niveau neuronal, sur les neurosciences sociales²⁴. Ces concepts sont par exemple récemment réapparus dans ma recherche avec une série de travaux expérimentaux sur la cognition sociale de duos musicaux improvisés (avec le musicologue Clément Canonne)²⁵. Nous avons montré que des duos d'improvisateurs peuvent communiquer des intentions sociales non-musicales (comme le fait d'être dominant, conciliant ou dédaigneux) via leur interaction musicale. Afin de montrer que la capacité de décoder ce type de comportements sociaux dans la musique ne repose pas seulement sur l'interprétation des signaux de l'un ou l'autre des musiciens (par exemple, jouer fort ou aigu), mais sur des indices présents au seul niveau de l'interaction (par exemple, jouer ensemble ou pas), nous avons enregistré chaque duo de musiciens dans 2 canaux audio séparés, et présenté ces stimuli à des participants dans trois conditions : soit uniquement le musicien 1, soit uniquement le musicien 2, soit les deux ensemble. Les intentions ne sont correctement reconnues que dans la dernière condition, signe de l'existence d'indices "émergents" au sens des systèmes dynamiques.

Toutefois, si ces concepts de l'Artificial Life ont une certaine force d'inspiration théorique, j'ai trouvé (au même titre que pour les algorithmes MIR) difficile de leur trouver une application pratique à la preuve de résultats dans le domaine des sciences cognitives. Cela tient, à mon avis, à une trop grande distance entre le type de données simulables par ces modèles et le type de comportements mesurables expérimentalement. Premièrement, l'espace dans lequel des simulations comme celles du mouvement du robot Miuro sont conduites est typi-

23. par exemple, le débat Frith-Di Paolo : Gallotti, M. & Frith, C. D. (2013). *Social cognition in the we-mode*. Trends in cognitive sciences, 17(4), 160-165 ; Di Paolo, E. A., De Jaegher, H. & Gallagher, S. (2013). *One step forward, two steps back—not the tango : comment on Gallotti and Frith*. Trends in cognitive sciences, 17(7), 303-304.

24. Dumas, G., Nadel, J., Soussignan, R., Martinerie, J. & Garnero, L. (2010). *Inter-brain synchronization during social interaction*. PloS one, 5(8), e12166

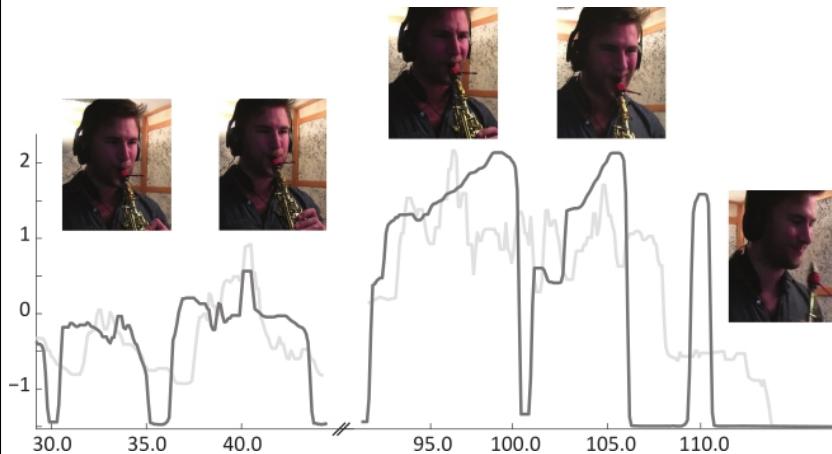
25. Aucourtier, J.J. & Canonne, C. (2017) *Musical friends and foes : the social cognition of affiliation and control in musical interactions*. Cognition, vol. 161, 94-108.



Performances de décodage (hit rates) de l'intention sociale exprimée dans 25 duos musicaux, présentés à des auditeurs musiciens, soit en condition stéréo (2 canaux audio), soit mono (1 seul canal audio), pour les 5 attitudes exprimées DOM : dominant, DIS : dédaigneux, CON : conciliant, CAR : prévenant, INS : insolent. (Aucourtier & Canonne, 2017)

quement une abstraction spatiale (des trajectoires de points dans un espace euclidien à 2 dimensions), et le résultat de ces simulations ne constitue donc pas un stimulus “écologiques”. Pour confronter certains facteurs de construction de ces modèles au jugement de participants humains (par exemple, voir si les phases de fort couplage entrée-sortie du réseau de neurones de l’Encadré 4 correspondent à un plus fort jugement d’agentivité par des observateurs), il faut soit demander aux participants de juger des trajectoires abstraites dans cet espace (ce qui présente certaines difficultés théoriques)²⁶, soit projeter ces trajectoires sur un artefact physique, par exemple une plateforme robotique, ce qui présente certaines difficultés pratiques : dans le cas de nos travaux avec le robot miuro, une fois adaptées aux constantes de temps du robot et à l’inertie physique du mouvement, les phases de couplage/découplages ne sont plus observables aussi précisément que dans la simulation²⁷.

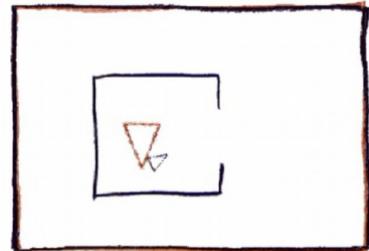
Encadré 5 : Coordination temporelle et interactions musicales



Séries temporelles superposées de l’énergie du signal sonore (RMS) de deux musiciens improvisant un duo dans lequel l’un (gris foncé : saxophone soprano, voir photos) a reçu la consigne de communiquer à l’autre (gris clair : piano) la notion de dominance (“c’est moi le chef”). L’analyse du duo comme un système dynamique couplé (avec la causalité de Granger) montre un fort flux d’information allant du musicien encodeur vers le musicien décodeur - un pattern que l’on retrouve dans une majorité des duos enregistrés ainsi ; le pattern inverse (du décodeur vers l’encodeur) est observé au contraire dans les duos avec la consigne d’être prévenant (“aux petits soins”).

Aucouturier, JJ. & Canonne, C. (2017) *Musical friends and foes : the social cognition of affiliation and control in musical interactions*. Cognition, vol. 161, 94-108.

26. Ceci ne serait pas sans rappeler certaines tâches d’empathie cognitive présentant à des enfants des animations vidéo abstraites, où plusieurs triangles bougent de façon à figurer des scènes sociales (par exemple une mère encourageant son enfant à sortir d’une pièce) et les enfants sont évalués sur leur capacité à attribuer des intentions à ces animations.



Castelli, F., Happé, F., Frith, U. & Frith, C. (2000). *Movement and mind : a functional imaging study of perception and interpretation of complex intentional movement patterns*. Neuroimage, 12(3), 314-325

27. Une démo vidéo pour s’en convaincre

<https://www.youtube.com/watch?v=CFP8ky6ssU8>

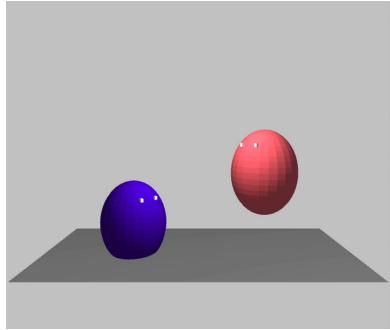


Deuxièmement, les comportements humains tels qu'on peut les mesurer en cognition sont difficiles à analyser de façon à les exprimer en paramètres de simulation dynamique (c'est-à-dire, d'en faire le "reverse engineering" sous forme de modèle computationnel). Au mieux, on peut caractériser ces comportements avec les mêmes mesures que l'on appliquerait à une simulation : dans les duos improvisés enregistrés par Aucouturier & Canonne (2017), on peut extraire les séries temporelles de l'énergie du signal sonore (RMS) de chaque musicien, et quantifier la force de leur couplage temporel avec les mêmes mesures de flux d'information ou de causalité de Granger que celles utilisées pour caractériser le système dynamique de Aucouturier & Ikegami (2009) (Encadré 5). Toutefois, ce diagnostic du comportement ne vaut pas modèle génératif : il existe une infinité de systèmes dynamiques couplés présentant les mêmes caractéristiques de transfert d'information, et il ne semble pas évident de "fitter" un modèle en particulier aux données comportementales, de la même façon qu'on "fitterait" une distribution statistique paramétrique à un échantillon de données²⁸

Il existe donc un fossé entre la flexibilité de simulation du système abstrait, et la concrétude des données qui sont collectables expérimentalement pour les problèmes qui m'intéressent. Une des pistes que nous explorons actuellement consiste à faire générer des comportements par l'humain directement dans un espace accessible à la simulation²⁹. Une autre piste consiste à construire des modèles génératifs capables de synthétiser des comportements directement interprétables par l'humain - c'est ce type de travail que je décris dans la section *Synthèse*.

28. On peut noter ici que les approches récentes de quantification de récurrence à base d'Oracles de Markov variables pourraient permettre de trouver un terrain intermédiaire entre les données et la spécification d'un modèle analytique, et ce sans avoir à reconstruire un espace de phase par *time-delay embedding*. Voir par exemple leur application à la reconnaissance d'émotions dans Mouawad, P. & Dubnov, S. (2016) *Symbolic Dynamics and Recurrence Quantification Analysis for Affect Recognition in Voice and Music*, International Conference on Perspectives in Nonlinear Dynamics

29. Par exemple permettre aux participants de manipuler le comportement de dyades de balles rebondissantes comme dans les simulations de Sievers et al (2013).



Sievers, B., Polansky, L., Casey, M. & Wheatley, T. (2013). *Music and movement share a dynamic structure that supports universal expressions of emotion*. Proceedings of the National Academy of Sciences, 110(1), 70-75.

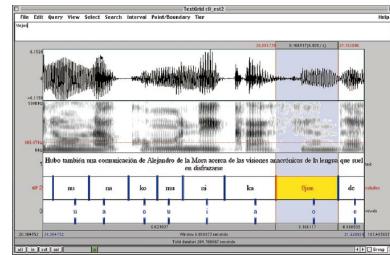
Analyse

MES PREMIERS TEMPS PASSÉS DANS DES LABORATOIRES purement expérimentaux (psychoacoustique chez Katsumi Watanabe, neurosciences chez Kazuo Okanoya au Japon, psychologie cognitive chez Emmanuel Bigand à l'Université de Bourgogne) m'ont permis d'apprendre de nouvelles méthodologies d'acquisition de données mais, paradoxalement, m'ont aussi forcé à trouver des utilisations beaucoup plus immédiates et pragmatiques de mon expérience de traiteur de signaux - il fallait bien contribuer, et mes moyens expérimentaux étaient encore maigres. L'une des utilisations les plus conventionnelles³⁰ du traitement du signal audio pour un travail de sciences cognitives consiste à analyser acoustiquement des corpus d'enregistrements pour en donner une caractérisation physique, complémentant ainsi leur caractérisation cognitive. De nombreux outils logiciels existent pour ce faire³¹, ce qui a l'intérêt de standardiser un peu les algorithmes et d'aider l'acceptation par la communauté expérimentale de telle ou telle feature (quand on parle de *jitter* et que l'on cite PRAAT, les *reviewers* savent de quoi l'on parle). Il existe 2 cadres classiques d'application de ces techniques dans la communauté de la cognition audio : soit les caractéristiques acoustiques servent de regresseurs pour "expliquer" des données comportementales - c'est une approche très utilisée dans la communauté des émotions vocales et musicales et dont je donnerai un exemple plus précis plus loin ; soit l'analyse automatique est utilisée pour découvrir des régularités dans de gros corpus de sons collectés en dehors de conditions expérimentales - ces régularités pouvant ensuite être soumise à une validation en laboratoire.

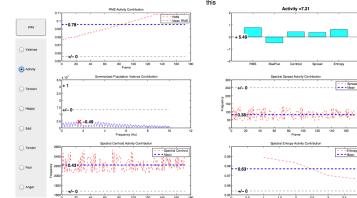
Un exemple de cette deuxième approche est une collaboration se poursuivant depuis 2010 avec Yulri Nonaka et Kazuo Okanoya sur l'analyse d'un corpus de cris de 30 bébés japonais, enregistrés chacun pendant 1 an dans leur environnement domestique, chaque cri étant annoté par la mère selon le contexte qu'elle pense lui être associé (faim, besoin d'être changé, besoin d'interaction, etc. au total environ 60,000 vocalisations contextualisées). La question scientifique sous-jacente est de savoir comment les enfants pré-linguistiques en-

30. qui, il y a 10 ans, n'était pas si courante que ça : la première fois que j'ai demandé à mes collaborateurs japonais de m'envoyer les données audio pour analyse, j'ai reçu le lendemain par Fedex une pile de spectrogrammes imprimés et surlignés à la main...

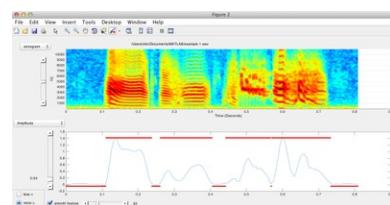
31. On peut citer Praat de P. Boersma pour les données de parole <http://www.praat.org/>,



MIR Toolbox d'Olivier Lartillot pour les données musicales <https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox>,

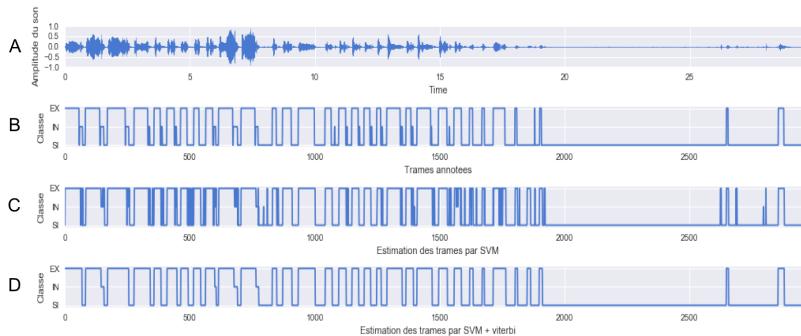


et Sound Analysis Pro de l'équipe d'Ofer Tchernichovski pour la communication animale <http://soundanalysispro.com>



codent un état émotionnel ou une intention sociale dans l'acoustique de leur cri, et en particulier comment ce code, qui peut-être considéré comme un précurseur développemental du langage articulé³², évolue au cours du temps et diffère d'individu en individu. Le travail acoustique sur ce type de corpus est un mélange d'exploration *brute force* avec des outils génériques et de développement d'algorithmes spécifiques. Avant de soumettre le corpus à l'analyse de données avec Praat (durée moyenne, *pitch*, *jitter*, *shimmer*, rapport harmonique/bruit, etc.) et de chercher des régularités en fonction de l'âge et du contexte, nous avons par exemple développé un algorithme supervisé permettant de segmenter automatiquement les cris en phases inspirées et expirées, afin d'analyser ces dernières uniquement (Encadré 6).

Encadré 6 : Segmentation automatique des phases expirées et inspirées de cris de bébés



Exemple de segmentation d'un enregistrement de 30 sec. (A), segmenté (B) à la main par un expert en biolinguistique avec Sound Analysis Pro en phases expirées (EX), inspirées (IN) et silence (SI). Afin de reproduire et généraliser ces annotations, on découpe le signal en courtes trames temporelles (50ms) et on extrait de chacune une paramétrisation du contenu spectral par coefficients mel-cepstraux (MFCCs). Une machine à vecteur support (SVM) apprend ensuite une fonction de discrimination entre ces trois types de phases. La classification trame à trame (C) est ensuite améliorée par un décodage de Viterbi (D) qui permet de prendre en compte les probabilités de transition typiques entre les 3 classes.

Aucouturier, JJ., Nonaka, Y., Katahira, K. & Okanoya, K. (2011) *Segmentation of expiratory and inspiratory sounds in baby cry audio recordings using hidden Markov models*, Journal of Acoustical Society of America, 130(5), pp. 2969-2977

Outre la nécessité de convaincre la communauté psychologique du bien-fondé des caractéristiques acoustiques que l'on utilise (là où la communauté MIR se satisferait de bons indicateurs de précision, il s'agit ici de montrer que les caractéristiques ont un sens cognitif - j'y

32. Le premier mot apparaît vers 18 mois - Nazzi, T. & Bertoni, J. (2003). *Before and after the vocabulary spurt : two modes of word acquisition ?*. Developmental Science, 6(2), 136-142.

Quelques autres résultats de cette période :

Nonaka, Y., Aucouturier, JJ., Katahira, K. & Okanoya, K. (2015) *Developmental diversity in infant cry through maternal interactions*, Tōkyō Lectures in Language Evolution, Tōkyō, Japan.

Fourer, D., Shochi, T., Rouas, J. L., Aucouturier, JJ., & Guerry, M. (2014). *Going ba-na-nas : Prosodic analysis of spoken Japanese attitudes*. In Speech Prosody, Dublin, Ireland, 2014.

Fourer, D., Guerry, M., Shochi, T., Rouas, J. L., Aucouturier, J. J., & Rilliard, A. (2014). *Analyse prosodique des affects sociaux dans l'interaction face à face en japonais*. In XXXèmes Journées d'études sur la parole, Reims, France.

Nonaka, Y., Aucouturier, JJ., Katahira, K. & Okanoya, K. (2013) *Developmental differentiation in human infant cry through dynamic interaction with caregivers*, 33rd International Ethological Conference (Behaviour'13), Newcastle Gateshead, UK.

Hegde, S., Aucouturier, J.-J., Ramamujam, B. & Bigand, E. (2012) *Variations in Emotional Experience During Phases of Elaboration of North Indian Raga Performance*, Proceedings of the 2012 International Conference on Music Perception and Cognition

reviendrai), ce type de travail implique également une plus grande sophistication statistique que dans la stricte communauté signal : il ne suffit pas d'exhiber une augmentation de telle ou telle caractéristique au cours du temps (par exemple, un allongement de la durée moyenne des cris avec l'âge, prédict par une augmentation de la capacité pulmonaire avec la croissance), mais il s'agit d'en tester la significativité statistique (c'est-à-dire, évaluer si les différences observées sur l'échantillon de données qu'on analyse correspondent à une différence réelle au niveau de la population "de tous les bébés possibles"). On pourra pour cela utiliser traditionnellement une ANOVA à mesures répétées (avec l'âge en facteur intra-sujet) ou, préférentiellement depuis quelques années, des modèles linéaires à effets mixtes (en représentant le bébé comme facteur aléatoire, et l'âge comme facteur ordonné). Tous ces aspects interprétatifs, en partie liés à l'épistémologie de la démarche (établir un phénomène, plutôt que construire un algorithme) mais aussi aux conventions de la communauté scientifique ciblée (utiliser Praat car 12,452 autres articles l'on fait avant moi³³) sont un bon exemple du savoir-faire, important mais rarement explicité, qu'il est nécessaire d'acquérir pour appliquer les technologies du son aux problèmes de sciences cognitives.

Avec le recul, deux sources d'incompréhension me semblent particulièrement importantes à éviter quand on applique le traitement du signal sonore à l'analyse de phénomènes de communication, dans le domaine de la linguistique, biolinguistique ou de l'expressivité musicale. Premièrement, les *features* audio que l'on utilise en reconnaissance de formes peuvent facilement passer, auprès de la communauté psychologique, pour des propositions de mécanismes cognitifs, alors qu'elle ne le sont pas. Il est en effet courant, dans la communauté computationnelle, de lire par exemple que les *features* MIR sont inspirées par les propriétés du système auditif humain. Un algorithme comme celui des Coefficients Mel-Cepstraux (MFCCs)³⁴ est souvent décrit comme reproduisant l'échelle tonotopique non-linéaire de la cochlée (car les fréquences sont converties en unité Mel) ainsi que la réponse dynamique des cellules ciliées de la membrane basilaire (car on prend un logarithme pour compresser l'amplitude des données). Cependant, cette description n'est que partiellement correcte. Si certaines propriétés de la physiologie humaine (assez datées, d'ailleurs³⁵) servent effectivement de point de départ, d'autres parties de l'algorithme lui ont été adjointes dans le seul but d'améliorer son utilité computationnelle dans un contexte de *machine learning*, et pas du tout sa pertinence cognitive. Par exemple, la transformée en cosinus discrète finale est utilisée pour réduire les corrélations entre coefficients et simplifier leur modélisation statistique³⁶. Cela est également vrai, par exemple, de features comme la *spectral skewness*, le troisième moment spectral, et

33. source : Google Scholar, 29/06/2017

34. Pour mémoire, les MFCCs sont dérivés ainsi :

1. Prendre la transformée de Fourier d'une trame du signal
2. Intégrer le spectre de puissance obtenu ci-dessus sur l'échelle Mel, avec des fenêtre fréquentielles en forme de triangles recouvrants
3. Prendre le log des puissances à chaque fréquence Mel
4. Prendre la transformée en cosinus discrète (DCT) de la liste des log des puissances Mel, comme si c'était un signal temporel
5. Les MFCCs sont les amplitudes du spectre résultant de la DCT

Zheng, F., Zhang, G., & Song, Z. (2001). *Comparison of different implementations of MFCC*. Journal of Computer Science and Technology, 16(6), 582-589.

35. L'échelle Mel remonte probablement à Koenig, W. (1949) *A new frequency scale for acoustic measurements*, Bell Telephone Laboratory Record, vol. 27, pp. 299-301.

36. Logan, B. (2000). *Mel Frequency Cepstral Coefficients for Music Modeling*. In Proc. ISMIR 2000, Plymouth, MA

la *kurtosis*, quatrième moment, qui ne doivent leur existence qu'au fait que le 2ème moment, le *spectral centroid* est cité en psychoacoustique et qu'il était mathématiquement logique³⁷ de continuer à dériver d'autres moments statistiques au delà de deux, et ce malgré leur probable absence de signification psychologique. Ce problème de l'absence de modularité cognitive des outils classiques du traitement du signal audio est d'autant plus saillant que de nombreuses techniques existent pour sélectionner et combiner automatiquement les features de façon à prédire n'importe quelle annotation. Une boîte à outil comme MIR-Toolbox permet de faire de la régression multivariée sur près de 300 combinaisons de caractéristiques ; d'autres techniques plus avancées parlent d'en explorer des millions³⁸. Cette tendance (plus de *features*, plus compliquées et plus combinées), qui culmine aujourd'hui dans les méthodes d'apprentissage profond, est sans aucun doute à créditer pour l'amélioration continue de la précision des algorithmes MIR sur les problèmes de classification et de régression audio, mais elle rend également de plus en plus difficile de dériver des intuitions psycho-logiques à partir de ce que la communauté présente comme de bons "prédicteurs". Je prends un exemple d'un corpus de jugements de *valence* et d'*arousal* émotionnel collectés sur des participants adultes pour un petit nombre d'extraits musicaux³⁹. Nous avons calculé une regression multiple automatique de ces données avec MIRToolbox dans l'espoir d'identifier quelles caractéristiques acoustiques des musiques peuvent expliquer les émotions qu'elles provoquent.

Regression for valence	
Feature	β
tonal_chromagram_peak_PeakMagPeriodEntropy	-0.75
tonal_mode_Mean	0.13
spectral_mfcc_PeriodAmp_8 (600 Hz +/- 66)	0.12
spectral_ddmfcc_PeriodEntropy_1 (133.3)	-0.11

Regression for arousal	
Feature	β
spectral_mfcc_Std_6 (466.6)	-0.34
spectral_mfcc_Mean_3 (266.6)	0.28
tonal_keyclarity_Std	-0.28
spectral_mfcc_Std_7 (533.3)	0.24

Les résultats montrent que la valence des musiques est très bien “expliquée” par *l’entropie de la période de l’amplitude du pic maximum détecté toutes les 50ms dans le chromagramme du signal*⁴⁰. L’excitation, par contre, est expliquée par *la variance du 6ème coefficient Mel-cepstral et la moyenne du 3ème* (mais attention, pas l’inverse!). Ces explications sont très clairement artificielles (l’effet sur un public d’informaticiens est presque humoristique), mais que se passerait-il si un psychologue cognitiviste essaie de les prendre à la lettre ? Il serait confronté à un mécanisme d’une complexité formidable, impliquant

37. Spectral centroid :

$$\mu = \sum f \tilde{S}(f) \quad (1)$$

Spectral spread :

$$\sigma^2 = \sum (f - \mu)^2 \tilde{S}(f) \quad (2)$$

Spectral skewness :

$$\gamma^3 = \sum (f - \mu)^3 \tilde{S}(f) / \sigma^3 \quad (3)$$

Spectral kurtosis :

$$\gamma^4 = \sum (f - \mu)^4 \tilde{S}(f) / \sigma^4 \quad (4)$$

où $\tilde{S}(f)$ est l'amplitude normalisée $S(f) / \sum_f S(f)$ du spectre S à la fréquence f

38. Pachet, F. & Roy, P. (2007). *Exploring billions of audio features*. In Proc. IEEE International Workshop on Content-Based Multimedia Indexing (CBMI); voir aussi Fröhlich, B., Rodner, E. & Denzler, J. (2012). *Semantic Segmentation with millions of Features : Integrating Multiple Cues in a Combined Random Forest Approach*. In Proc. 11th Asian conference on Computer Vision

39. Bigand, E., Delb  , C., Tillmann, B., G  rard, Y. (2011) *Categorisation of extremely brief auditory stimuli : Domain-specific or domain-general processes ?* PLoS ONE 6(10)

40. Le chromagramme repr  sente l'intensit   de chaque classe de pitch (do, do#, etc.) pr  sente dans le signal   un instant donn  e. Bartsch, M. A. & Wakefield, G. H. (2001). *To catch a chorus : Using chroma-based representations for audio thumbnailing*. in proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.

des processus aussi divers que l'apprentissage statistique (“l'entropie”), l'entrainement rythmique (“de la période”), l'intégration temporelle (“du pic maximum”), l'analyse harmonique (“du chromagramme”) et nécessitant même l'acculturation du sujet dans la culture occidentale (car le chromagramme est basé sur le système de 12 tons à tempérament égal de la musique classique). Ces algorithmes (des mécanismes computationnels) ont certes un fort pouvoir prédictif, mais leur intérêt en tant que mécanismes cognitifs est à peu près le même que celui d'un jet à réaction pour les biologistes essayant de comprendre la biomécanique d'un oiseau en vol : ce n'est pas parce que ça “fait la même chose” que ça fonctionne pareil.

Un deuxième problème lié à l'utilisation d'outils d'analyse acoustiques pour la preuve dans le domaine des sciences cognitives est que les résultats qui en découlent sont d'ordre descriptifs, et non causaux. Il existe un biais très fort dans la littérature expérimentale pour les études qui ne font pas qu'observer des corrélations (entre, ici, des comportements et leurs caractéristiques acoustiques) mais confirment que ces corrélations révèlent bien des mécanismes cognitifs. L'idée sous-jacente est simple (la corrélation n'implique pas la causalité), mais souvent obscurcie par des expressions conventionnelles qu'il m'a fallu des années pour décoder : une étude sera par exemple rejetée comme étant “trop descriptive” ou “pas assez mécanistique” (en anglais, *lack of mechanistic insights*), alors que ce que l'on attend en fait est simplement une forme de confirmation expérimentale où l'on manipule l'un des facteurs acoustiques révélés par l'analyse et observe un changement du comportement associé dans la direction prédictée⁴¹. J'en ai eu un exemple récent lors de la première soumission de notre travail sur les improvisations musicales⁴², où nous avions montré par analyse acoustique que des interactions affiliatives ou dominantes étaient caractérisées par plusieurs types de coordination entre les 2 canaux des musiciens, dont leur degré de consonance/dissonance et la causalité de Granger (cf. Encadré 5). Le manuscrit fut rejeté par l'un des lecteurs, avec une réponse qui pouvait sembler sans appel (“While it is a neat study with sophisticated acoustical analyses, the mechanisms behind the observed effects are not elucidated in psychological terms. For this reason I believe that there are grounds to question the work's theoretical relevance to the topic of cognition”), mais à laquelle, correctement interprétée, il fut relativement facile de répondre : il s'agissait d'établir un lien causal, et pas seulement corrélational, entre la dissonance et la synchronisation de la musique et le degré perçu d'affiliation et de dominance. Nous avons donc rapidement construit deux versions dégradées de nos stimuli, l'une en désaccordant un des canaux d'un demi-ton (transformant ainsi une interaction consonante en dissonante), l'autre en retardant l'un des canaux de 1 seconde par rapport

41. “For example, consider a cytokine expression study in which an increase in a specific inflammatory mediator is inferred to be important because its expression changes during infection. Such an inference cannot be made on correlation alone, since correlation does not necessarily imply a causal relationship. The study might be labeled ‘descriptive’ and assigned low priority. On the other hand, imagine the same study in which the investigators use the initial data to perform a specific experiment to establish that blocking the cytokine has a certain effect while increasing expression of the cytokine has the opposite effect. By manipulating the system, the investigators transform their study from merely descriptive to hypothesis driven. Hence, the problem is not that the study is descriptive per se but rather that there is a preference for studies that provide novel mechanistic insights”. Casadevall, A. & Fang, F. C. (2008). *Descriptive science*. Infection and immunity, 76(9), 3835-3836

42. Aucouturier, JJ. & Canonne, C. (2017) *Musical friends and foes : the social cognition of affiliation and control in musical interactions*. Cognition, vol. 161, 94-108.

à l'autre (transformant ainsi un musicien *leader* en *follower*), et fait un court test d'écoute pour évaluer l'impact des ces manipulations sur les jugements d'auditeurs. Le premier type de manipulation baissa l'affiliation perçue, et le deuxième, la dominance. “*With these new results*”, nous répondîmes, “*our study is no longer merely descriptive, but provides novel mechanical insights into how listeners process musical interactions*”. L'article fut accepté au 2ème tour de relecture. Cette logique expérimentale (utiliser l'analyse acoustique pour faire une hypothèse, puis manipuler le son et valider le mécanisme), en soi pas très éloignée de la démarche *training - testing* du *machine learning*, se heurte pourtant en pratique à plusieurs difficultés. D'une part, cette confirmation paraît souvent superflue au chercheur en informatique, car la plupart des outils utilisés en analyse acoustique sont déjà en soi prédictifs : la sélection de caractéristiques “à la MIRToolbox”, par exemple, fonctionne par validation croisée et semble donc correctement prédire des étiquettes de données dites “non observées”. Toutefois si, pour l'apprentissage automatique, toutes les données sont équivalentes et peuvent successivement être considérées pour l'apprentissage ou le test, la science expérimentale est plus soucieuse des conditions de test : des données collectées avec un plan d'expérience exploratoire (par exemple, un corpus que l'on soumet ensuite à l'analyse acoustique) ne doivent pas être utilisée ensuite pour confirmer un mécanisme - il convient de refaire une autre collecte, avec un design approprié (par exemple, un nombre égal d'exemples et de contre-exemples) pour répondre à cette deuxième question. D'autre part, le passage de l'analyse à la manipulation de caractéristiques audio dans des stimuli est souvent difficile, car la plupart des *features* audio ne sont pas des modèles génératifs. Trouver par exemple qu'un comportement est corrélé à une augmentation du centre de gravité spectrale ou, pire, du 2ème coefficient MFCC ne donne aucune indication sur la façon de modifier un signal pour augmenter spécifiquement ces caractéristiques, et pas les autres⁴³. C'est pour répondre à cette difficulté que mes travaux récents ont décidé de privilégier les outils de synthèse à ceux de l'analyse.

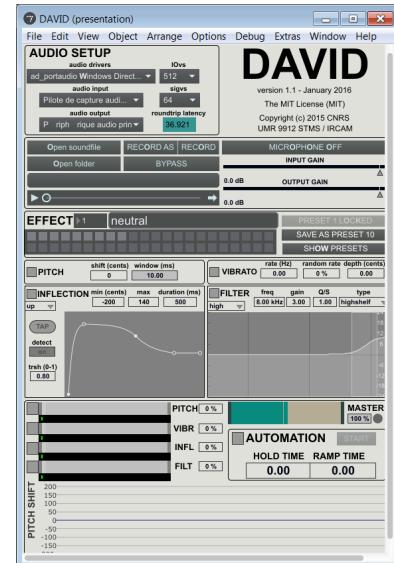
43. On peut procéder par interpolation, comme Terasawa, H., Slaney, M., Berger, J. (2005) *The thirteen colors of timbre*. In : Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, mais au prix d'une resynthèse qui dégrade rapidement la plausibilité des stimuli.

Synthèse

LA NÉCESSITÉ RÉCURRENTE QUI EST CONFRONTÉE aux techniques de reconnaissance (chap.I), de simulation (chap.II) et d'analyse du signal (chap. III) d'établir des "liens mécanistiques" entre les propriétés du son et les comportements humains qui en résultent m'a amené à développer depuis quelques années une quatrième approche qui me semble un cadre fructueux pour appliquer les technologies du son à ce type de travaux. Il s'agit d'utiliser la synthèse de son, et plus exactement l'algorithme de transformation de signaux sonores, comme technique de manipulation expérimentale des stimuli. Dans cette approche, l'informatique est utilisée pour étendre les possibilités de l'approche expérimentale classique. D'une part, avec un algorithme de synthèse ou de transformation sonore, nous sommes aujourd'hui⁴⁴ capables de manipuler finement les paramètres acoustiques d'un stimulus, là où l'approche classique se bornerait par exemple à sélectionner une voix d'acteur ou un extrait de Mozart correspondant plus ou moins bien aux hypothèses à tester. De plus, l'informatique, et en particulier les systèmes temps-réels, permettent de créer de nouvelles situations expérimentales, et donc potentiellement de mettre en évidence des phénomènes psychologiques nouveaux. Un exemple de cette approche (et celui qui m'a permis de la formaliser le mieux) est notre mise en évidence expérimentale du mécanisme psychologique de "rétroaction vocale émotionnelle". Nous avons créé (avec Petter Johansson et Lars Hall de l'Université de Lund en Suède, Katsumi Watanabe de l'Université Waseda au Japon et Marco Liuni et Laura Rachman de l'IRCAM) un algorithme de traitement du signal (appelé DAVID) permettant de modifier subrepticement le ton émotionnel de la voix d'un participant, en temps-réel, au fur et à mesure qu'il ou elle parle, pour la faire paraître plus joyeuse, triste ou anxieuse. Pour des auditeurs extérieurs, les transformations sont émotionnellement claires et naturelles, pourtant les locuteurs dont la voix est ainsi modifiée ne détectent pas la manipulation. De plus, dû au fait de s'écouter parler avec un ton de voix modifié, nous avons montré que l'état émotionnel des participants change dans la direction de l'émotion manipulée : en

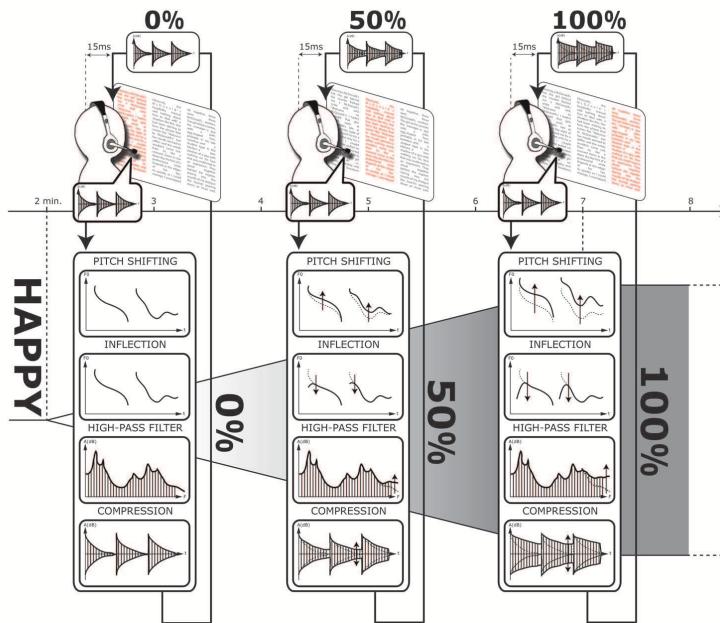
44. Cela est rendu possible par une maturité récente des algorithmes qui permettent un très haut niveau de qualité du rendu sonore, en naturel de la transformation (artefacts) comme en qualité sonore (dégradation). On peut donc prétendre utiliser des sons manipulés sans que les participants expérimentaux ne le remarquent ni adoptent une stratégie de traitement particulière (voir sur ce point Chap.5, p. 42-45)

Le logiciel DAVID ("*Da amazing voice inflection device*") permet d'appliquer des transformations de pitch paramétrique à un flux audio, comme une voix parlée, avec une latence inférieure à 20 millisecondes lui permettant de s'insérer dans la boucle sensorimotrice d'un locuteur sans altérer sa fluence verbale :



s'entendant parler de façon plus joyeuse, les participants deviennent plus joyeux eux-mêmes. Ce résultat a été publié dans la revue PNAS en Janvier 2016 (Encadré 7).

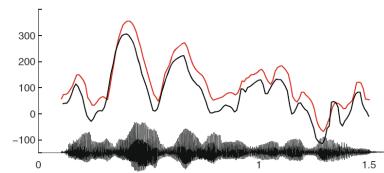
Encadré 7 : Transformation temps-réel de l'émotion d'une voix parlée



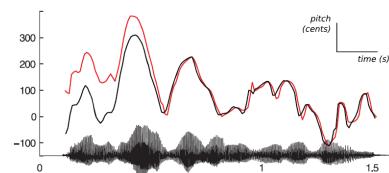
Dispositif expérimental de rétroaction vocale émotionnelle : les participants lisent un texte à voix haute, tout en s'entendant au casque. A leur insu, nous modifions avec des algorithmes de traitement du signal le retour sonore de leur propre voix pour la rendre progressivement plus expressive, par exemple ici plus positive : la hauteur de la voix est élevée de 50 cents, avec une inflection supplémentaire au début de chaque énonciation, pour la faire paraître plus dynamique ; les hautes fréquences sont amplifiées pour la faire paraître plus brillante. Ces changements, réalisés avec moins de 20ms de latence, ne sont détectés que par une minorité de locuteurs (moins de 15%), pourtant leur état émotionnel (mesuré par des questionnaires avant et après la lecture, et par une mesure d'activité électrodermique pendant la lecture) change en accord avec la manipulation : en s'entendant parler de façon plus joyeuse, les participants deviennent eux-mêmes plus joyeux.

Aucouturier, J.J., Johansson, P., Hall, L., Segnini, R., Mercadié, L. & Watanabe, K. (2016) *Covert Digital Manipulation of Vocal Emotion Alter Speakers' Emotional State in a Congruent Direction*, Proceedings of the National Academy of Science, 113(4).

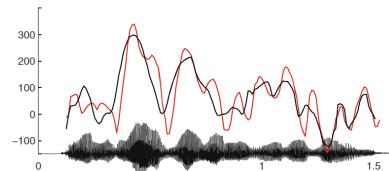
Trois des transformations de hauteur réalisées par DAVID : (1) une élévation ou diminution statique de la hauteur (*pitch shifting*), paramétrée en amplitude (par exemple, + 50 cents),



(2) une modification dynamique de la hauteur au début des utterances paramétrée avec des tronçons d'exponentielle et une durée (par exemple, 500ms), et déclenchée après un passage au silence ("inflection"),



(3) une modulation périodique de la hauteur, paramétrée en fréquence (par exemple, 8Hz) et en profondeur (par exemple, 50cents) ("vibrato").



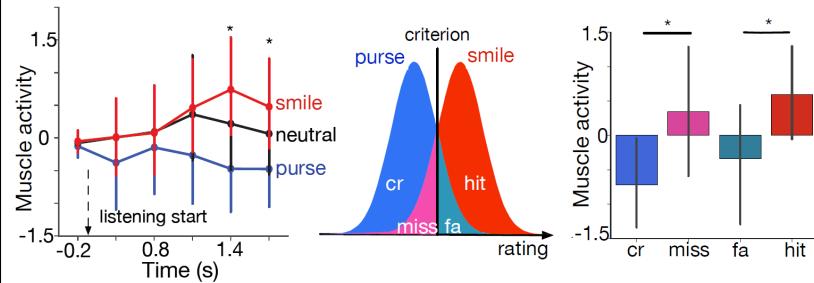
Cette approche reprend la vision développée au fil de mes travaux de *Reconnaissance sonore* (Chap.1), selon laquelle ces algorithmes sont moins utiles en tant que modèles cognitifs (modèles de réception de stimuli humains) qu'en tant que moyen de contrôle expérimental (modèles de production, que l'on soumet ensuite à la cognition humaine). Elle répond également aux limites identifiées de l'approche de *Simulation* (chap.2) de constituer des systèmes génératifs capables de synthétiser des comportements directement interprétables par les participants expérimentaux : les paramètres du modèle font par exemple bouger la hauteur de la voix ou la fréquence de son vibrato, qui sont des comportements sonores émotionnels directement réalisés par l'humain. Enfin, elle répond de façon immédiate aux reproches faits aux approches d'*Analyse* de ne pas tester de mécanismes causaux : avec des algorithmes comme ceux de DAVID, on manipule directement les paramètres du stimulus (par exemple, rendre sa propre voix plus joyeuse), afin de mettre en évidence un changement du comportement dans une direction prédictive (par exemple, devenir soi-même plus joyeux).

Comme toute démarche expérimentale, cette utilisation des outils de synthèse sonore doit passer, avant même de développer l'outil nécessaire, par la formulation d'une hypothèse liant certaines caractéristiques du son (que l'on va manipuler) à certains comportements (que l'on va mesurer). L'étude de la voix émotionnelle se prête particulièrement bien à ce cheminement intellectuel. Même si la plupart des approches existantes, justement basées plus sur l'analyse de corpus que de la synthèse, identifie un faisceau de paramètres acoustiques vocaux associé à chaque type d'expression⁴⁵, il est en fait relativement facile d'identifier des comportements vocaux émotionnels, biologiquement bien définis, qui méritent d'être modélisés et étudiés isolément. Un exemple est le phénomène du sourire vocal : le geste facial du sourire (l'étirement des coins de la bouche bilatéralement par les muscles zygomatiques majeurs) est un des éléments les mieux étudiés du répertoire expressif humain. Les sourires sont reconnus et imités dès les premiers mois de vie⁴⁶, ils sont utilisés comme signal visuel d'émotion positive dans toutes les cultures humaines⁴⁷ et sont traités par le cerveau de façon remarquablement robuste et automatique⁴⁸. Or, ce geste facial a également des conséquences acoustiques : en étirant les lèvres, le sourire change la forme du résonateur vocal et provoque des changements du timbre de la voix qui sont, par exemple, audibles au téléphone⁴⁹. Quand nous nous sommes intéressés à ce phénomène avec mon doctorant Pablo Arias, les mécanismes cognitifs du traitement de ces "sourires auditifs" étaient totalement inconnus. Nous avons modélisé le sourire vocal comme une transformation paramétrique agissant sur l'enveloppe spectrale de la parole en la déformant (avec un algo-

45. Par exemple, une voix joyeuse est dite associée à une vitesse rapide, une intensité moyenne à haute, une énergie haute fréquence moyenne, un *pitch* élevé et très variable, des contours de *pitch* croissants, des attaques rapides et "peu de régularités microstructurelles" - Juslin, P. N. & Laukka, P. (2003). *Communication of emotions in vocal expression and music performance : Different channels, same code ?* Psychological bulletin, 129(5), 770.
46. Oostenbroek, J., Suddendorf, T., Nielsen, M., Redshaw, J., Kennedy-Costantini, S., Davis, J. & Slaughter, V. (2016). *Comprehensive longitudinal study challenges the existence of neonatal imitation in humans*. Current Biology, 26(10), 1334-1338.
47. Ekman, P., Sorenson, E. & Friesen, M. (1969). *Pan-cultural elements of facial displays of emotions*. Science, 164(3875), 86-88
48. Dimberg, U., Thunberg, M. & Elmehed, K. (2000) *Unconscious facial reactions to emotional facial expressions*. Psychological science, 11(1) :86-89.
49. Basso, F. & Oullier, O. (2010). *'Smile down the phone' : Extending the effects of smiles to vocal social interactions* (Comment on target article). Behavioral and Brain Sciences, 33(6), 436

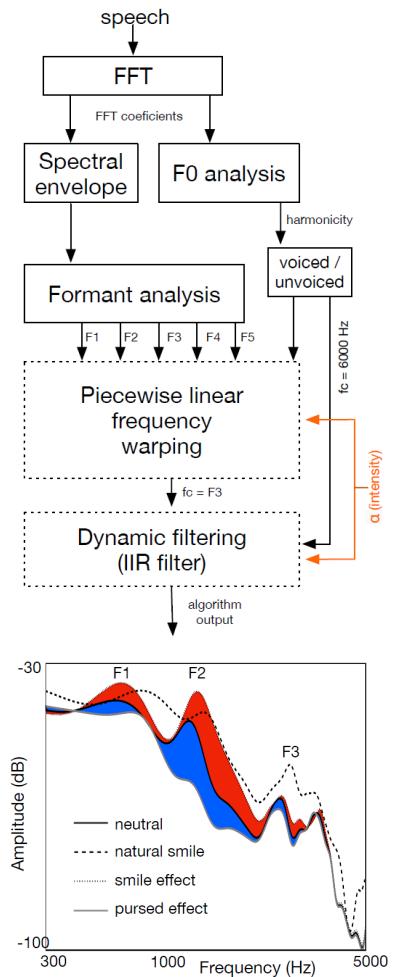
rithme de *frequency warping* intégré à une architecture de vocodeur de phase) de façon à décaler vers le haut la zone des trois premiers formants, et amplifier l'énergie des formants F2 à F5. . Grâce à ce modèle, qui permet de contrôler paramétriquement et de façon très réaliste la quantité de sourire perçue dans des enregistrements de parole, nous avons pu tester l'hypothèse, pour la première fois, que l'écoute de parole souriante (sans indice visuel) est suffisante pour déclencher chez l'auditeur une réaction d'imitation faciale (*facial mimicry*). Cette méthodologie a permis non seulement de montrer que c'est le cas, mais aussi que cette réaction est en partie inconsciente (Encadré 8).

Encadré 8 : Imitation faciale inconsciente du sourire perçu dans la voix parlée



Nous avons équipé des participants adultes d'électrodes EMG sur les muscles *Zygomaticus Major* (ci-dessus) et *Corrugator Supercilii* (non montrés ici), et leur avons demandé de noter le caractère souriant ou pas d'enregistrement de phrases manipulés algorithmiquement pour présenter, ou non, les caractéristiques acoustiques d'un sourire parlé. Les phrases souriantes ont généré plus d'activité musculaire zygomatique que les phrases non souriantes (gauche), démontrant un phénomène d'imitation faciale similaire à celui observé avec des stimuli visuels de visages souriants. Cependant, cette activité zygomatique chez l'auditeur est indépendante du jugement : elle existe pour les essais souriants et reconnus comme tels (*hits*), mais aussi pour les essais souriants non reconnus (*miss*) et pas pour les essais non-souriants identifiés à tort comme étant souriants (*false alarms*). Ces résultats montrent une surprenante capacité inconsciente des auditeurs à faire le *reverse engineering* de gestes articulatoires comme le sourire à partir de leur seule résultante acoustique.

Arias, P., Belin, P. & Aucouturier, J.J. (2017) *Auditory smiles trigger unconscious facial imitations*, en relecture.

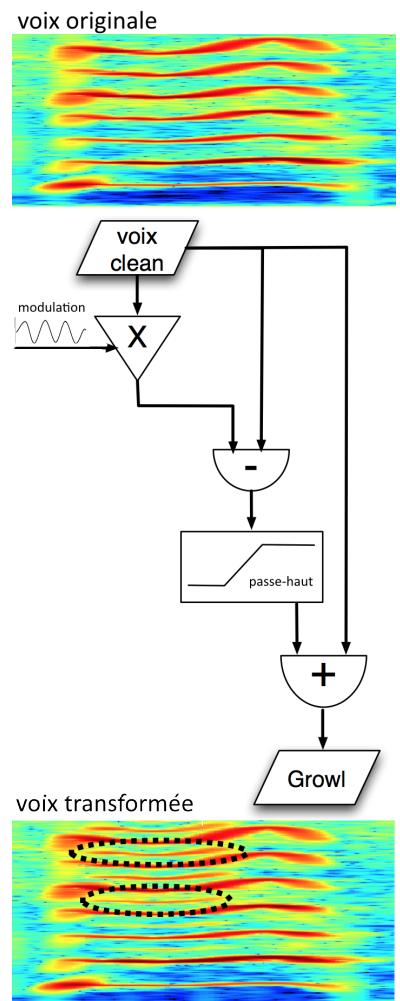


Arias, P., Soladié, C., Bouaffif, O., Roebel, A., Séguier, R. & Aucouturier, J.J. (2017) *Realistic transformation of facial and vocal smiles in real-time audiovisual streams*, en relecture

Un autre exemple de comportement vocal émotionnel que nous étudions avec cette méthodologie de transformation acoustique est la signalisation de l'excitation émotionnelle par la rugosité vocale. Plusieurs études récentes ont montré que la rugosité et l'inharmonicité vocale sont des indices acoustiques de danger et d'alarme qui sont traités de façon privilégiée à un niveau sous-cortical⁵⁰ et qu'ils sont présents non seulement dans les cris animaux et humains, mais aussi utilisés dans certains sons musicaux, comme les guitares saturées de la musique *metal*⁵¹. Cependant, il n'existe pas encore de modèle computationnel génératif capable de manipuler de façon paramétrique ces indices dans un signal sonore, ce qui limite les possibilités de comprendre expérimentalement comment ils influent sur la cognition et les émotions de l'auditeur. Nous avons donc développé (avec mes collègues Marco Liuni et Luc Ardaillon) un algorithme de transformation vocale capable de simuler les inharmonicités liées à la production vocale en régime rugueux, basé sur la production de sous-harmoniques par modulation de fréquence, et avons plusieurs expériences en cours l'utilisant pour tester la causalité de ce genre d'indices sur le comportement de participants. Parmi nos autres travaux, publiés ou en cours de publication, qui procèdent de la même vision, on peut également citer une technique de manipulation de l'impression de force physique dans la voix (basée sur la réduction de la dispersion des premiers formants), qui nous a permis de montrer que les médecins régulateurs du SAMU sont influencés dans leurs décisions par le son de la voix du patient au téléphone (Scientific Reports 2016, avec Laurent Boidron) ; et un algorithme de randomisation de la prosodie vocale, qui nous a permis de mettre en évidence des représentations mentales robustes de ce que doit être une intonation dominante ou digne de confiance, partagées entre hommes et femmes (en relecture 2017, avec Emmanuel Ponsot et Pascal Belin ; voir aussi Chap. 5, p. 40-42).

Outre sa puissance expérimentale, cette méthodologie de contrôle des stimuli sonores par transformation algorithmique présente également un fort potentiel de ré-emploi dans d'autres expériences. En effet, une fois qu'une technique de synthèse/transformation est développée et utilisée pour répondre à une hypothèse initiale (par exemple DAVID pour les travaux de l'Encadré 7, le sourire vocal pour ceux de l'Encadré 8), l'outil logiciel qui en résulte ouvre de nombreuses possibilités expérimentales supplémentaires, au même titre que les boîtes à outils d'analyse du type Praat ou MIRToolbox ne servent pas qu'au premier set de données qu'elles permettent de traiter. Tout d'abord, le fait que de tels outils permettent de synthétiser des comportements sonores de façon à la fois paramétrique et réaliste permet de s'en servir dans des paradigmes psychoacous-

Algorithme de transformation de la rugosité d'une voix, par mixage dans les hautes fréquences d'une modulation avec une porteuse basse-fréquence générant des sous-harmoniques (cercles pointillés).



50. Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L. & Poeppel, D. (2015). *Human screams occupy a privileged niche in the communication soundscape*. Current Biology, 25(15), 2051-2056.

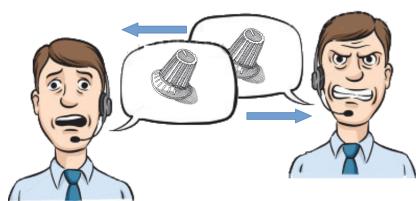
51. Blumstein, D. T., Bryant, G. A. & Kaye, P. (2012). *The sound of arousal in music is context-dependent*. Biology letters, rsbl20120374.

tiques. Les études classiques de la prosodie vocale, par exemple, utilisent généralement des vocalisations enregistrées par des acteurs pour imiter telle ou telle émotion, et cette méthodologie souffre de problèmes de typicalité (exprime-t-on toujours la joie au quotidien comme sur une scène de théâtre ?) et de co-variation de paramètres (un acteur, si bon soit-il, peut-il incarner seulement l'aspect tonal d'une expression triste, sans faire varier le timbre de sa voix ?).⁵²



Avec des outils comme DAVID, un expérimentateur peut échantillonner un large espace de variations prosodiques (par exemple toutes les fréquences de vibrato de 1Hz à 10Hz), ou manipuler sélectivement l'amplitude de certains paramètres (par exemple la hauteur indépendamment de l'intensité), une possibilité expérimentale qui, si elle est aujourd'hui courante dans le domaine

visuel⁵³, n'existe pas encore pour les sciences cognitives du son. Ceci peut être utilisé pour générer des variations prosodiques plus naturelles pour des techniques de *reverse correlation*⁵⁴ (voir Chap.5, p. 40-42), pour manipuler la difficulté de décodage des stimuli de façon continue dans le cadre de procédures psychophysiques itératives, ou pour produire des stimuli interculturels qui utilisent exactement les mêmes indices acoustiques, plutôt que de les faire produire par des locuteurs natifs de chaque culture.

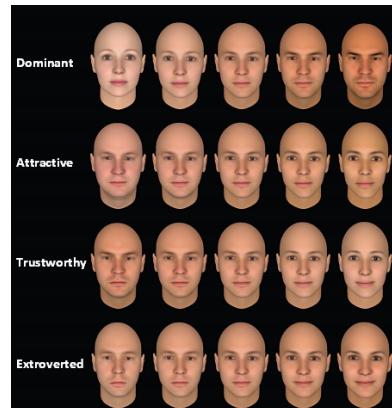


D'autre part, le fait que ces transformations sont réalisables en temps-réel ouvre la possibilité de manipuler l'expression émotionnelle en direct, au cours d'interactions sociales. La façon habituelle en psychologie sociale de contrôler cette expression lors de comportements de groupe

consiste soit à en donner l'instruction explicite aux participants⁵⁵, soit à les amener à exprimer spontanément cette émotion grâce à une *cover story* plus ou moins alambiquée⁵⁶. Avec des outils comme DAVID, on peut étudier des relations causales dans des interactions sociales en laissant les participants interagir librement (par exemple, au téléphone) et en modifiant l'expression de leur voix dans des directions congruentes ou incongruentes, et ce sans effet de demande expérimentale. Cette procédure pourra être utilisée, par exemple, pour étudier les stéréotypes émotionnels, la perception de la volonté de coopérer ou l'impact des processus émotionnels sur la productivité de groupe.

52. Jürgens, R., Grass, A., Drolet, M. & Fischer, J. (2015). *Effect of acting experience on emotion expression and recognition in voice : Non-actors provide better stimuli than expected.* Journal of Nonverbal Behavior, 39(3), 195–214

53. On peut citer par exemple les plateformes de synthèse paramétriques de visages expressifs utilisées par les groupes des psychologues Alexander Todorov à Princeton et Philippe Schyns à Glasgow - Yu, H., Garrod, O & Schyns, P. G. (2012) *Perception-driven facial expression synthesis*, Computers & Graphics, vol. 36, no. 3, pp. 152–162



54. comme, dans le domaine visuel, par Mangini, M. C. & Biederman, I. (2004). *Making the ineffable explicit : Estimating the information employed for face classifications.* Cognitive Science, 28(2), 209–226.

55. Tice, D. M. (1992). *Self-concept change and self-presentation : The looking glass self is also a magnifying glass.* Journal of Personality and Social Psychology, 63(3), 435.

56. Van Doorn, E. A., Heerdink, M. W., & Van Kleef, G. A. (2012). *Emotion and the construal of social situations : Inferences of cooperation versus competition from expressions of anger, happiness, and disappointment.* Cognition & Emotion, 26(3), 442–461.

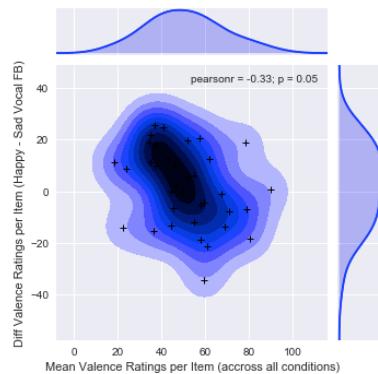


Enfin, le fait que certaines transformations soient si rapides (<20ms pour l'outil DAVID) permet d'envisager des paradigmes expérimentaux comme la rétroaction vocale émotionnelle qui étaient jusqu'à présent impossibles à concevoir. La recherche sur la régulation émotionnelle implique souvent de demander à des participants de se remémorer des événements ou de décrire des situations personnelles en utilisant un langage ou des gestes expressifs⁵⁷. Dans de tels paradigmes, il est difficile de séparer l'effet du ré-engagement émotionnel de celui de la production de telles expressions. Avec des outils comme DAVID, l'expérimentateur peut demander au participant de se remémorer un événement, tout en manipulant à son insu le ton émotionnel de sa voix, afin de tester par exemple si l'impact de souvenirs négatifs est atténué par un ton de voix positif. De façon similaire, dans la lignée de l'hypothèse des marqueurs somatiques de Damasio⁵⁸, certains travaux sur la prise de décision ont montré que le fait de donner à des participants de fausses informations à propos de leur rythme cardiaque ou leur ressenti émotionnel peut changer leurs jugements sur des vignettes morales ou la confiance qu'ils ont dans ces jugements⁵⁹. Des outils de transformation vocale temps-réel tels que ceux que nous développons peuvent donc être utilisés pour tester, sans effet de demande, si la voix fonctionne elle aussi comme un marqueur somatique pour la prise de décision.

Si un stimulus est lu avec une voix triste, et un autre avec une voix joyeuse, la prédiction est que les participants seraient orientés plus positivement vis à vis du second. Nos travaux actuels (avec Louise Goupil, ci-contre) tendent à montrer que c'est le cas : dans une réPLICATION *within-subject* du paradigme de rétroaction vocale de Aucouturier et al.(2016), nous avons demandé à des participants de décrire à voix haute quel serait leur ressenti s'ils faisaient l'expérience d'un certain nombre de situations émotionnelles (la perte d'un être cher, recevoir une promotion, etc.), et nous avons manipulé leur voix de façon triste ou joyeuse pendant qu'ils s'exprimaient. La valence de la manipulation vocale a alors influencé les jugements émotionnels qu'ils ont portés sur les situations imaginées, d'une façon qui suggère un traitement conjoint de la voix et du texte. Cette approche d'induction émotionnelle par sa propre voix peut être particulièrement intéressante dans des situations où le fonctionnement cognitif auto-référentiel est perturbé, par exemple dans la dépression ou les symptômes de stress post-traumatique.⁶⁰ Plus que de moyen de contrôle expérimental pour la recherche fondamentale, les technologies de transformation de la voix se verraient alors servir d'intervention dans un cadre cli-

57. Slatcher, R. B. & Pennebaker, J. W. (2006). *How do i love thee ? let me count the words the social effects of expressive writing*. Psychological Science, 17(8), 660–664
58. Damasio, A. R., Everitt, B. J. & Bishop, D. (1996). *The somatic marker hypothesis and the possible functions of the prefrontal cortex*. Philosophical transactions : Biological sciences, 348(1313-1420).
59. Shahidi, S. & Baluch, B. (1991). *False heart-rate feedback, social anxiety and self-attribution of embarrassment*. Psychological Reports, 69(3), 1024–1026; Arminjon, M., Preissmann, D., Chmetz, F., Duraku, A., Ansermet,F. & Magistretti, P. J. (2015). *Embodied memory : unconscious smiling modulates emotional evaluation of episodic memories*. Frontiers in Psychology, 6, 650

Influence d'une manipulation vocale joyeuse ou triste sur l'évaluation émotionnelle de vignettes lues à voix haute : un feedback vocal positif amoindrit le plaisir des situations positives, mais diminue aussi le déplaisir des situations tristes (Goupil et al., en prép.).



60. Grimm, S., Ernst, J., Boesiger, P., Schuepbach, D., Hell, D., Boeker, H., & Northoff, G. (2009). *Increased self-focus in major depressive disorder is related to neural abnormalities in subcortical-cortical midline structures*. Human Brain Mapping, 30(8), 2617–2627

nique. Illustration récente de ce potentiel, nous venons d'obtenir un financement ANR 2017 sur l'application du feedback vocal pour traiter l'alexithymie dans les symptômes post-traumatiques.⁶¹

Mon ambition aujourd'hui est de continuer d'avancer simultanément sur ces deux plans (développement d'outils logiciels pour la communauté d'une part, et utilisation de ces logiciels pour nos propres travaux de cognition), et de positionner mon équipe de recherche comme pionnière d'une méthodologie qui, je l'espère, contribuera à changer durablement la façon dont on étudie comment le cerveau traite la musique et le son. Afin de favoriser ces applications, nos travaux récents se doublent d'une stratégie de "science ouverte", par laquelle nos logiciels sont mis à la disposition de la communauté scientifique sous forme gratuite et open-source⁶², sans s'interdire d'autre part d'explorer leur éventuelle valorisation industrielle hors du cadre académique (par exemple, avec le dépôt en Oct 2016 d'un brevet CNRS sur le sourire vocal visant l'industrie des centres de relation-client par téléphone et la robotique de loisirs).

61. ANR REFLETS ("Rétroaction Faciale et Linguistique et Stress Traumatique", Oct. 2017-Oct.2021), avec CentraleSupélec (coord.), IRCAM, Service de Santé des Armées / Hôpital militaire Percy, Chanel, Dynamixyz, HumanEvo.

62. Premier de cette série, le logiciel DAVID a été mis en ligne en Mars 2017 sur la plateforme Forumnet de l'IRCAM (ci-dessous), où il bénéficie d'un suivi de versions et d'un forum utilisateur



[Download](#) [Documentation](#)

LATEST NEWS

[Windows release 2.18.2 of SuperVP for Max](#)
April 21, 2017

[New release 2.18.2 of SuperVP for Max](#)
December 23, 2016

[Modals 3.4.3](#)
September 6, 2016

LATEST DISCUSSION GROUP POSTS // DAVID

[FAQ] Optimal hardware set-up for DAVID
acouturier

Welcome to the DAVID user-group !

acouturier

RELATED TRIBUTES



[Report of the Forum Ircam Spring 2017](#)
palumbo

DAVID, a real-time emotional voice transformation tool

"Like an auto-tune, but for emotions" (Brian Resnick, for Vox.com)

DAVID (*Da Amazing Voice Inflection Device*) is a free, real-time voice transformation tool able to "colour" any voice with an emotion that wasn't intended by its speaker. DAVID was especially designed with the affective psychology and neuroscience community in mind, and aims to provide researchers with new ways to produce and control affective stimuli, both for offline listening and for real-time paradigms. For instance, we have used it to create real-time emotional vocal feedback in Aucouturier et al. 2016.

Technically, DAVID is implemented as an open-source patch for the close-source audio processing platform Max (Cycling74), and, like Max, is available for most Windows and Mac OS configurations. DAVID is available for free to the Forum community after registration (follow download link to proceed).

What does it do ?

DAVID is a software tool able to "add" emotion to a speech recording, i.e. it can make that American chap

or that French young lady

sound anxious,

Interrogations

TOUT AU LONG DE CE PARCOURS allant des algorithmes de reconnaissance de formes, de simulation dynamique, d'analyse à ceux de la synthèse et transformation du son, c'est la question de comment ces technologies peuvent être utilisées pour "faire preuve scientifique" qui est posée. Nous pouvons aujourd'hui considérer que nous avons mis en place un méthodologie (la manipulation expérimentale de comportements sonores émotionnels à l'aide de modèles génératifs - cf chap. Synthèse) qui semble vouée à une certaine reconnaissance dans la communauté cognitive. Un indicateur de cette reconnaissance est l'adoption en cours de certains de nos outils, comme le logiciel DAVID, dans les laboratoires de chercheurs importants du domaine, comme ceux de Robert Zatorre à Montréal, Elvira Brattico à Aarhus, Marcel Zentner à Innsbruck, Petter Johansson à Lund et Katsumi Watanabe et Kazuo Okanoya à l'Université de Tōkyō. Toutefois, de nombreuses questions se posent encore sur la direction à donner à ces travaux. Certaines de ces questions sont techniques, concernant par exemple la modélisation du temps en informatique ou la perception du caractère naturel ou non d'une transformation. D'autres sont d'ordre "méta", par exemple sur le besoin de dériver une question théorique commune à tous les mécanismes cognitifs traités par ces outils, ou l'à-propos du choix interdisciplinaire pour une carrière académique. J'en liste ici, en guise de conclusion, un certain nombre, que je pose au lecteur autant qu'à moi-même - façon de prendre rendez-vous avec l'avenir : où cette interdisciplinarité en sera-t'elle dans 10 ans ?

Q1 : Êtes-vous informaticien, ou psychologue ?

La question est classique, voire un peu énervante, et tout travail interdisciplinaire y est irrémédiablement confronté un jour⁶³. Ironiquement, si mes premiers travaux étaient rejettés de la communauté psychologique comme ceux d'un informaticien ("vous devriez soumettre cet article à une revue d'ingénierie"), c'est aujourd'hui la question inverse qui m'est souvent (op)posée. Elle est plus profonde : à quel point dans des travaux comme la paire {outil DAVID / expérience de rétro-

63. Elle est quelque peu structurelle, aussi. Après avoir été recruté en 2012 au CNRS par la Commission Interdisciplinaire CID44 "sciences cognitives", j'avais été frappé que la première question qui m'a été posée était celle de mon rattachement principal soit à la section 7 (informatique), soit à la section 26 (cognition).

action vocale} ou {outil smile / expérience d'imitation faciale} fait-on vraiment oeuvre de recherche en informatique, ou bien ne se borne-t'on qu'à mettre les technologies informatiques existentes au service de l'expérimentation ? Comment avec des travaux tels que ceux décrits dans le chap. Synthèse de ce mémoire peut-on garantir que l'on fait avancer les deux disciplines autant l'une que l'autre ?

Notre article Cognition (2017), publié grâce à la manipulation de 5 chansons détunées et décalées dans le temps avec le logiciel Audacity, est un cas extrême, mais symptomatique : la communauté cognitive n'exige pas en soi de dépasser l'état de l'art informatique. Le cas du logiciel DAVID est également intéressant : la transformation expressive de la voix en temps-réel est un sujet d'étude tout à fait contemporain dans la communauté des interfaces hommes-machines et de l'informatique affective, et l'IRCAM possède dans ce domaine une expertise sur des paradigmes comme celui du vocodeur de phase⁶⁴ qui permettrait *a priori* de traiter cette question en dépassant l'état de l'art. Cependant, la plupart de ces méthodes sont basées sur une analyse spectrale qui nécessite par construction l'acquisition d'un buffer audio d'une durée permettant une résolution fréquentielle suffisante (en pratique, au moins 50ms). Or, la boucle sensorimotrice de la parole est sévèrement perturbée pour des délais de plus de 20 ou 30ms⁶⁵. Les besoins de l'expérience, et les propriétés du système biologique que nous voulions étudier, nous ont donc contraint à développer un logiciel de manipulation de hauteur dans le domaine temporel (à base de lignes de retard multiples⁶⁶, une technologie des années 1970...) qui, s'il s'est avéré parfaitement adapté au besoin expérimental, est très en deça de l'état de l'art du point de vue de la qualité sonore. DAVID a été publié comme une nouvelle technologie expérimentale dans la revue *Behavior Research Methods*, mais il aurait été difficile d'en tirer un article dans une revue informatique du type IEEE.

Un meilleur cas de figure est notre algorithme de simulation algorithmique du sourire audio développé pour traiter pour la question de l'imitation faciale (cf Encadré 8). Ce problème correspond une question expérimentale pour laquelle la solution n'existeit encore au niveau technologique. Le développement de cet algorithme a nécessité d'étendre le paradigme de vocodeur de phase et de *warping* fréquentiel pour en adapter les paramètres en temps-réel aux caractéristiques formantiques de la voix. Ce travail, fait en collaboration avec Axel Roebel de l'équipe Analyse/Synthèse de l'IRCAM, a pu faire l'objet, en soi et sans le "soutien" de l'expérience cognitive qui lui est associée, d'un article soumis dans la revue IEEE Affective Computing et d'un brevet d'invention. De la même façon, nos premières expériences sur les duo improvisés dominants ou prévenants (cf Encadré 5) posent la question de concevoir un système interactif temps-réel capable d'ajus-

64. Liuni, M. & Roebel, A. (2013). *Phase vocoder and beyond*. Musica, Tecnologia, 7, 73-120.

65. Stuart, A., Kalinowski, J., Rastatter, M. P., & Lynch, K. (2002). *Effect of delayed auditory feedback on normal speakers at two speech rates*. The Journal of the Acoustical Society of America, 111(5), 2237-2241.

66. Bode, H. (1984). *History of electronic sound modification*. Journal of the Audio Engineering Society, 32(10), 730-739

ter son comportement sur un utilisateur humain de façon à toujours le précéder et lui imposer une direction - quelque chose que l'homme du métier ne sait pas aujourd'hui réaliser, et qui constituerait en soi une contribution d'ordre informatique.

Dans ces cas les plus favorables, la causalité entre recherche informatique et expérimentation cognitive devient difficile à déterminer : la question cognitive provoque un dépassement de l'état de l'art computationnel, et en retour, le nouvel objet informatique permet de poser des questions expérimentales inédites, et de les poser mieux⁶⁷. Nous donnons dans les deux paragraphes suivants deux exemples supplémentaires de questions qui émergent directement de la communauté informatique et que notre expérience expérimentale permet d'adresser.

Q2 : Comment modéliser le temps ?

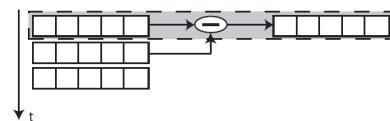
Une question informatique qui est posée en filigrane par la plupart des travaux décrits dans ce mémoire est celle de la modélisation du temps. La voix et la musique n'ont un effet cognitif sur l'auditeur qu'en ce qu'elles se déplacent et sont perçues "dans le temps". Pourtant, le moins qu'on puisse affirmer est que les sciences du son, et celles abordées dans mes travaux en particulier (analyse/synthèse et reconnaissance de formes), n'ont pas toujours traité cet aspect de façon satisfaisante.

En reconnaissance (Chap.1), la chaîne classique d'analyse extrait des caractéristiques acoustiques sur des fenêtres temporelles courtes (typiquement 50ms) et en modélise la distribution de façon statistique. La dimension temporelle peut y être traitée à deux niveaux : soit elle est intégrée aux *features* en construisant des vecteurs qui agglomèrent d'une façon ou d'une autre les caractéristiques de plusieurs fenêtres consécutives (par exemple, avec des *delta-coefficients*), soit on délègue au modèle le soucis de capturer des régularités statistiques à partir de la séquence de *features*, par exemple avec des modèles de Markov cachés qui mesurent des probabilités de transition d'un type de trame à un autre. De la même façon, en analyse acoustique (Chap.3), ce ne sont que des statistiques temporelles qui sont utilisées : on conclut sur la moyenne ou la variance du pitch (par exemple, plus haut en moyenne et plus varié quand la parole est joyeuse) mais pas du tout sur les formes temporelles que prennent ces caractéristiques (si le pitch est plus varié, est-ce à cause d'un vibrato constant, ou d'une brusque inflection de temps en temps?). Ces méthodologies permettent de prendre en compte des relations court-terme, trame à trame, dans la mesure où elles restent avérées statistiquement sur toute la durée du signal (par exemple, une augmentation continue). Or, si l'on sait depuis peu que la perception auditive opère en effet dans certains cas

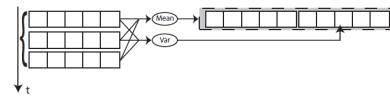
67. "Perhaps the question most often asked of computational modelers by empirical researchers is, 'What's the point?' My reply is always this : Good computational models inform empirical research, and good empirical research informs computational modeling". Robert M. French (2016), Reflections on Connectionist Modeling.

Trois façons d'intégrer la dimension temporelle dans les vecteurs de caractéristiques utilisés en reconnaissance de formes :

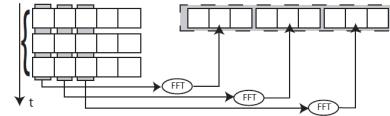
(1) *delta coefficients* : on adjoint aux caractéristiques instantanées leur dérivée temporelle à l'ordre 1 ou 2,



(2) *texture windows* : on substitute aux caractéristiques leurs statistiques sur des fenêtres plus longues,



(3) *dynamic features* : on substitute aux caractéristiques leur transformée de Fourier sur une fenêtre plus longue.



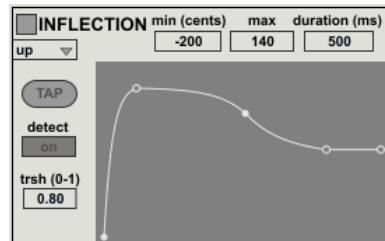
sur de tels résumés statistiques du signal⁶⁸, ce mode de perception ne suffit pas pour des signaux structurés comme la voix et la musique. Ce paradigme a par exemple beaucoup de mal à prendre en compte des événements isolés qui apparaissent dans le temps avec une certaine régularité, comme une augmentation de hauteur à la fin d'une phrase (pour en marquer le caractère interrogatif) ou une syncope à la fin de chaque mesure.

Parce qu'elles naissent du même creuset méthodologique que l'analyse, les méthodes de Synthèse (chap. 4) ont le même inconfort à incorporer de la structure temporelle aux transformations qu'elles opèrent sur le signal. Par exemple, l'outil DAVID applique des transformations fixes, en continu : la hauteur de la parole est augmentée de 50 cents, et le vibrato est de 8Hz, que le flux d'entrée en soit à une consonne, une voyelle, un soupir ou un silence. Cela résulte en partie de contraintes de temps-réel, qui oblige à prédire "à l'avance"⁶⁹ tout changement de régime. Par exemple, pour réaliser de courtes inflections de hauteur au début des utterances, DAVID détecte les passages au silence avec un seuil de niveau, puis déclenche l'application, dès que l'on dépasse ce seuil, d'une transformation de hauteur de forme fixe (par exemple, une fonction exponentielle décroissante tronquée après 500ms). D'une façon similaire, mais plus sophistiquée, l'algorithme de transformation de voix souriante (Arias et al., 2017) adapte les points de coupure de la déformation fréquentielle aux fréquences des formants de la voix, calculées à chaque trame : quand le premier formant augmente, la déformation de l'enveloppe suit proportionnellement, avec un ratio fixé à l'avance. Un désavantage majeur de ce type d'algorithmes est qu'ils sont évidemment complètement heuristiques et nécessitent de découvrir d'abord quelles sont ces régularités temporelles que l'on doit paramétriser - ce qui, dans la plupart des cas, est loin d'être trivial.

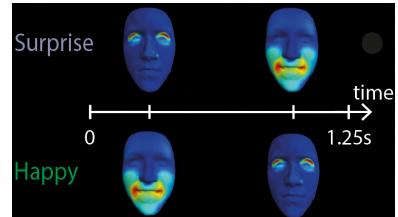
Une façon d'interroger comment la cognition du son déploie son attention au court du temps, et à quelle forme temporelle elle est réactive, est d'utiliser le paradigme psychophysique de la corrélation inverse (*reverse correlation*). Cette méthode, introduite pour caractériser les systèmes physiologiques (comme les champs récepteurs des neurones), présente aux participants un grand nombre de stimuli dont, faute de mieux, on fait varier les paramètres de façon aléatoire au cours du temps. On utilise alors les réponses des participants à chaque stimuli pour faire l'ingénierie inverse des caractéristiques du bruit qui entraîne telle ou telle réponse. Ce paradigme expérimental est utilisé intensivement depuis quelques années par les groupes de Philippe Schyns à Glasgow et Frédéric Gosselin à Montréal pour étudier les expressions faciales émotionnelles, et permet de distinguer des émotions (par exemple, la joie et la surprise) qui utilisent les mêmes muscles faciaux (yeux et bouche) mais avec des décours temporels différents.

68. McDermott, J. H., Schemitsch, M. & Simoncelli, E. P. (2013). *Summary statistics in auditory perception*. Nature neuroscience, 16(4), 493-498.

69. "Predictions can be very difficult — especially about the future" - attribué à Niels Bohr

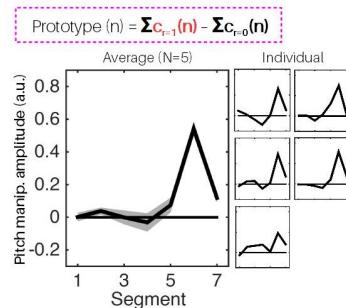
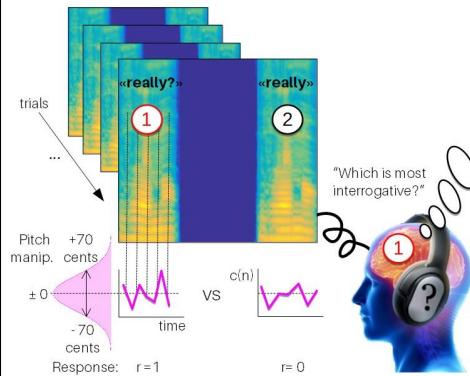


Le module d'inflection de DAVID, qui applique une forme temporelle de transformation de hauteur après chaque passage au seuil.



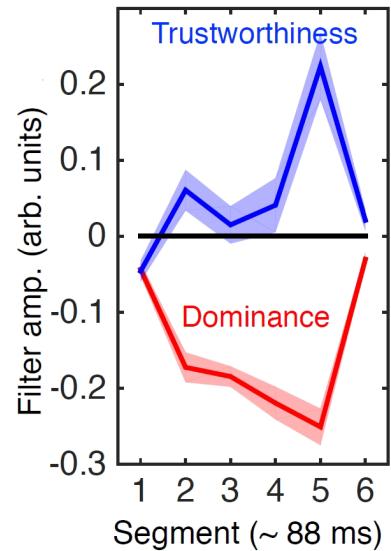
Les méthodes de *reverse correlation* présentent un très grand nombre d'animations faciales aléatoires aux participants, et d'exploiter leur jugement pour extraire quelles dimensions (dans le temps et l'espace) permettent, par exemple, de distinguer une expression de joie de celle d'une expression de surprise. (Delis, I., Chen, C., Jack, R. E., Garrod, O. G. B., Panzeri, S. & Schyns, P. G. (2016). *Space-by-time manifold representation of dynamic facial expressions for emotion categorization*. Journal of Vision 16(8) :14).

Encadré 9 : Réprésentations mentales de la forme temporelle de la prosodie interrogative



La technique de la *reverse correlation* peut être utilisée avec des techniques de transformations du signal vocal pour révéler les représentations mentales de la forme temporelle de certains types de prosodies expressives, comme celle d'une prononciation interrogative. À gauche : Différentes prononciations d'un même mot ("vraiment ?") sont manipulées avec une technique de vocodeur de phase pour les doter d'un contour de pitch aléatoire. On présente aux participants des paires de prononciations manipulées, et leur demande de juger laquelle paraît la plus interrogative. À droite : on calcule ensuite le contour moyen des stimuli jugés comme interrogatifs, auquel on soustrait celui de ceux jugés déclaratifs. Le prototype obtenu correspond au "filtre" appliqué aux données par les participants pour effectuer la tâche : les jugements interrogatifs sont associés à des prosodies montantes à la fin de la deuxième syllabe.

Ponsot, E., Buried, J.J., Belin, P. & Aucouturier, J.J. (2017) *Cracking the social code of speech prosody using reverse-correlation*, soumis.



Représentations mentales de la forme prosodique d'une prononciation dominante (rouge) et digne de confiance (bleu) du mot "bonjour", construites par *reverse correlation*.

Nous avons récemment appliqué ce paradigme à la question de la prosodie expressive. Si les linguistes et les anthropologues ont depuis longtemps noté des régularités dans les contours de hauteur utilisés pour exprimer certaines attitudes propositionnelles (l'ironie, le doute) ou sociales (l'arrogance, la conciliation)⁷⁰, on ne sait toujours pas quelles exactes formes temporelles sont utilisées dans ce type de communication. Nous avons donc conçu un algorithme de transformation de voix capable de générer un nombre arbitrairement grand de transformations prosodiques à partir d'un seul enregistrement. Il procède en découplant l'enregistrement en fenêtres (par exemple, 6 segments de 100ms dans un enregistrement du mot "vraiment" d'une durée de 600ms), et en appliquant une modification de hauteur à

70. R. L. Mitchell and E. D. Ross (2013). *Attitudinal prosody : What we know and directions for future study*. Neuroscience & Biobehavioral Reviews, 37 :471-479.

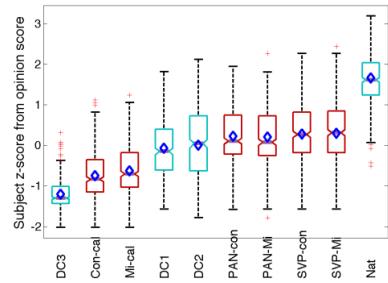
chaque segment, d'une amplitude tirée aléatoirement d'une loi normale. Ces stimuli peuvent ensuite être utilisés dans des paradigmes expérimentaux de *reverse-correlation* (Encadré 9), qui permettent de révéler les formes temporelles prototypiques sur lesquelles se basent les participants pour faire leurs jugements. Nous avons appliqué cette technique (avec E. Ponsot, J.J. Burred et P. Belin) pour révéler les profils temporels d'une prosodie interrogative, dominante et digne de confiance⁷¹, et sommes en train d'appliquer le même paradigme à la perception de la justesse dans les mélodies chantées. Cette technique ne résoud pas intégralement la question du temps dans la prosodie (en particulier, reste à découvrir comment ces formes se déplient sur des séquences de mots : sont-elles appliquées à chaque mot successivement, ou globalement, ou co-articulées, etc.?), mais permet au moins de l'opérationnaliser pour avancer sur la question de sa modélisation computationnelle.

Q3 : Qu'est-ce qu'une transformation naturelle ?

Une deuxième question de recherche qui se situe au niveau de l'état de l'art à la fois dans le domaine informatique et dans le domaine psychologique est celle de la vraisemblance ou de la "naturalité" des transformations du signal que nous proposons d'utiliser. Les études perceptives, dite de "validation", deviennent en effet la norme dans toute la communauté de traitement du signal audio, dans le but d'évaluer la qualité sonore des algorithmes, mais ces études, souvent improvisées sans véritable compétence expérimentale, ne sont pas toujours conduites d'une façon qui satisfait aux exigences de la communauté psychologique.

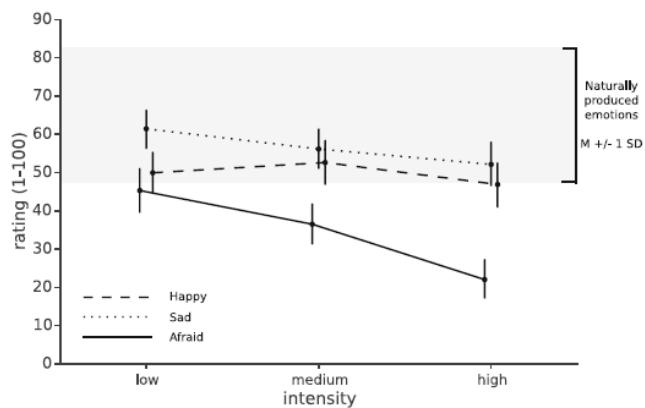
Nous nous sommes prêtés à l'exercice, avec ma doctorante L. Rachman, pour évaluer la naturalité des transformations de voix émotionnelles de DAVID (Encadré 10). Pour évaluer si les voix manipulées par l'algorithme peuvent passer pour de la parole authentique, nous avons présenté à N=20 participants des enregistrements de voix, un par un, extraits soit d'enregistrements neutres manipulés de façon émotionnelle par l'algorithme, soit d'authentiques enregistrements émotionnels des mêmes phrases, réalisés par des acteurs. Dans les instructions, les participants étaient informés que certains des enregistrements étaient produits par des locuteurs humains, et que d'autres avaient été manipulés par ordinateur. Pour chaque extrait, les participants devaient noter à quel point la voix leur paraissait naturelle, sur une échelle continue allant de "très artificielle/pas du tout naturelle" à "pas du tout artificielle/très naturelle".

71. Ponsot, E., Burred, J.J., Belin, P. & Aucouturier, J.J. (2017) *Cracking the social code of speech prosody using reverse-correlation*, en relecture



Un exemple typique d'évaluation subjective d'un algorithme de synthèse de chant, comparé à une série d'algorithmes "de contrôle" pris dans l'état de l'art. Pour un certain nombre d'extraits synthétisés par les algorithmes, on demande au participant/utilisateur d'évaluer "à quel point vous avez été satisfait de la qualité sonore", sur une échelle de 1 à 10. Feugère, d'Alessandro, Delalez, Ardaillon & Roebel (2016) *Evaluation of singing synthesis : methodology and case study with concatenative and performative systems*, in proc. Interspeech, September 8–12, 2016, San Francisco, USA

Encadré 10 : Evaluation de la naturalité des transformations de voix de DAVID



Pour évaluer la capacité des transformations algorithmiques de DAVID à passer pour de la “vraie” parole émotionnelle, nous avons demandé à des participants de noter le caractère naturel d’un ensemble de voix manipulées par DAVID, ainsi que d’enregistrements de voix authentiques. Même si les transformations ont été notées de façon moins naturelle que les enregistrements authentiques, les jugements de naturalité pour les effets *happy* et *sad* sont situés à moins d’un écart-type de la moyenne des jugements des extraits authentiques, et près d’un quart des “vraies” voix ont été jugées moins naturelles que les effets. La naturalité de l’effet *afraid* semble plus problématique, et ne se comporte comme les effets joyeux et tristes qu’à faible niveau d’intensité émotionnelle.

Rachman, L., Liuni, M., Arias, P., Lind, A., Johansson, P., Hall, L., Richardson, D., Watanabe, K., Dubal, S. & Aucouturier, J.J. (2017) DAVID : An open-source platform for real-time emotional speech transformations. *Behavior Research Methods*.

Même avec un protocole aussi simple, l’interprétation des résultats d’une telle étude est complexe. D’une part, il est important de comprendre que ce type de tâche ne se donne pas pour but d’évaluer la capacité maximale des participants à détecter la manipulation, mais leur performance typique. On peut toujours trouver une situation dans laquelle l’algorithme faillira. Par exemple, quand quelqu’un a la possibilité de comparer un enregistrement original avec plusieurs de ses transformations, il lui sera facile de remarquer que toutes les transformations reproduisent l’intonation et le rythme de l’original de façon exactement identique. Ce que les données de notre expérience montrent, c’est que, au moins dans certaines situations, certaines phrases manipulées peuvent être jugées autant voire plus naturelles

que des phrases authentiques. Cependant, l'acceptation d'enregistrements manipulés comme "véridiques" dépend considérablement du contexte d'écoute. Par exemple, dans nos études de rétroaction vocale où les participants doivent lire un texte à voix haute et où les transformations sont appliquées à leur insu, seul 14% des participants déclarent percevoir un artefact dans leur voix (Aucouturier et al., 2016). Si, dans la même situation, on avait informé les participants de la possibilité que leurs voix soient modifiées, ce taux de détection aurait vraisemblablement été plus élevé.

D'autre part, on remarque dans notre étude (Encadré 10) que les jugements de naturalité des vraies voix ne sont pas concentrés en haut de l'échelle, ce qui signifie que même des enregistrements authentiques peuvent être jugés comme étant artificiels. Ce phénomène est probablement dû au fait que nous informons les participants que certaines des voix qu'il auront à juger sont produites par ordinateur, sans leur dire lesquelles. Si l'on peut reprocher à ces instructions de réduire de façon artificielle la *baseline* à laquelle on compare les algorithmes, il semble pourtant difficile d'obtenir de nos participants expérimentaux des réponses cohérentes sans expliciter auprès d'eux ce qu'on entend par "naturel" (c'est-à-dire ici, le contraire de "manipulées par ordinateur"). En effet, juger un extrait comme "naturel" pourrait signifier également juger de l'authenticité de l'émotion⁷² (est-ce une expression sincère de joie, ou bien le locuteur fait-il semblant ?), de l'accord entre l'émotion exprimée et le contenu verbal de la phrase⁷³ (est-ce naturel d'être joyeux en déclamant "je voudrais acheter un nouveau réveil" ?), ou de la performance d'acteur du locuteur⁷⁴ (à quelle point cette tristesse est-elle "bien jouée"). Toutes ces possibilités de jugements alternatifs doivent être contrôlées dans les facteurs de l'expérience, pour être sûrs qu'on évalue bien ce que l'on veut évaluer.

Là encore, la question de l'évaluation parfaite de ce genre d'algorithmes n'est pas réglée, mais il apparaît que notre démarche scientifique nous place dans une situation privilégiée pour contribuer, par exemple en formulant des "bonnes pratiques" apprises de la communauté psychologique, à un problème qui est au cœur des préoccupations de la communauté de traitement du signal sonore.

- 72. Calvo, M. G., Gutiérrez-García, A., Avero, P. & Lundqvist, D. (2013). *Attentional mechanisms in judging genuine and fake smiles : Eye-movement patterns*. *Emotion*, 13(4), 792.
- 73. Nygaard, L. C. & Queen, J. S. (2008). *Communicating emotion : linking affective prosody and word meaning*. *Journal of Experimental Psychology : Human Perception and Performance*, 34(4), 1017.
- 74. Konijn, E. (2010). *Acting emotions*. Amsterdam University Press.

Q4 : Réductionnisme ?

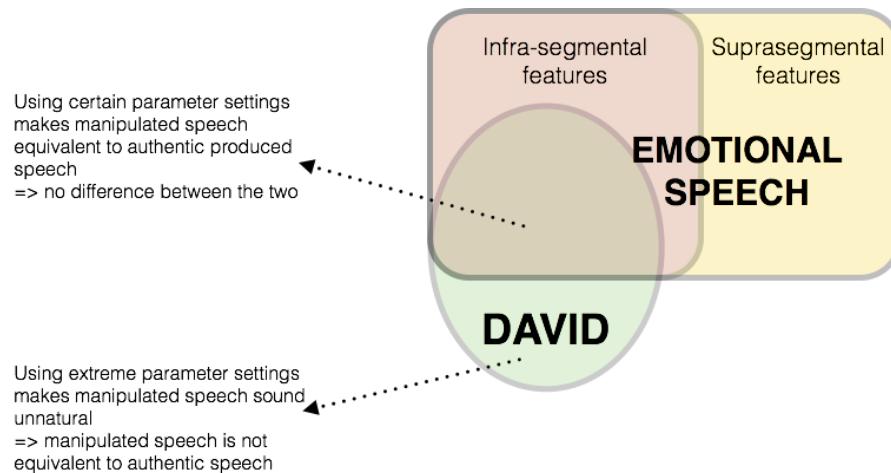
Plus mes travaux intègrent les modèles informatiques à des recherches expérimentales sur la cognition du son, et plus souvent l'on m'interroge sur le caractère "réductionniste" de cette méthodologie. L'idée de la réduction méthodologique, selon laquelle l'examen des parties et de cas spécifiques peut être utilisé pour faire des prédictions d'ordre global et général, remonte sans doute au 17ème siècle occidental, à Francis Bacon puis Descartes⁷⁵. Quand elle m'est posée, la question a presque toujours une nuance péjorative : le modèle informatique d'un comportement sonore est souvent vu non seulement comme un moyen de réduire la complexité, mais aussi de diminuer coupablement la grandeur des phénomènes que l'on étudie. D'autant plus, il me semble, quand l'on prétend toucher à quelque chose d'aussi profond que les émotions, la voix ou la musique.

Cette accusation peut se comprendre dans la mesure où un algorithme donné n'agit nécessairement que sur des aspects isolés du phénomène, ceux que l'on a pris la peine de modéliser : on agit par exemple sur le pitch ou le timbre de la voix, le rythme ou la consonance de la musique. Cette manipulation peut suffire à produire un effet, mais son caractère isolé peut aussi appauvrir l'expérience et manquer à révéler des interactions avec d'autres éléments potentiellement importants. Il me semble que cela est surtout affaire de bonne communication : quand on présente un algorithme nouveau et relativement complexe à une communauté - les neurosciences et psychologie cognitives - qui n'ont pas d'expertise en ingénierie, la tâche nous incombe en effet de ne pas "sur-vendre" les résultats (sous un vernis de technicité qui peut être impressionnant pour le non-expert) et de bien délimiter ce que peut l'outil, et ce qu'il ne peut pas. Par exemple, dans le cas de notre outil DAVID, il aurait été contre-productif de présenter l'algorithme comme étant capable de remplacer purement et simplement le recours aux voix d'acteurs pour la recherche en émotion. De nombreux facteurs contribuent à la perception d'émotions dans la voix, et ils ne sont pas tous manipulés par l'algorithme. En détail, celui-ci n'agit que sur les caractéristiques "infra-segmentales" de la voix, au niveau du phonème (cad sur le pitch, les inflections rapides, le vibrato et le timbre), mais ne manipule pas les caractéristiques supra-segmentales (par exemple, augmenter le pitch à la fin d'une phrase). Dans ce sous-ensemble des caractéristiques infra-segmentales, et pour un certain domaine de variation de ses paramètres, DAVID produit des transformations émotionnelles qui sont jugées raisonnablement authentiques par des auditeurs humains, cad que les voix produites par DAVID sont perceptivement équivalentes à ce que des locuteurs humains produiraient s'ils ne manipulaient que ce sous-ensemble de

75. *"Il faut diviser chaque difficulté en autant de parties qu'il est possible et nécessaire pour la solutionner".*
Descartes, R. 1637. *Discours de la méthode pour bien conduire sa raison, et chercher la vérité dans les sciences.*
I. Maire, Leiden, Netherlands

"Given the many factors which alter the perception of emotion in speech outside pitch and gross spectral shape, [...] if one of my PhD students based their research on DAVID manipulation rather than using real emotional speech they could not be certain to what extent the very blunt approximations offered by the system were responsible for the results rather than the effect of authentic emotional change". - Matthew Aylett (Univ. Edinburgh) réagissant à une première version de notre manuscrit *Behavior Research Methods* 2017.

facteurs. D'autre part, même restreintes à ce sous-ensemble de caractéristiques, les transformations deviennent non naturelles dès que l'on dépasse un certain domaine de variation des paramètres. Si l'on présente l'outil de cette façon, la critique de réductionnisme s'opérationalise : il s'agit de juger si cette approximation, certes simpliste, garde suffisamment de sens. Dans le cas de DAVID, vu le rôle important, déjà connu, des indices infra-segmentaux⁷⁶ et les effets déjà obtenus avec l'outil, on peut argumenter que c'est le cas.



Une autre façon de penser l'opposition entre un réductionnisme sensé découler de la dissection du signal sonore par l'informatique et le holisme présumé de l'étude de sa cognition par l'humain est, selon moi, la bonne vieille dialectique *top-down* vs *bottom-up*. Certaines utilisations des technologies du traitement du signal sonore en cognition décrites ici peuvent être *top-down*, partant de corpus d'enregistrements écologiques (pris "dans toute leur complexité") et cherchant à en dériver les principes explicatifs sous-jacents (c'est, par exemple, la démarche d'analyse de notre article *Cognition* 2017). D'autres utilisations peuvent être *bottom-up*, partant des propriétés acoustiques de comportements idéaux et en dérivant des modèles qui peuvent ensuite être testés et validés (par exemple la modélisation acoustique de l'effet du sourire sur la voix - Arias et al, 2017). La première approche commence avec la collecte de données et une description du phénomène, alors que la seconde part du mécanisme déjà formulé, mais l'important est qu'elle aboutissent toutes deux à un modèle de comportement en réponse à une perturbation acoustique (le positionnement en avance ou en retard de la réponse d'un musicien, le changement d'un phonème [a] quand on étire les lèvres pour sourire, etc.) qui puisse être testé de façon expérimentale. Les deux approches sont complémentaires, et la compréhension d'une hypothèse peut nécessiter des aller-retours entre l'une et l'autre⁷⁷.

76. Bachorowski, J. A. & Owren, M. J. (1995). *Vocal expression of emotion : Acoustic properties of speech are associated with emotional intensity and context*. Psychological science, 6(4), 219-224.

77. "Reductionism is most useful if observations made in a simplified system allow accurate predictions, or at least the generation of hypotheses, to be made when returning to the complex natural world. [On the other hand,] interpreting observations from holistic studies may require mechanistic insights gained from earlier reductionistic work or may generate hypotheses that are amenable to testing through reductionistic experimental approaches. [...] We conclude that one approach is not necessarily better than another. Observations made in test tubes that have no correlates in the real world may not be very useful biology, but the mere creation of large datasets without interpretation, or holistic cartoon models that fail to achieve concordance with empirical reality, is also of little value". Fang & Casadevall (2011) *Reductionistic and Holistic Science*. Infection and Immunity, vol. 79(4), 1401–1404

Q5 : Quelle unité théorique à tout cela ?

Le mode de travail décrit dans le Chap. 4 (Synthèse) tient un peu du *design pattern* : on identifie une question expérimentale (par exemple, le fait de s'entendre parler de façon joyeuse nous rend-il joyeux ?), on conçoit un modèle génératif de ce comportement sonore (par exemple, un algorithme de transformation temps-réel de la voix), on fait et publie l'expérience grâce à l'outil dans la communauté cognitive (Aucouturier et al. PNAS 2016), enfin on publie une description algorithmique et une validation expérimentale de l'outil dans une revue technique (Rachman et al, *Behav. Res. Methods*, 2017) pour en encourager l'usage dans d'autres expériences. Puis, on ré-itère (imitation faciale/outil *smile*, etc). Peu à peu, on imagine une boîte à outil de logiciels en open-source qui grandit, sur le site du labo au moins et - peut-être - dans les usages de la communauté.

La question difficile à laquelle on est rapidement confronté alors est : y a-t'il une unité théorique à toutes ces expériences ? Quel est le ou les (quelques) problèmes cognitifs qui traversent nos travaux⁷⁸ ? La tentation, en effet, est grande de faire des expériences parce qu'on sait faire l'outil correspondant, ou en tout cas car il devient logique d'ajouter tel ou tel comportement à la boîte à outil (un modèle génératif de rire, de la prosodie du *motherese*, de chant collectif paramétrées par la taille et la cohésion du groupe, pour ne citer que quelques possibilités discutées dans mon équipe ces dernières semaines). Si le domaine cognitif d'interprétation de tous ces comportements est grossièrement le même (les neurosciences sociales et affectives), les mécanismes sous-jacents sont assez divers⁷⁹ et le socle théorique de chaque article, et à travers lui la communauté visée, sont potentiellement nouveaux à chaque fois. La question se pose alors de la lisibilité de ces travaux par la communauté psychologique - quand la seule unité est technologique, on redevient rapidement technicien.

Une façon de répondre à cette question est de reconnaître que le fait d'interroger la cognition humaine avec des modèles de comportements sonores prescrit en partie quels domaines cognitifs sont les plus pertinents à étudier : le lien entre communication et systèmes perceptifs (car, contrairement aux voix d'acteurs, l'outil paramétrique permet l'exploration systématique de l'espace, et donc la caractérisation des systèmes impliqués par des méthodes psychophysiques comme la *reverse-correlation*), la détection d'erreur et la métacognition (car l'outil de transformation permet de faire percevoir au participant une intention ou un état qu'il n'a pas produit lui-même), et l'interaction sociale (car l'outil temps-réel permet de manipuler les signaux transmis et reçus, indépendamment). Dans chacun de ces domaines, il me semble opportun, pour la suite, de ne pas trop vite "quitter" chacun

78. “To paraphrase an old saying, beware of the man of one method or one instrument, either experimental or theoretical. He tends to become method-oriented rather than problem-oriented. The method-oriented man is shackled; the problem-oriented man is at least reaching freely toward what is most important”- Platt, J. R. (1964). *Strong inference*. Science, 146(3642), 347-353.

79. Mécanismes perceptifs émotionnels sous-corticaux (rugosité de la voix, avec M. Liuni), systèmes sensori-moteurs (imitation faciale, avec P. Arias), cognition sociale (prosodie de la *trustworthiness*, avec E. Ponsot), meta-cognition (*feedback vocal*, avec L. Rachman et L. Goupil), prise de décision, etc.

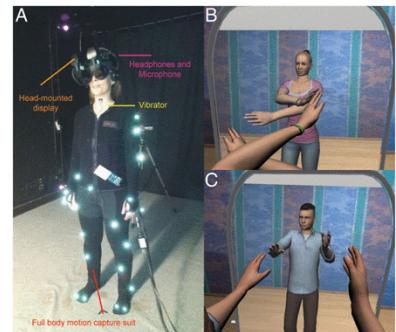
des effets expérimentaux mis en évidence, mais au contraire de les explorer chacun en profondeur, afin d'en tirer le plus de compréhension théorique possible et de les constituer en paradigmes expérimentaux réutilisables (transformer l'outil expérimental en outil théorique, en somme). Par exemple, l'effet émotionnel du feedback vocal est-il un effet d'*appraisal* ("j'entends un stimuli joyeux"), un effet de contagion ("la personne qui me parle est joyeuse") ou d'introspection ("je m'entends exprimer la joie") - autant de questions que l'outil DAVID permet de poser, et sur lesquelles on ne pourra pas faire l'impasse. Entre explorer en largeur expérimentale (plus d'outils) et en profondeur théorique (établir les plus pertinents en paradigmes), il faudra sans doute faire des choix.

Q6 : Quitter le son ?

De toute évidence, un autre parti-pris implicite de tous ces travaux est qu'ils portent, depuis toujours, sur le son. Or, même si la communauté de la cognition visuelle se nourrit depuis plus longtemps de hautes technologies de traitement de l'image⁸⁰, il apparaît rapidement que la méthodologie de transformation de comportements développée ici pourrait également servir à étudier les mécanismes de cognition visuelle. Par exemple, l'utilisation du modèle génératif de sourire parlé nous a permis de créer une tâche de jugement au seuil de perception, et de faire une analyse de détection du signal (cf. Encadré 8). Cette technique a permis de montrer que l'imitation faciale auditive apparaît sur les *hits* et les *miss*, mais pas les *false alarms*, c'est-à-dire que le comportement des muscles suit le signal et pas le jugement. Ce mécanisme de dissociation entre jugement et imitation n'a jamais été montré dans le domaine visuel malgré des centaines d'articles sur la *mimicry*, peut-être parce qu'il est difficile de créer des stimuli faciaux ambigus sans algorithme de synthèse. Nous collaborons donc depuis quelques mois avec Catherine Soladié et Renaud Séguier (Centrale Supélec Rennes) pour développer une version audiovisuelle de notre algorithme de transformation de sourire, afin de tester l'existence de ce mécanisme.

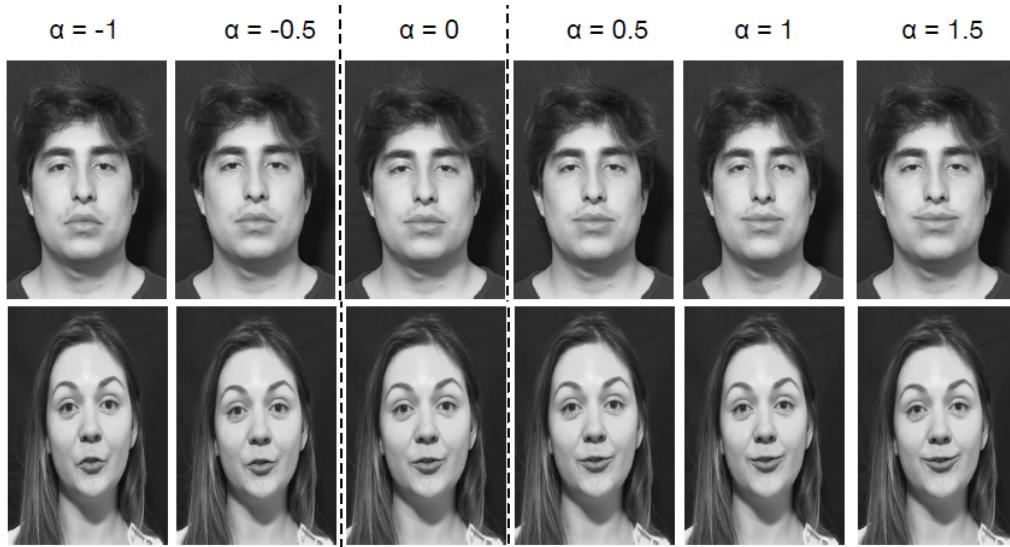
De la même façon, notre étude sur la signification sociale de la coordination temporelle et harmonique dans les improvisations musicales a suggéré que les mécanismes d'interprétation de ces indices sont probablement amodaux et sont également recrutés, par exemple, dans le *back-channel* vocal ou gestuel pendant une conversation (hocher la tête en rythme, etc.). Nous sommes donc actuellement en train de modéliser ces comportements conjointement dans la modalité visuelle et dans la modalité sonore, avec des animations visuelles contrôlées par des *sliders* par les participants (avec B. Sievers, Dartmouth).

80. Deux illustrations de cette technicité plus habituelle en vision, en plus de celle déjà mentionnée de la synthèse de visages expressif chez Schyns ou Todorov : (1) un système de réalité virtuelle dans lequel l'avatar du participant est manipuler son sentiment de *body ownership* - D Banakou, M Slater (2014) *Body ownership causes illusory self-attribution of speaking and influences subsequent real speaking*, Proceedings of the National Academy of Sciences 111 (49), 17678-17683



et (2) un miroir japonais déformant l'image du participant pour lui donner l'impression de sourire - Yoshida, S., Tanikawa, T., Sakurai, S., Hirose, M. and Narumi, T. (2013) *Manipulation of an emotional experience by real-time deformed facial feedback*, in Proceedings of the 4th Augmented Human International Conference. ACM, pp. 35–42





Mes deux doctorants Pablo et Laura, se prêtant au jeu de la transformation de sourire audiovisuelle - Arias, P. et al. *Realistic transformation of facial and vocal smiles in real-time audiovisual streams*, en relecture, 2017

Faut-il donc étendre l'argument de ce mémoire aux technologies de la transformation de signaux visuels ? Il s'agit d'une piste pour l'avenir, en particulier quand ces modèles computationnels multimodaux peuvent être construits de façon générique (même chaîne de traitement informatique et/ou même réaction mesurée) et permettent d'établir l'existence de mécanismes cognitifs amodaux ou partagés entre modalités. Là encore, il s'agit sans doute de distinguer exploration en largeur (appliquer la même métaphore aux comportements visuels, quels qu'ils soient, pour la seule raison que c'est possible et excitant d'un point de vue technologique) et en profondeur (mieux comprendre le comportement audio en étudiant ce qu'il a de commun avec le comportement visuel).

Q7 : Faut-il cultiver l'interdisciplinarité en soi, ou entre soi ?

A l'heure de conclure ce mémoire, une dernière question s'impose à moi sur le caractère interdisciplinaire des travaux décrits ici. Il me semble que l'on peut opposer deux façons de construire un parcours interdisciplinaire : soit l'on essaie de maîtriser les techniques et les connaissances des deux domaines soi-même, et donc de construire l'interdisciplinarité “en soi”, soit l'on reste dans son domaine et on collabore avec des collègues de l'autre discipline - construisant une interdisciplinarité interactive, “entre soi”. Si les deux modèles ont des avantages, le second présente celui d'une plus grande clarté institutionnelle et, vraisemblablement, d'une plus grande expertise, acquise plus

rapidement, par les praticiens de chacun des domaines. Il est clair que le parcours présenté dans ce mémoire obéit, et ce obstinément depuis quelques années, à la première de ces deux approches.

Ce n'est pas du tout un regret. Avec le recul, je ne saurais en effet comment procéder pour acquérir les codes de la communauté des sciences cognitives autrement que comme je l'ai fait, en "me prenant pour" un expérimentateur et en m'évaluant aux critères de cette communauté : manipier, me tromper sur le *design* ou les statistiques, me faire reprendre par un relecteur comme le débutant que je suis, comprendre pourquoi, refaire. Je ne vois pas comment j'en serais arrivé à faire le genre de travaux que je fais aujourd'hui si j'avais gardé au fil des années mon seul rôle d'informaticien-collaborant-avec-des-expérimentateurs - parce que je ne saurais pas en quoi je peux les aider, et eux ne sauraient pas non plus à quoi je peux leur servir. Mais tout est sans doute question de maturité des domaines : nos deux disciplines des sciences et technologies du son et des sciences cognitives sont encore très peu perméables, et les scénarios d'interaction entre les deux sont mal balisés. Il est plaisant de me dire que les travaux décrits ici rendront peut-être cette interaction plus facile à l'avenir.

Ce n'est pas un regret, mais ce n'est pas non plus une profession de foi. En effet, cette question, posée dans le contexte de la formation par la recherche, devient presque une question éthique, à laquelle je n'ai pas de réponse : est-ce un service à faire aux apprentis chercheurs qui travaillent avec moi que de les former à une méthodologie interdisciplinaire qui est pour moi le résultat d'un parcours personnel, et qui a pris quinze ans à se construire. En serais-je là aujourd'hui si je n'avais pas reçu une formation principale dans un seul domaine, l'informatique ? Comment se construiront par la suite mes jeunes collègues, formés à mes côtés à faire interagir des modèles computationnels qu'ils ne maîtrisent pas autant que s'ils avaient fait une pure thèse d'informatique, avec des protocoles expérimentaux et des méthodes d'analyse de données qu'ils ne maîtrisent pas autant que s'ils avaient fait une pure thèse de psychologie ou neurosciences cognitives ? Y a t-il une substance dans ce dialogue interdisciplinaire qui compense le temps passé à ne pas "boxer" dans une seule discipline, pour une seule section de concours, au contact d'une seule communauté de recherche ? Je n'ai pas encore, à mon niveau personnel⁸¹, le recul de suivi de carrière de mes propres étudiants permettant de répondre à cette question. Ma seule réponse possible aujourd'hui consiste à décrire ces apports méthodologiques, à les faire accepter, adopter, pour que leur explication de parcours à eux ne prennent pas 50 pages, mais 1 seule.

Et pour qu'ils puissent passer à la suite, plus vite et plus loin que moi.

81. Cette question se pose cependant tout aussi vivement au niveau de mon unité, qui investit des intersections disciplinaires de même nature que celles décrites ici. A ce titre, l'exemple de l'équipe de Frédéric Bevilacqua, étudiant les interactions entre informatique et mouvement, est encourageant, car deux des derniers doctorants viennent d'être recrutés cette année en section 7 CNRS.

Travaux cités

- Alluri, V. & Toiviainen, P. (2010) *Exploring perceptual and acoustic correlates of polyphonic timbre*. Music Perception 27(3), 223-241.
- Arminjon, M., Preissmann, D., Chmetz, F., Duraku, A., Ansermet,F. & Magistretti, P. J. (2015). *Embodied memory : unconscious smiling modulates emotional evaluation of episodic memories*. Frontiers in Psychology, 6, 650
- Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L. & Poeppel, D. (2015). *Human screams occupy a privileged niche in the communication soundscape*. Current Biology, 25(15),2051-2056
- Bachorowski, J. A. & Owren, M. J. (1995). *Vocal expression of emotion : Acoustic properties of speech are associated with emotional intensity and context*. Psychological science, 6(4), 219-224.
- Banakou, D. & Slater, M. (2014) *Body ownership causes illusory self-attribution of speaking and influences subsequent real speaking*, Proceedings of the National Academy of Sciences 111 (49), 17678-17683
- Basso, F. & Oullier, O. (2010). *'Smile down the phone' : Extending the effects of smiles to vocal social interactions* (Comment on target article). Behavioral and Brain Sciences,33(6), 436
- Bernard, Claude (1865).*Introduction à l'étude de la médecine expérimentale*. Paris : JB Baillière et Fils.
- Bigand, E., Delbé, C., Tillmann, B. & Gérard, Y. (2011) *Categorisation of extremely brief auditory stimuli : Domain-specific or domain-general processes ?* PLoS ONE 6(10) (2011)
- Blumstein, D. T., Bryant, G. A. & Kaye, P. (2012). *The sound of arousal in music is context-dependent*. Biology letters, rsbl20120374.
- Bode, H. (1984). *History of electronic sound modification*. Journal of the Audio Engineering Society, 32(10), 730-739
- Bonini, F. (2009) *All the pain and joy of the world in a single melody : A transylvanian case study on musical emotion*. Music Perception 26(3), 257-261.
- Calvo, M. G., Gutiérrez-García, A., Avero, P. & Lundqvist, D. (2013). *Attentional mechanisms in judging genuine and fake smiles : Eye-movement patterns*. Emotion, 13(4), 792.
- Casadevall, A. & Fang, F. C. (2008). *Descriptive science*. Infection and immunity, 76(9), 3835-3836
- Castelli, F., Happé, F., Frith, U. & Frith, C. (2000). *Movement and mind : a functional imaging study of perception and interpretation of complex intentional movement patterns*. Neuroimage, 12(3), 314-325
- Damasio, A. R., Everitt, B. J. & Bishop, D. (1996). *The somatic marker hypothesis and the possible functions of the prefrontal cortex*. Philosophical transactions : Biological sciences, 1413-1420.

- De Boer, B. & Kuhl, P. (2003) *Investigating the role of infant-directed speech with a computer model*. Acoustics Research Letters Online 4(4), 129-134.
- Delis, I., Chen, C., Jack, R. E., Garrod, O. G. B., Panzeri, S. & Schyns, P. G. (2016). *Space-by-time manifold representation of dynamic facial expressions for emotion categorization*. Journal of Vision, 16(8) :14
- Descartes, R. (1637). *Discours de la méthode pour bien conduire sa raison, et chercher la vérité dans les sciences*. I. Maire, Leiden, Netherlands.
- Dimberg, U., Thunberg, M. & Elmehed, K. (2000) *Unconscious facial reactions to emotional facial expressions*. Psychological science, 11(1) :86-89.
- Di Paolo, E. A., De Jaegher, H. & Gallagher, S. (2013). *One step forward, two steps back—not the tango : comment on Gallotti and Frith*. Trends in cognitive sciences, 17(7), 303-304.
- Dumas, G., Nadel, J., Soussignan, R., Martinerie, J. & Garner, L. (2010). *Inter-brain synchronization during social interaction*. PloS one, 5(8), e12166
- Ekman, P., Sorenson, E. & Friesen, M. (1969). *Pan-cultural elements of facial displays of emotions*. Science, 164(3875), 86-88
- Fang & Casadevall (2011) *Reductionistic and Holistic Science*. Infection and Immunity, vol. 79(4), 1401–1404
- Feugère, d'Alessandro, Delalez, Ardaillon & Roebel (2016) *Evaluation of singing synthesis : methodology and case study with concatenative and performative systems*, in proc. Interspeech, September 8–12, 2016, San Francisco, USA.
- Firestone, C. & Scholl, B. J. (2016). *Cognition does not affect perception : Evaluating the evidence for 'top-down' effects*. Behavioral and brain sciences, 39.
- French, Robert M. (2016), *Reflections on Connectionist Modeling*, Computational Modeling of Cognition and Behavior.
- Fröhlich, B., Rodner, E. & Denzler, J. (2012). *Semantic Segmentation with millions of Features : Integrating Multiple Cues in a Combined Random Forest Approach*. In Proc. 11th Asian conference on Computer Vision
- Gallotti, M. & Frith, C. D. (2013). *Social cognition in the we-mode*. Trends in cognitive sciences, 17(4), 160-165.
- Gigerenzer, G. & Todd, P.M. (1999) *Simple heuristics that make us smart*. New York : Oxford University Press.
- Greenwald, A. G. (2012). *There is nothing so theoretical as a good method*. Perspectives on Psychological Science, 7(2), 99-108.
- Grey, J. M. (1977). *Multidimensional perceptual scaling of musical timbres*. the Journal of the Acoustical Society of America, 61(5), 1270-1277.
- Grimm, S., Ernst, J., Boesiger, P., Schuepbach, D., Hell, D., Boeker, H., & Northoff, G. (2009). *Increased self-focus in major depressive disorder is related to neural abnormalities in subcortical-cortical midline structures*. Human Brain Mapping, 30(8), 2617–2627
- Jürgens, R., Grass, A., Drolet, M. & Fischer, J. (2015). *Effect of acting experience on emotion expression and recognition in voice : Non-actors provide better stimuli than expected*. Journal of Nonverbal Behavior, 39(3), 195–214

- Juslin, P. N. & Laukka, P. (2003). *Communication of emotions in vocal expression and music performance : Different channels, same code ?*. Psychological bulletin, 129(5), 770.
- Karydis, I., Radovanovic, M., Nanopoulos, A., & Ivanovic, M. (2010). *Looking Through the 'Glass Ceiling' : A Conceptual Framework for the Problems of Spectral Similarity*. In Proc. ISMIR (pp. 267-272).
- Koenig, W. (1949) *A new frequency scale for acoustic measurements*, Bell Telephone Laboratory Record, vol. 27, pp. 299-301.
- Konijn, E. (2010). *Acting emotions*. Amsterdam University Press.
- Lee, J. H., Jones, M. C. & Downie, J. S. (2009). *An Analysis of ISMIR Proceedings : Patterns of Authorship, Topic, and Citation*. In Proc. ISMIR (pp. 57-62).
- Lewicki, M. (2002). *Efficient coding of natural sounds*. Nature Neuroscience 5(4), 356-363
- Lewin, K. (1951). *Field theory in social science : Selected theoretical papers* (D. Cartwright, Ed.). New York, NY : Harper & Row.
- Liu, D. & Zhang, H.J. (2006) *Automatic mood detection and tracking of music audio signal*. IEEE Transactions on Speech and Audio processing 14(1), 5-18.
- Liuni, M. & Roebel, A. (2013). *Phase vocoder and beyond*. Musica, Tecnologia, 7, 73-120.
- Logan, B. (2000). *Mel Frequency Cepstral Coefficients for Music Modeling*. In Proc. ISMIR 2000, Plymouth, MA
- Mangini, M. C.& Biederman, I. (2004). *Making the ineffable explicit : Estimating the information employed for face classifications*. Cognitive Science, 28(2), 209–226.
- McAdams, S., Winsberg, S., Donnadieu, S., Soete, G., & Krimphoff, J. (1995). *Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes*. Psychological research, 58(3), 177-192
- McDermott, J. H., Schemitsch, M. & Simoncelli, E. P. (2013). *Summary statistics in auditory perception*. Nature neuroscience, 16(4), 493-498.
- Meltzoff, A.& Moore, K. (1977). *Imitation of facial and manual gestures by human neonates*. Science, 198(4312), 75-78
- Mitchell R. L. and Ross E. D. (2013). *Attitudinal prosody : What we know and directions for future study*. Neuroscience & Biobehavioral Reviews, 37 :471-479
- Molnar, Kaplan, Roy, Pachet, Pongracz, Doka & Miklosi (2008) *Classification of dog barks : a machine learning approach*. Animal Cognition 11(3), 389-400.
- Nazzi, T. & Bertoni, J. (2003). *Before and after the vocabulary spurt : two modes of word acquisition ?*. Developmental Science, 6(2), 136-142.
- Nygaard, L. C. & Queen, J. S. (2008). *Communicating emotion : linking affective prosody and word meaning*. Journal of Experimental Psychology : Human Perception and Performance, 34(4), 1017.
- Pachet, F. (2003). *The continuator : Musical interaction with style*. Journal of New Music Research, 32(3), 333-341.
- Pachet, F. & Roy, P. (2007). *Exploring billions of audio features*. In Proc. IEEE International Workshop on

- Content-Based Multimedia Indexing (CBMI).
- Platt, J. R. (1964). *Strong inference*. Science, 146(3642), 347-353.
- Radovanović, M., Nanopoulos, A. & Ivanović, M. (2010). *Hubs in space : Popular nearest neighbors in high-dimensional data*. Journal of Machine Learning Research, 11(Sep), 2487-2531.
- Serra, J., Corral, A., Boguna, M., Haro, M. & Arcos, J.L. (2012) *Measuring the evolution of contemporary western popular music*. Scientific Reports 2.
- Shahidi, S. & Baluch, B. (1991). *False heart-rate feedback, social anxiety and self-attribution of embarrassment*. Psychological Reports, 69(3), 1024-1026.
- Sievers, B., Polansky, L., Casey, M. & Wheatley, T. (2013). *Music and movement share a dynamic structure that supports universal expressions of emotion*. Proceedings of the National Academy of Sciences, 110(1), 70-75.
- Slatcher, R. B. & Pennebaker, J. W. (2006). *How do i love thee ? let me count the words. The social effects of expressive writing*. Psychological Science, 17(8), 660-664
- Stuart, A., Kalinowski, J., Rastatter, M. P., & Lynch, K. (2002). *Effect of delayed auditory feedback on normal speakers at two speech rates*. The Journal of the Acoustical Society of America, 111(5), 2237-2241.
- Teglas, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J.B. & Bonatti, L.L. (2011) *Pure reasoning in 12-month-old infants as probabilistic inference*. Science 332, 1054-1059.
- Terasawa, H., Slaney, M., Berger, J. (2005) *The thirteen colors of timbre*. In : Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA
- Tice, D. M. (1992). *Self-concept change and self-presentation : The looking glass self is also a magnifying glass*. Journal of Personality and Social Psychology, 63(3), 435.
- Tomašev, N., & Buza, K. (2015). *Hubness-aware kNN classification of high-dimensional data in presence of label noise*. Neurocomputing, 160, 157-172.
- Van Doorn, E. A., Heerdink, M. W. & Van Kleef, G. A. (2012). *Emotion and the construal of social situations : Inferences of cooperation versus competition from expressions of anger, happiness, and disappointment*. Cognition & Emotion, 26(3), 442-461.
- Yoshida, S., Tanikawa, T., Sakurai, S., Hirose, M. and Narumi, T. (2013) *Manipulation of an emotional experience by real-time deformed facial feedback*, in Proceedings of the 4th Augmented Human International Conference. ACM, pp. 35-42
- Yu, H., Garrod, O & Schyns, P. G. (2012) *Perception-driven facial expression synthesis*, Computers & Graphics, vol. 36, no. 3, pp. 152-162
- Zheng, F., Zhang, G., & Song, Z. (2001). *Comparison of different implementations of MFCC*. Journal of Computer Science and Technology, 16(6), 582-589

Publications de l'auteur

(Les collaborateurs soulignés sont des étudiant(e)s, doctorant(e)s ou postdoctorant(e)s sous ma supervision ou co-supervision)

Revues à comité de lecture

- (i) Arias, P., Belin, P. & Aucouturier, JJ. (en relecture, 2017) *Auditory smiles trigger unconscious facial imitations.*
 - (ii) Arias, P., Soladié, C., Séguier, R. & Aucouturier, JJ. (en relecture, 2017) *Realistic manipulation of facial and vocal smiles in real-world video streams.*
 - (iii) Ponsot, E., Burred, JJ., Belin, P. & Aucouturier, JJ. (en relecture, 2017) *Cracking the social code of speech prosody using reverse correlation*
 - (iv) Ponsot, E., Arias, P. & Aucouturier, JJ. (en relecture, 2017) *Mental representations of smile in the human voice.*
1. Rachman, L., Liuni, M., Arias, P., Lind, A., Johansson, P., Hall, L., Richardson, D., Watanabe, K., Dubal, S. & Aucouturier, JJ. (2017) *DAVID : An open-source platform for real-time emotional speech transformations.* Behavior Research Methods (in press)
 2. Aucouturier, JJ. & Canonne, C. (2017) *Musical friends and foes : the social cognition of affiliation and control in musical interactions.* Cognition, vol. 161, 94-108.
 3. Aucouturier, JJ., Johansson, P., Hall, L., Segnini, R., Mercadié, L. & Watanabe, K. (2016) *Covert Digital Manipulation of Vocal Emotion Alter Speakers' Emotional State in a Congruent Direction,* Proceedings of the National Academy of Science, 113(4).
 4. Boidron, L., Boudenia, K., Avena, C., Boucheix, J.M. & Aucouturier, JJ. (2016) *Emergency medical triage decisions swayed by manipulated cues of physical dominance in caller's voice,* Scientific Reports, 6, 30219.
 5. Aucouturier, JJ., Fujita, M., & Sumikura, H. (2015). *Experiential response and intention to purchase in the cocreative consumption of music : The Nine Inch Nails experiment.* Journal of Consumer Behaviour.
 6. Canonne, C. & Aucouturier, JJ. (2015). *Play together, think alike : Shared mental models in expert music improvisers.* Psychology of Music, 44(3).
 7. Lagrange, M., Lafay, G., Defreville, B., & Aucouturier, JJ. (2015). *The bag-of-frames approach : a not so sufficient model for urban soundscapes, after all.* Journal of the Acoustical Society of America, 138 (EL.487).
 8. Mercadié, L., Caballe, J., Aucouturier, JJ., & Bigand, E. (2014). *Effect of synchronized or desynchronized music listening during osteopathic treatment : An EEG study.* Psychophysiology, 51(1), 52-59.

9. Hemery, E., & Aucouturier, JJ. (2014). *One hundred ways to process time, frequency, rate and scale in the auditory cortex : a pattern-recognition meta-analysis*. Frontiers in Computational Neuroscience, 9(80).
10. Aucouturier, JJ. (2013) *All class communication, public : Using Twitter in lieu of LMS*, International Journal on Learning and Media, 4 (1), 1-7.
11. Aucouturier, JJ. & Bigand, E. (2013) *Seven problems that keep MIR from attracting the interest of cognition and neuroscience*, Journal of Intelligent Information Systems, 41 (3), 483-497
12. Ferreri, L., Aucouturier, JJ., Muthalib., M., Bigand, E. & Bugajska, A. (2013) *Music improves verbal memory encoding while decreasing prefrontal cortex activity : a fNIRS study*, Frontiers in Human Neuroscience 7.
13. Lüthy, M., & Aucouturier, JJ. (2013). *Content Management for the Live Music Industry in Virtual Worlds : Challenges and Opportunities*. Journal For Virtual Worlds Research, 6(2).
14. Aucouturier, JJ., Nonaka, Y., Katahira, K. & Okanoya, K. (2011) *Segmentation of expiratory and inspiratory sounds in baby cry audio recordings using hidden Markov models*, Journal of Acoustical Society of America, 130(5), pp. 2969-2977
15. Aucouturier, JJ. & Daudet, L. (2010) *Pattern Recognition of Non-Speech Audio*. Pattern Recognition Letters, 31(12), p. 1487-1488
16. Aucouturier, JJ. & Ikegami, T. (2009) *The Illusion of Agency : Two Engineering Approaches to Compromise Reactivity and Autonomy in an Artificial System*. Adaptive Behavior, vol.17(5).
17. Aucouturier, JJ. (2009) *Re-inventing Fourier*. Leonardo Transactions, 42(5), p. 472-473
18. Aucouturier, JJ. & Defreville, B. (2009) *Judging the similarity of soundscapes does not require categorization : Evidence from spliced stimuli*. Journal of the Acoustical Society of America, 125(4), pages 2155-61.
19. Aucouturier, JJ. & Pampalk, E. (2008) *From Genres to Tags : A little epistemology of Music Information Retrieval research*. Journal of New Music Research, vol. 32(7).
20. Aucouturier, JJ. (2008) *The hypothesis of self-organization for musical tuning systems*. Leonardo Music Journal, vol. 18, pp. 63—69.
21. Aucouturier, JJ., Ogai, Y. & Ikegami, T. (2008) *Using Chaos to Trade Synchronization and Autonomy in a Dancing Robot*. IEEE Intelligent Systems, 23(2).
22. Aucouturier, JJ., Defreville, B. & Pachet, F. (2007) *The bag-of-frame approach to audio pattern recognition : A sufficient model for urban soundscapes but not for polyphonic music*. Journal of the Acoustical Society of America, 122(2) :881-91.
23. Aucouturier, JJ. & Pachet, F. (2007) *A scale-free distribution of false positives for a large class of audio similarity measures*. Pattern Recognition, vol. 41(1), pp. 272-284.
24. Aucouturier, JJ. & Pachet, F. (2007) *The influence of polyphony on the dynamical modelling of musical timbre*. Pattern Recognition Letters, vol. 28 (5), pp.654-661.
25. Aucouturier, JJ. & Pachet, F. (2006) *Jamming with plunderphonics : Interactive contatenative synthesis of music*. Journal of New Music Research, vol. 35(1), pp. 35-50.
26. Aucouturier, JJ., Pachet, F. & Sandler, M. (2005) *The Way It Sounds : Timbre Models For Analysis and Retrieval of Polyphonic Music Signals*. IEEE Transactions of Multimedia, 7(6) :1028-1035.
27. Pachet, F., La Burthe, A. & Aucouturier, JJ. (2005) *Editorial Metadata in the Sony Music Browser : Between Universalism and Autism*. Journal of New Music Research, 34(2) :173-184.

28. Pachet, F., La Burthe, A., Zils, A. & Aucouturier, JJ. (2004) *Popular Music Access : The Sony Music Browser*. Journal of the American Society for Information Science, 55(12) :1037 - 1044.
29. Aucouturier, JJ. & Pachet F. (2004) *Improving Timbre Similarity : How high is the sky ?*. Journal of Negative Results in Speech and Audio Sciences, 1(1).
30. Pachet, F., Aucouturier, JJ., La Burthe, A., Zils, A. & Beurivé, A. (2004) *The Cuidado Music Browser : an end-to-end Electronic Music Distribution System*. Multimedia Tools and Applications.
31. Aucouturier, JJ. & Pachet, F. (2003) *Representing Musical Genre : A State of the Art*. Journal of New Music Research, 32(1).

Actes de colloques à comité de lecture

1. Liuni, M., Ponsot, E. & Aucouturier, JJ. (2017). Not so scary anymore : Screaming voices embedded in harmonic contexts are more positive and less arousing. European Society for the Cognitive Sciences of Music, Ghent, Belgium, July 2017 (accepted).
2. Aucouturier, JJ. & Canonne, C. (2017). Is musical consonance a signal of social affiliation ? European Society for the Cognitive Sciences of Music, Ghent, Belgium, July 2017 (accepted).
3. Aucouturier, JJ. & Canonne, C. (2015) *Music does not only communicate intrapersonal emotions, but also interpersonal attitudes*, Fifth International Conference on Music and Emotions, (ICME5), Geneva, Suisse.
4. Rachman, L., Liuni, M., Arias, P. & Aucouturier, JJ. (2015) *A new tool to synthesize speech-like emotional expression onto music, speech or any pre-existing audio signal*, Fifth International Conference on Music and Emotions, (ICME5), Geneva, Suisse.
5. Aucouturier, JJ. & Canonne, C. (2015) *Music does not only regulate, but directly and reliably communicates social attitudes*, Ninth Triennial Conference of the European Society for the Cognitive Sciences of Music, Manchester, UK.
6. Nonaka, Y., Aucouturier, JJ., Katahira, K. & Okanoya, K. (2015) *Developmental diversity in infant cry through maternal interactions*, Tōkyō Lectures in Language Evolution, Tōkyō, Japan.
7. Fourer, D., Shochi, T., Rouas, J. L., Aucouturier, JJ., & Guerry, M. (2014). Going ba-na-nas : Prosodic analysis of spoken Japanese attitudes. In Speech Prosody, Dublin, Ireland, 2014.
8. Fourer, D., Guerry, M., Shochi, T., Rouas, J. L., Aucouturier, J. J., & Rilliard, A. (2014). Analyse prosodique des affects sociaux dans l'interaction face à face en japonais. In XXXèmes Journées d'études sur la parole, Reims, France.
9. Nonaka, Y., Aucouturier, JJ., Katahira, K. and Okanoya, K. (2013) *Developmental differentiation in human infant cry through dynamic interaction with caregivers*, 33rd International Ethological Conference (Behaviour'13), Newcastle Gateshead, UK.
10. Aucouturier, J.-J. and Bigand, E. (2012) *Mel Cepstrum & Ann Ova : The Difficult dialog between Music Information Retrieval and Cognitive Psychology*, Proceedings of the 2012 International Conference on Music Information Retrieval.
11. Hegde, S., Aucouturier, J.-J., Ramanujam, B. and Bigand, E. (2012) *Variations in Emotional Experience During Phases of Elaboration of North Indian Raga Performance*, Proceedings of the 2012 International Conference on Music Perception and Cognition

12. Becker, A., Aucouturier, J.-J., Mougenot, C. and Yamanaka, T. (2010) *A situated experimental protocol to study emotional responses to an interactive object*, Proceedings of the 3rd International Workshop on Kansei.
13. Mougenot, C., Aucouturier, J.-J., Yamanaka, T. and Watanabe, K. (2010) *Comparing the effects of auditory stimuli and visual stimuli in design creativity*, Proceedings of the 3rd International Workshop on Kansei.
14. Aucouturier, J.-J. (2010) “*Legrand in Second-Life*” : *A virtual laboratory for real music business*, (electronic) Proceedings of Virtual World Best Practice in Education
15. Lüthy, M. and Aucouturier, J.-J. (2009) *MIR when all recordings are gone : Recommending live music in real-time*. Workshop on the Future of MIR, International Conference on Music Information Retrieval (ISMIR).
16. Aucouturier, J.-J. (2008) *What is it like to be a byte ? Analysis of student self-reports after a role-playing learning activity*. Temple University Japan Linguistics Symposium.
17. Aucouturier, J.-J. and Defreville, B. (2008) *Differences in the cognitive processing of music and soundscapes revealed by performance on spliced stimuli*, Proceedings of the 10th International Conference on Music Perception and Cognition (ICMPC), Sapporo, Japan.
18. Segnini, R., Aucouturier, J.J., Johansson, P., Hall, L. and Watanabe, K. (2008) *Feedback Effects of Emotional Voice Transformation on Self-Rated Emotion Experience*, International Conference of the Association for the Scientific Study of Consciousness (ASSC), Taipei, Taiwan.
19. Aucouturier, J.-J., Ogai, Y. and Ikegami, T. (2007) *Making a robot dance to music using chaotic itinerancy in a network of FitzHugh-Nagumo neurons*. Proceedings of the 14th International Conference on Neural Information Processing (ICONIP), Kitakyushu, Japan.
20. Aucouturier, J.-J. (2007) *The hypothesis of self-organization for musical tuning systems*. Proceedings of the European Conference on Artificial Life (ECAL), Workshop on Music and Artificial Life, Lisbon, Portugal, Sept. 2007.
21. Aucouturier, J.-J., Pachet, F., Roy, P. and Beurivé, A. (2007) *Signal + context = Better classification*. Proceedings of the International Conference on Music Information Retrieval (ISMIR), Vienna, Austria.
22. Aucouturier, J.-J. and Defreville, B. (2007) *Differences in the cognition of urban soundscapes and polyphonic music : a pattern-recognition point of view*. Proceedings of the 36th International Congress on Noise Control Engineering (INTER-NOISE), Istanbul, Turkey.
23. Aucouturier, J.-J. and Defreville, B. (2007) *Sounds like a park : A computational technique to recognize soundscapes holistically, without source identification*. Proceedings of the 19th International Congress on Acoustics (ICA), Madrid, Spain.
24. Aucouturier, J.-J. (2006) *Investigating embodiment mechanisms in music perception : The case of traditional dance music*. Proceedings of the 11th Japanese Auditory Research Forum, Shiga, Japan.
25. Aucouturier, J.-J. (2006) *The cognitive implausibility of statistical pattern recognition algorithms for music*. Proceedings of the 11th Japanese Auditory Research Forum, Shiga, Japan.
26. Aucouturier, J.-J. and Pachet, F. (2005) *Ringomatic : A Real-Time Interactive Drummer Using Constraint Satisfaction and Drum Sound Descriptors*. Proceedings of the International Conference on Music Information Retrieval (ISMIR), London, UK.

27. Roy, P., Aucouturier, J.-J., Pachet, F. and Beurivé, A.(2005) *Exploiting the Tradeoff Between Precision and Cpu-time to Speed Up Nearest Neighbor Search*. Proceedings of the International Conference on Music Information Retrieval (ISMIR), London, UK.
28. Aucouturier, J.-J. and Pachet, F. (2004) *Tools and Architecture for the Evaluation of Similarity Measures : Case Study of Timbre Similarity*. Proceedings of the International Conference on Music Information Retrieval (ISMIR), Barcelona, Spain.
29. Aucouturier, J.-J., Pachet, F. and Hanappe, P. (2004) *From Sound Sampling To Song Sampling*. Proceedings of the International Conference on Music Information Retrieval (ISMIR), Barcelona, Spain.
30. Laburthe, A., Pachet, F. and Aucouturier, J.-J. (2003) *Editorial Metadata in the Cuidado Music Browser : Between Universalism and Autism*. Proceedings of the WedelMusic Conference, Liverpool, UK.
31. Aucouturier, J.-J. and Pachet, F (2002) *Finding Songs that Sound the Same*. Proceedings of IEEE Benelux Workshop on Model-Based Processing and Coding of Audio, Leuven, Belgium.
32. Aucouturier, J.-J. and Pachet, F. (2002) *Scaling up Music Playlist Generation*. Proceedings of IEEE International Conference on Multimedia and Expo (ICME), Lausanne, Switzerland.
33. Aucouturier, J.-J. and Pachet, F. (2002) *Music Similarity Measures : What's the Use ?*. Proceedings of the International Symposium on Music Information Retrieval (ISMIR), Paris, France.
34. Aucouturier, J.-J. and Sandler, M. (2002) *Finding Repeating Patterns in Acoustic Musical Signals*. Proceedings of the 22nd International AES Conference on Virtual, Synthetic and Entertainment Audio, Espoo Finland.
35. Aucouturier, J.-J. and Sandler, M. (2001) *Segmentation of Musical Signals Using Hidden Markov Models*. Proceedings of the 110th AES Convention, Amsterdam, Netherlands.
36. Aucouturier, J.-J. and Sandler, M. (2001) *Using Long-term structure to Retrieve Music : Representation and Matching*. Proceedings of the International Symposium on Music Information Retrieval (ISMIR), Bloomington (IN), USA.
37. Reiss, J., Aucouturier, J.-J. and Sandler, M. (2001) *Efficient Multidimensional Searching Routines for Music Information retrieval*. Proceedings of the International Symposium on Music Information Retrieval (ISMIR), Bloomington (IN), USA.

Publications dans des revues sans comité

1. Aucouturier, JJ. (2016) L'apport des technologies de la musique pour la recherche en neurosciences. L'Étincelle.
2. Aucouturier, JJ (2014) Acteur de la science, ds la première année d'université. Eduquer, 108, 23-25

Chapitres de livres

1. Aucouturier, J.-J. (2011) *The hypothesis of self-organization for musical tuning systems*. In (E. Miranda, ed.) *Music and Computer Models of Living Systems*, The Computer Music and Digital Audio Series, Middleton : A-R Editions.
2. Aucouturier, J.-J. (2009) *Sounds like Teen Spirit : Computational Insights into the Grounding of Every-day Musical Terms*. In (Minett, J. and Wang, W. eds.) *Language, Evolution and the Brain, Frontiers in Linguistics Series*, Taipei : Academia Sinica Press

Numéros spéciaux de revues

1. Aucouturier, J.-J and Daudet, L., Eds. (2010) *Future Trends in the Pattern Recognition of Non-Speech Audio*. Special Issue Pattern Recognition Letters, Volume 31, Issue 12, 2010.
2. Aucouturier, J.-J and Pampalk, E., Eds. (2008) *From genres to tags : Music Information Retrieval in the era of folksonomies*. Special Issue Journal of New Music Research, vol. 32(7).
3. Aucouturier, J.-J., Ed. (2008) *Cheek to Chip : Dancing robots and the future of AI*, IEEE Intelligent Systems, Trends and Controversies, vol. 23, No. 2.

Brevets d'invention

1. **CNRS 09652-01** : Méthode et appareil de modification du timbre de la voix par décalage formantique d'enveloppes spectrales (2017), avec P. Arias, A. Roebel
2. **US2008040362** : Hybrid Audio-visual Categorization System and Method (2008-04), avec F. Pachet, P. Roy
3. **JP2006106754** : Mapped Meta-data Sound-Reproduction Device and Audio Sample Processing System Usable Therewith (2006-10), avec F. Pachet
4. **WO2006037786** : A Content Management Interface (2006-03) avec F. Pachet, P. Roy