Rapport de stage

Imitations vocales du timbre des sons environnementaux

Ali Jabbari

Maître de stage :
Guillaume Lemaitre

Co-encadrants :
Patrick Susini
Nicolas Misdariis

Ecole Phelma, Grenoble INP
Ircam, équipe Perception et Design Sonores
Mars - Juin 2014

Remerciements

Je tiens à remercier tout d'abord mon maître de stage Guillaume Lemaitre pour tout ce qu'il m'a appris, pour sa grande disponibilité et sa pédagogie exemplaire. Ce fut un réel plaisir de travailler sous sa direction et de profiter de son expertise scientifique. Ses conseils ont été vraiment encourageants et pertinents dans le cadre de ce stage, mais le resteront aussi par la suite.

Je remercie également tous les membres de l'équipe Perception et Design Sonore qui m'ont permis de travailler dans d'excellentes conditions pendant ces quatre mois de stage. Je remercie particulièrement Nicolas Misdariis et Patrick Susini pour leur gentillesse, la qualité de l'encadrement qu'ils ont su m'offrir tout au long de ce stage.

Table des matières

Chapitre 1

Introduction

1.1 Cadre et déroulement du stage

Ce stage de fin d'étude s'est déroulé à l'Ircam, au sein de l'équipe Perception et Design Sonores co-dirigée par Nicolas Misdariis et Patrick Susini. Soutenue par une politique de partenariats nationaux et internationaux, la mission de l'Ircam investit trois grands champs : Création musicale et sonore, Recherche dans le domaine des Sciences et Techniques de la Musique et du Son (STMS) et la pédagogie. Les travaux de l'équipe Perception et Design Sonores s'orientent vers l'étude et l'analyse de la perception des phénomènes sonores afin d'établir des connaissances sur les mécanismes perceptifs associés à la description et à la signification des sons, jusqu'à l'identification de sources sonores et à la perception de séquences sonores. [1] Elle applique ses recherches dans le champ du design sonore, réalisant ainsi une articulation art-science dans le domaine de la création sonore appliquée. Les travaux en perception sont articulés avec des applications en design sonore qui alimentent aussi de nouveaux axes de recherche.

Le projet SkAT-VG (Sketching Audio Technologies using Vocalizations and Gestures) [2] dans lequel ce stage s'inscrit, vise à permettre aux designers sonores d'utiliser directement leur voix et leurs gestes, pour réaliser l'esquisse sonore du son d'un produit, ce qui rend plus facile l'exploitation des possibilités fonctionnelles et esthétiques de son. Le noyau de ce cadre est un système capable d'interpréter les intentions des utilisateurs à travers des gestes et des vocalisations et de sélectionner des modules de synthèse sonore appropriées, comme il est habituellement fait dans les premières étapes du processus de conception.

1. Site web : www.ircam.fr
2. Site web : www.skatvg.eu

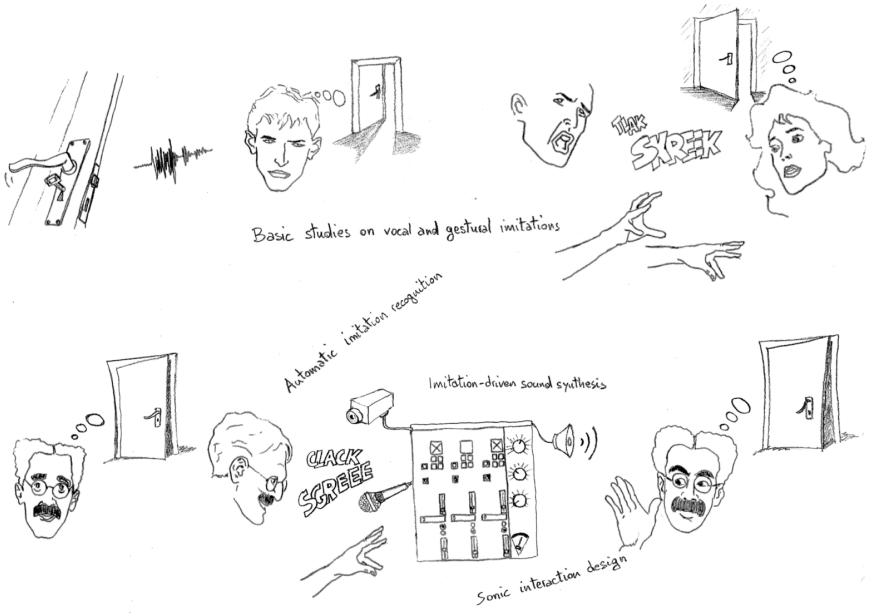

FIGURE 1.1 – Imitation vocales et gestes tant qu'outils de réaliser l'esquisse sonore du son d'un produit

Pour atteindre son objectif, le projet Skat-VG est basé sur un mélange original d'expertises complémentaires : la reproduction vocale, l'analyse du geste, la psychologie cognitive, l'apprentissage automatique, le design d'interaction, et développement d'applications audio.

Mon travail a été encadré par Guillaume Lemaitre, Nicolas Misdariis et Patrick Susini. A la suite de chaque phase d'expérience, une réunion était organisée pour présenter les derniers résultats, aborder de nouvelles problématiques et échanger des idées avec d'autres membres de l'équipe.

Apres une première période de documentation et réflexion, une expérience pilote a été mené en utilisant les bases sonores de l'équipe. L'analyse des résultats préliminaires de cette expérience nous a permis de tracer le chemin pour la suite du stage.

1.2 Contexte

La pratique de l'esquisse est à la base de la plupart des démarches de conception. Le papier et le crayon sont les outils primaires pour le concepteur (designer) ou l'architecte lorsqu'il veut matérialiser rapidement un concept ou une idée dans le

but de partager ou de transmettre. Dans le domaine du design sonore, cette transmission d'idées est rendue difficile par la nature immatérielle du sujet (le son) ainsi que par les relatives difficultés à associer des mots aux sons et la complexité des logiciels de création sonore. Par ailleurs, l'analogie graphique du papier/crayon est possible, parfois utilisée, mais s'avère néanmoins limitée lorsqu'il s'agit d'expliciter de manière fine les propriétés et les formes sonores d'une réalisation ou d'une maquette. Pour cela, les imitations vocales se révèlent être une alternative naturelle et efficace, du fait, notamment, que cette faculté d'imiter les sons est couramment utilisée par les êtres humains dans leur vie quotidienne. De plus, dans une étude récente [Lemaitre et al., 2011], il a été mis en évidence que les imitations vocales véhiculent les informations sonores nécessaires pour distinguer les grandes classes de sources sonores.

Bien qu'inscrit dans le projet SkAT, mon travail ne porte pas du tout sur l'aspect gestuel du projet. L'étude présente porte donc sur les mécanismes d'imitations vocales dans les processus d'échange et de communication sur le son ; elle s'intéresse plus particulièrement à comparer le timbre d'un ensemble de sons et le timbre de leurs imitations. Elle comporte principalement les étapes suivantes :

— élaborer un corpus de sons pour lequel un espace des timbres sera déterminé par une approche expérimentale
— déterminer les descripteurs acoustiques pertinents permettant de décrire les dimensions de l'espace,
— mettre en œuvre, la production et l'enregistrement d'imitations vocales des sons du corpus,
— effectuer une tâche d'identification (choix forcé à plusieurs alternatives) à partir des imitations.
— révéler l'espace des timbres associé aux sons imités afin de le comparer avec l'espace initial des sons réels. L'objectif est de révéler les descripteurs acoustiques les plus pertinents pour décrire le timbre du corpus étudié.

Les résultats de cette étude visent la mise en évidence d'invariants acoustiques entre les sons réels et les sons imités ; ces invariants étant les propriétés sonores qui représentent l'information générique du son à imiter.

Chapitre 2

Etat de l'art

Par définition [1], les sons environnementaux sont tous les phénomènes acoustiques audibles causés par des mouvements dans l'environnement d'humain. Les sources de ces sons sont des événements réels, ils sont plus complexes que les sinusoïdes synthétisées, ils sont significatifs, c'est-à-dire qu'ils signifient un événement dans l'environnement, et ces sons ne font pas partie des systèmes de communication humains. Ces sons sont perçus de façon littérale, sans aucune interprétation symbolique. Ainsi, ce genre de sons sont complexes et donc ils sont de nature multidimensionnelle de point de vue acoustique et perceptif.

La capacité humaine à reconnaitre des sources de sons environnementaux est essentielle dans la vie quotidienne. Grace aux techniques expérimentales d'évaluation et de mesure, nous pouvons caractériser et évaluer les propriétés acoustiques et perceptives de ces sources sonores. En plus, certaines propriétés acoustiques contribuent d'une manière importante à la construction de la représentation mentale des sources de sons. Dans cette partie de l'étude nous avons essayé de mieux comprendre la contribution des dimensions acoustiques du timbre. Nous allons ensuite étudier l'application des imitations vocales comme un outil d'exploration du timbre.

2.1 Timbre : l'attribut multidimensionnel du son

Par définition, le timbre est l'attribut perceptif qui distingue deux sons de même hauteur, sonie et durée[2]. Autrement dit, c'est le timbre musical qui distingue deux sons de même note et de même durée et intensité, joués par deux différents instruments. Cela implique la relation entre la perception du timbre et

l'identification de la source sonore. Les études préliminaires sur l'identification des instruments à partir du timbre montrent que pour les sons courts l'intégrité de la structure temporelle est cruciale pour identifier l'instrument. Cependant, pour les sons longs la partie soutenue est suffisante pour la reconnaissance[3]. Cela implique que la reconnaissance de la source du son s'appuie sur les informations temporelles et spectrales ou spectrotemporelles. Nous pouvons donc conclure que différents paramètres acoustiques peuvent être regroupé sous le terme du timbre.

Le timbre est fréquemment décrit comme l'attribut multidimensionnel des sons complexes. Nous avons adopté cette hypothèse qu'au contraire de la hauteur qui dépend largement de la fréquence fondamentale du son, et la sonie qui dépende grossièrement de l'intensité du son, le timbre s'appuie sur plusieurs dimensions acoustiques. Le but des recherches en perception du timbre est de découvrir la nature de ces dimensions.

Deux approches majeures ont été utilisées : les différentiels sémantiques[4] et plus souvent analyse multidimensionnel de proximités (MDS) des attributs de dissemblance[5]. Dans cette étude, nous avons choisi cette dernière approche et nous allons nous en servir pour la suite.

2.2 Analyse multidimensionnelles de proximités

L'idée de base de MDS est de prendre l'ensemble des proximités parmi les membres d'une collection de données, par exemple une collection de timbres et de les modéliser dans un espace euclidien avec le moins de dimensions possibles. Dans le cas du timbre, les données de dissemblance sont obtenues par des expériences psychoacoustiques dans lesquelles les sujets humains évaluent leur perception de la dissemblance entre deux stimuli d'une paire donnée. La MDS est un outil productif pour étudier les relations perceptives entre les stimuli et pour analyser les attributs sous-jacents utilisés par les sujets lorsqu'ils font des jugements de dissemblance sur les paires de stimuli. Le but de MDS est de représenter ces relations parmi un ensemble de stimuli dans un espace géométrique avec un nombre optimal de dimensions (Euclidien par défaut) telle que les distances entre les stimuli représentent les dissemblances perceptives. Une distinction importante entre différentes techniques de MDS est le choix de la typologie du modèle spatial utilisé pour représenter les distances entre les paires de stimuli.

2.2.1 Modèles MDS

L'échelle multidimensionnelle des jugements de dissemblance a été l'outil préféré pour explorer la représentation perceptive du timbre ([6],[5], etc.). Il existe diverses raisons pour ce choix : (1) les jugements sont faciles à faire pour les sujets ; (2) la technique ne fait aucune présomption sur la nature des dimensions de la représentation perceptive utilisée par les sujets pour comparer les timbres de deux sons ; (3) la représentation géométrique obtenue peut s'afficher tout de suite dans un modèle spatial.

Le modèle classique de MDS a été proposé par Torgerson[7] et Gower[**?**]. Ce modèle est implémenté dans les programmes MDSCAL et KYST. La distance euclidienne dans un espace à R dimensions, $d_{jj'}$, entre le stimulus j et j' est donnée par

$$d_{jj'} = \left[\sum_{r=1}^{R} (x_{jr} - x_{j'r})^2\right]^{\frac{1}{2}}$$

où x_{jr} est le coordonnée du stimulus j sur la dimension $r(j, j' = 1, ..., J)$. Dans ce modèle la distance entre une paire de stimuli ne dépend ni de la source des données ni du sujet. Dans ce contexte, la forme classique de MDS a été conçue pour interpréter un seul ensemble de dissemblances (la moyenne sur tous les sujets) parmi les éléments.

Le modèle INDSCAL (Individual Difference Scaling), mis au point par Carroll et Chang[9], a été la première solution robuste à ce problème. Il prend en compte les différences individuelles dans la détermination des « poids » (w) des dimensions. Ces différences individuelles déterminent l'orientation des axes résultants, ce qui exclut leur rotation. Dans ce modèle, la distance $d_{jj'}$ entre les stimuli j et j' pour le sujet n est donnée par

$$d_{jj'n} = \left[\sum_{r=1}^{R} w_{nr} (x_{jr} - x_{j'r})^2\right]^{\frac{1}{2}}$$

où x_{jr} est le coordonnée du stimulus j sur la dimension $r(j, j' = 1, ..., J)$, et w_{nr} est le poids de la dimension r associé au sujet $n(n = 1, ..., N; w_{nr} \leq 0)$.

Parmi les modèles MDS adaptés pour multiples sujets, CLASCAL proposé par Winsberg et De Soete est un modèle récent dans l'étude du timbre[10]. Le but de ce modèle est de minimiser l'erreur d'approximation dans l'équation suivante :

$$d_{ijj'} = \left[\sum_{r=1}^{R} w_{C(i),r} \left(x_{jr} - x_{j'r} \right)^2 \right]^{\frac{1}{2}}$$

où x_{jr} est le coordonnée du stimulus j sur la dimension $r(j, j' = 1, ..., J)$, et $w_{C(i),r}$ est le poids de la dimension r pour « la classe latente » $C(i)$ à laquelle CLASCAL a attribuée le sujet i. Les classes latentes sont définies pour représenter les groupes de sujets qui exercent des stratégies d'évaluation similaires. Le nombre de classes latentes utilisées est un compromis entre sur-paramétrisation du modèle INDSCAL, qui attribue chaque sujet à sa propre classe, et sur-généralisation, autrement dit ignorer les différences entre les sujets en prenant la moyenne de toutes les matrices de dissemblances. Les poids des classes peuvent être interprétés comme les stratégies d'évaluation de chaque classe. Un poids relativement élevé pour une dimension particulière implique que cette dimension joue un rôle plus important dans les évaluations des sujets de la classe.

Un autre problème avec la méthode traditionnelle de MDS c'est qu'elle assume que toute la variance parmi les données peut être expliquée par les dimensions partagées par tous les stimuli. Cela n'est pas toujours valide pour le timbre : dans beaucoup de cas, le timbre comprend des composants spécifiques. Une version plus complexe de CLASCAL sépare ces composants, appelés les spécificités, en utilisant l'équation suivante :

$$d_{ijj'} = \left[\sum_{r=1}^{R} w_{C(i),r} \left(x_{jr} - x_{j'r} \right)^2 + v_{C(i)}(s_j + s_{j'}) \right]^{\frac{1}{2}}$$

où s_j et $s_{j'}$ représentent les spécificités des stimuli j et j' et $v_{C(i)}$ est le poids des spécificités données par les sujets de la classe $C(i)$.

2.2.2 Choix du modèle MDS approprié

Dans la plupart des cas, le nombre des classes latentes du modèle CLASCAL n'est pas connu à priori . Habituellement, nous nous intéressons à vérifier si une solution à T+1 classes donne une adaptation significativement meilleure que celle à T classes. Si ce n'est pas le cas, T est considéré comme le nombre suffisant des classes pour décrire les données. Pour cela, nous utilisons un teste de Monte Carlo proposé par Hope[11] et Aitken[12].

Une fois le nombre des classes latentes est déterminé par la procédure de Hope, nous pouvons choisir le modèle de distances, avec ou sans spécificités et avec le bon nombre de dimensions communes à l'aide d'une comparaison des valeurs d'un critère d'information telle que le critère d'information d'Akaike (1977) ou bien le critère d'information bayésien proposé par Schwarz (1978).

$$BIC = -2ln(L) + ln(N)k$$

$$AIC = -2ln(L) + 2k$$

avec L la vraisemblance du modèle estimée, N le nombre d'observations dans l'échantillon et k le nombre de paramètres du modèle. Le modèle avec les valeurs les plus petites de ces critères est considéré comme la meilleure représentation des données. Dans leurs travaux, Winsberg et De Soete (1993) suggèrent l'utilisation du critère BIC.

Etant donné que les deux modèles courants (INDSCAL et CLASCAL) engendrent des espaces perceptifs dont l'invariance rotationnelle est éliminée, on pourrait supposer que les deux modèles conduiraient aux dimensions perceptives similaires. Alors que la présence des spécificités dans le modèle CLASCAL peut modifier le sens psychologique des dimensions. En effet, le fait qu'une partie des distances euclidiennes soit étalé par les spécificités conduit à la modification de la proportion d'information exprimée par les dimensions. Par conséquence les dimensions obtenues par les deux modèles ne seront pas nécessairement pareilles.

2.3 Etudes du timbre avec l'approche MDS

Dans le contexte d'étude des timbres musicaux, la méthode MDS initialement utilisée par Plomp [6] a cette avantage qu'elle ne fait aucune présomption sur les dimensions acoustiques sous-jacentes. Lorsqu'elle est appliquée aux évaluations de dissemblances des timbres, MDS produit un espace perceptif connu comme l'espace du timbre. Cet espace est dérivé des tests de dissemblances des sujets et les distances entre les timbres dans l'espace sont des distances perceptives. La dernière étape d'une telle étude de l'espace du timbre est d'interpréter les dimensions perceptives du point de vue des dimensions acoustiques sous-jacentes.

Pour que le modèle de l'espace de timbre soit correct, il faut une bonne correspondance entre l'espace physique et l'espace perceptif obtenue. Miller et Carterette[13] ont montré que les sujets peuvent se servir de la fréquence fondamentale (une propriété spectrale), l'enveloppe d'amplitude (une propriété temporelle) et le nombre des harmoniques (une propriété spectrale) dans leur évaluation.

Ils ont conclu une prédominance perceptive des caractéristiques spectrales dans les jugements de dissemblance du timbre. Dans ces études, les gammes de variation des différents paramètres n'ont pas été égalisées. Ainsi, étant donné que la variation de la fréquence fondamentale a été plus saillante, les sujets ont été empêchés d'utiliser d'autres paramètres tels que la structure harmonique dans leur prise de décision.

Samson, Zatorre, et Ramsay (1997) ont affirmé que le centre de la gravité spectrale (SCG), la moyenne en fréquence du spectre d'énergie et le temps d'attaque ont servi dans l'évaluation des dissemblances du timbre. Cependant, au contraire des études préliminaires, une plus grande contribution des propriétés temporelles et spectrotemporelles a été découverte par Grey (1977), Wessel (1979) et Iverson et Krumhansl (1993). Etant donné les corrélations acoustiques, la plupart des études amenées ([5], [14], [15]) mettent l'accent sur le rôle du centre de la gravité spectrale (SCG), la moyenne en fréquence du spectre d'énergie et le temps d'attaque et en plus, une troisième dimension qui est plus difficile à interpréter et dépend du stimuli. Des études ultérieures ont suggéré une mesure de l'irrégularité d'enveloppe spectrale[2] ou le flux spectral[14] pour la troisième dimension.

Une étude récent[16] a validé la significativité perceptive du SCG en échelle linéaire, le logarithme du temps d'attaque et l'irrégularité spectrale (modélisée en atténuation des harmoniques paires en échelle linéaire) et a rejeté celle du flux spectral. Dans cette étude les modèles CLASCAL et CONSCAL ont été comparés directement. Globalement les deux sont assez adéquats pour modéliser les estimations de dissemblance. Théoriquement, CONSCAL nécessite moins de paramètre que CLASCAL et les fonctions psychophysiques produites par analyses de CONSCAL peuvent fournir plus d'information sur le mapping des propriétés du signal dans le système auditif. Par ailleurs, les dimensions moins saillantes comme le flux spectral ont été mieux capturées en utilisant le modèle CONSCAL.

2.4 Imitation vocale en tant qu'outil d'exploration du timbre

Notre étude est basée sur l'hypothèse que la manière de description et communication de sons par des sujets humains révèle les dimensions perceptivement importantes du timbre. Globalement, les gens utilisent deux méthodes de description de sons : la description par les mots et l'imitation vocale. Les travaux préliminaires démontrent l'efficacité de chacune de ces méthodes.

Les travaux de Lemaitre, Houix et al.[17] montrent que les sujets naïfs décrivent

les sons à partir de ce qu'ils identifient comme la source des sons. La description des sons non-linguistiques (ceux qui n'appartiennent à aucun système linguistique) est une tâche assez compliquée quand la source du son référent est difficile à communiquer. Lorsqu'ils n'arrivent pas à identifier les sources, ils s'appuient sur les métaphores synesthésiques (« Le son est rugueux », etc.) ou bien ils essaient d'imiter les sons. L'imitation vocale s'apparaît donc comme un moyen adéquat pour décrire et communiquer les sons. Il y a deux façons différentes de produire une imitation vocale : l'imitation standardisée d'une langue (onomatopées) et l'imitation non-conventionnelle et créative. Les onomatopées sont très similaires aux mots et sont acoustiquement similaires aux sons qu'elles représentent. Les onomatopées sont le type le plus étudié des imitations vocales ([18], [19], [20] et d'autres).

Par contre, les imitations créatives sont acoustiquement similaires aux sons référents. Alors, ce type d'imitation n'est contraint que par la capacité vocale des locuteurs et ne dépend pas des conventions symboliques.

Le répertoire humain est capable de produire une grande variété des sons et certains (beatboxers, comédiens, etc.) ont développé des techniques vocales qui leur permettent d'imiter des sons très variés. Mais dans tous les cas, certaines contraintes persistent. Le système vocal humain peut être approximé comme un modèle source – filtre. La limite principale de ce que l'on peut produire par la voix vient du signal glottique. Il est produit par un seul système vibratoire (les cordes vocales) et donc les signaux vocaux sont normalement périodiques et monophoniques, malgré le fait que des exceptions existent. En plus, la plage de fréquence fondamentale de la voix humaine est en général contrainte entre 80 et 1100 Hz en moyenne. Cette gamme de fréquences est beaucoup plus étroite que celle des sons environnementaux. Gygi et al.[21] ont montré que la plage de 1200-2400 Hz est la gamme la plus importante pour reconnaître les sons du quotidien et donc la voix humaine n'est pas capable de reproduire une tranche très importante de fréquences de sons quotidiens.

Une autre contrainte vient de la langue maternelle des locuteurs. Les locuteurs montrent une meilleure performance à produire les sons de leur langue maternelle et rencontres souvent des difficultés à produire des sons d'une langue étrangère[22]. Alors, la reproduction de locuteur est influencée selon l'agencement des phonèmes de chaque langue.

Dans les études préliminaires[23], l'efficacité des deux méthodes de communication des sons a été étudiée avec quatre différents catégories de sons : évènements complexes identifiables (briquet, chute de pièces de monnaie, etc.), interactions mécaniques élémentaires (égouttement, rebondissement, etc.), effets sonores artificiels (sinusoïdes simples, sinusoïdes modulées en amplitude ou en fréquence, etc.) et

les sons mécaniques non-identifiables (sons de contact, etc.). Donc les sons correspondent à quatre zones d'identifiabilité. Dans la première phase de l'expérience, les sujets ont spontanément enregistré leur imitation de ces corpus. Dans la deuxième phase, les participants ont sélectionné les meilleurs description pour chaque son référent. Dans la troisième phase, pour chaque type de description le taux de succès des sujets à reconnaître les sons référents a été mesuré. Le meilleur taux de reconnaissance a été trouvé pour les sons mécaniques élémentaires décrits par verbalisation. Globalement, pour les corpus identifiables aucune différence n'a été observée pour les deux types de description. Par contre, l'imitation vocale a été significativement plus efficace dans les cas des deux corpus non identifiables. Le taux de reconnaissance correcte a exposé une forte corrélation avec la confidence en identification. L'imitation vocale donne toujours un taux de reconnaissance correcte assez bon et même un meilleur taux lorsque les sons sont plus difficiles à identifier. L'analyse des imitations efficaces et les moins efficaces montre que le taux de reconnaissance correcte est maximal lors que l'imitation reproduit en même temps le contenu spectral (hauteur, spectre d'énergie, etc.) et temporel (enveloppe temporelle, patterns temporels, etc.) des sons. Lorsque l'imitation ne reproduit pas fidèlement le contenu spectral, le taux de reconnaissance baisse mais reste toujours assez élevé. Par contre, la reconnaissance est dégradée en cas de perte du pattern temporel, même si l'information spectrale est conservée. Une autre étude menée par Lemaitre et Heller[24] suggère que les sujets montrent des meilleures performances à identifier les actions (communiqué surtout par les information temporelles) que les matériaux (communiqué partiellement par les information spectrales) des évènements sonores mécaniques.

2.5 Conclusion

Dans ce chapitre, nous avons présenté la définition et les approches d'exploration du timbre. Nous avons vu que l'imitation vocale est un outil de description du son qui permet de reconnaitre les sons référents, indépendamment de la source et son identifiabilité. Nous avons donc adopté cette hypothèse que les dimensions du timbre d'un son se conservent dans l'imitation vocale. Pour la montrer, nous nous servirons de l'approche MDS. Nous allons obtenir les espace du timbre pour un corpus homogène des sons référents et leurs imitations. Ensuite, nous allons trouver les corrélations entre les dimensions des deux espaces. Nous espérons de retrouver les mêmes dimensions perceptives dans les imitations. Si le corpus est homogène et égal en sonie, hauteur et durée, nous pouvons constater que le timbre des sons est conservé et communiqué dans la description de son par imitation vocale.

Nous pourrons aussi vérifier si les imitations vocales sont correctement reconnues même si elles ne portent pas les plus hautes fréquences du son référent. Cela implique que l'information temporelle est cruciale pour la reconnaissance du son et peut être même plus important que l'information spectrale, une nouvelle hypothèse qui reste à vérifier.

Chapitre 3

Expériences

Dans ce chapitre nous allons détailler les expériences que nous avons menées pour valider notre hypothèse. Les résultats de chaque expérience nous a permis d'envisager les prochaines étapes.

3.1 Expérience pilote

Nous avons sélectionné trois corpus de sons préexistant dans la base de sons de l'équipe dont nous avons eu l'espace du timbre. Nous avons sélectionné des corpus homogènes et que nous pensions faciles à imiter. Nous avons mené une première expérience pour valider notre choix de corpus au niveau de faisabilité d'imitation et de reconnaissance.

3.1.1 Enregistrement d'imitations

Après quelques essais avec les membres de l'équipe, nous avons enregistré des imitations faites par des sujets externes pour les corpus sélectionnés. Ici, nous allons expliquer chaque étape de cette expérience pilote.

Sélection de sujets

Pour cette expérience, nous avons recruté 3 sujets (deux hommes et une femme). Les sujets n'ont reporté aucun problème de voix ou d'audition. Ils n'ont

aucune expertise en chant ou d'autre performance vocale. Ils ont reçu une indemnisation forfaitaire pour leur participation.

Stimuli et dispositifs

L'expérience s'est déroulée dans une cabine d'audiométrie IAC. L'interface graphique pour la diffusion de stimuli et l'enregistrement des imitations a été programmée en MAX/MSP et exécutée sur un Apple Mac Pro a travers d'une carte son RME Fireface 800 et des enceintes Yamaha MSP5. Pour enregistrer les imitations vocales produites par les sujets, nous avons utilisé un microphone Schoeps CMC 5 U.

Les stimuli ont été composés de trois corpus de sons, chaque corpus a compris 16 sons :

Sons de moteur de voiture : ils ont deux composants discriminables : une composente harmonique avec une fréquence fondamentale relativement basse produit par le moteur, et une composente bruitée produite par la turbulence de l'air.

Sons de klaxons : ce sont des signaux large bande avec un spectre harmonique. Ils ont une fréquence fondamentale qui se trouve entre 300Hz et 600Hz.

Sons d'impacts : ils sont similaires aux sons instrumentaux percussifs. En plus, la structure temporelle de ces sons est discriminable.

Procédure

Au départ, le sujet a écouté l'ensemble de 16 sons afin d'avoir une impression de la variété des sons. Ensuite, chaque son est présenté dans un ordre aléatoire et le sujet a enregistré son imitation du son correspondant. Il n'y a pas de limite pour réécouter le son original et l'imitation avant passer au son suivant. Cette procédure est similaire à celle qui sera utilisée par les sujets experts (voir paragraphe 3.3.1)

3.1.2 Test de reconnaissance des imitations

Procédure

Nous avons fait ensuite un test de reconnaissance des imitations enregistrées. Les sujets de ce test ont été les membres de l'équipe. Pendant des séances d'une

heure à peu près, les sujets écoutaient chaque imitation et leur tâche a été de naviguer entre les sons originaux du corpus pour retrouver le son correspondant. Cette interface a été conçue en MATLAB (PsychToolbox[29]) et les résultats ont été enregistrés automatiquement.

Analyse des résultats

Pour chaque corpus, nous avons calculé le taux de reconnaissance correcte :

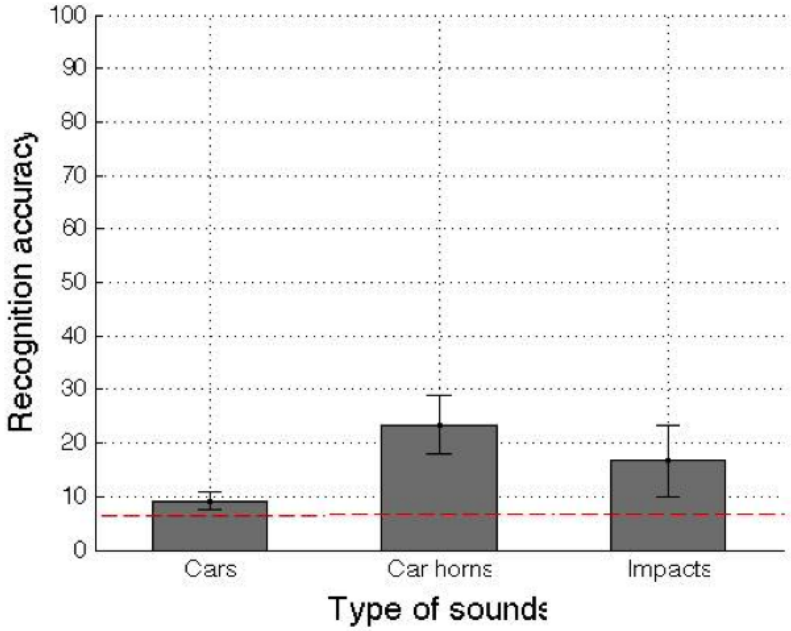

FIGURE 3.1 – Taux de reconnaissance correcte de chaque ensemble de sons. La ligne rouge indique le niveau du hasard. L'intervalle de confiance est également indiquée pour chaque enemble

Les taux de reconnaissance correcte sont justes au dessus du hasard (8 %, la ligne pointillé) pour tous les trois ensembles. L'intervalle de confiance à 95 % est indiqué sur le diagramme, ce qui permet d'évaluer grossièrement la significativité des résultats.

Nous pouvons constater que les sons de voitures ont été les plus difficiles à imiter (en accord avec l'avis des sujets après leur passage) avec le taux de reconnaissance le plus faible (inférieur à 10%).

On peut bien observer un recouvrement entre les intervalles de confiance des sons de klaxons et ceux des impacts, ce qui suggère que la différence de taux

de reconnaissance n'est pas significative. Vu le nombre très faible de sujets, nous ne pouvons pas utiliser des tests tels que t-test ou équivalents pour mesurer la significativité des résultats. En plus, c'est bien possible que le taux soit plus élevé pour ces deux ensembles parce que les sujets ont généralement distingué ces sons à partir d'autres propriétés que celles du timbre (dans sa définition standard : ni la hauteur, ni la durée, ni la sonie). C'est la hauteur dans le cas des klaxons et la durée dans le cas des impacts.

En suite, nous avons généré les matrices de confusion pour chaque ensemble. La matrice de confusion est un outil servant à visuqliser les erreurs commises d'un système de classification. Ici, l'axe vertical représente les sons originaux (la classe réelle) et l'axe horizontal représente les imitations (la classe estimée). Chaque ligne de la matrice représente le nombre de fois où chaque imitation a été associée à chaque son (en %, la somme de la ligne égale à 100 %). Dans le cas idéal, toutes les imitations sont associées aux sons originaux correspondants et la matrice de confusion est une matrice identité (confusion minimale).

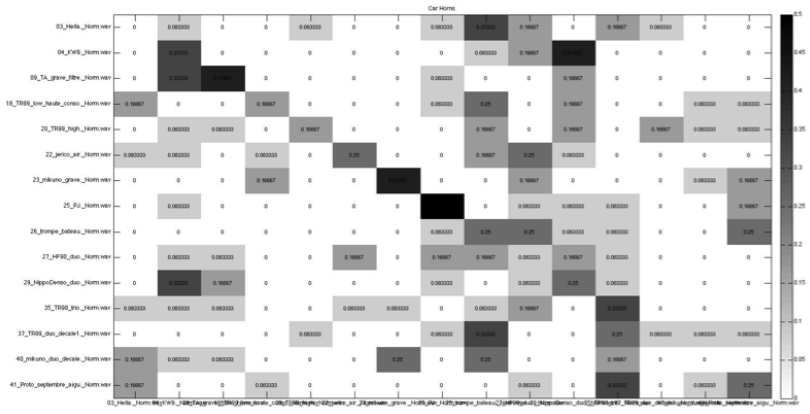

FIGURE 3.2 – La matrice de confusion pour les sons de klaxons

La matrice de confusion pour les sons de klaxons possède une diagonale vague (figure 3.2) mais les taux de reconnaissance sont très bas (toujours inférieur à 50 %). Cela est même pire pour les sons d'impacts et ceux de voiture où aucune diagonale n'est observable. La seule reconnaissance correct avec un taux relativement élevé, c'est le son d'impact « 1P62A5 1.wav » dont le taux est supérieur à 70 %. La durée significativement courte de ce son explique le succès des sujets à reconnaître les imitations correspondantes car il est en quelque sorte « unique ».

Afin de vérifier l'effet de la performance du locuteur (imitateur), la performance de chaque locuteur a été examinée pour chaque ensemble de sons :

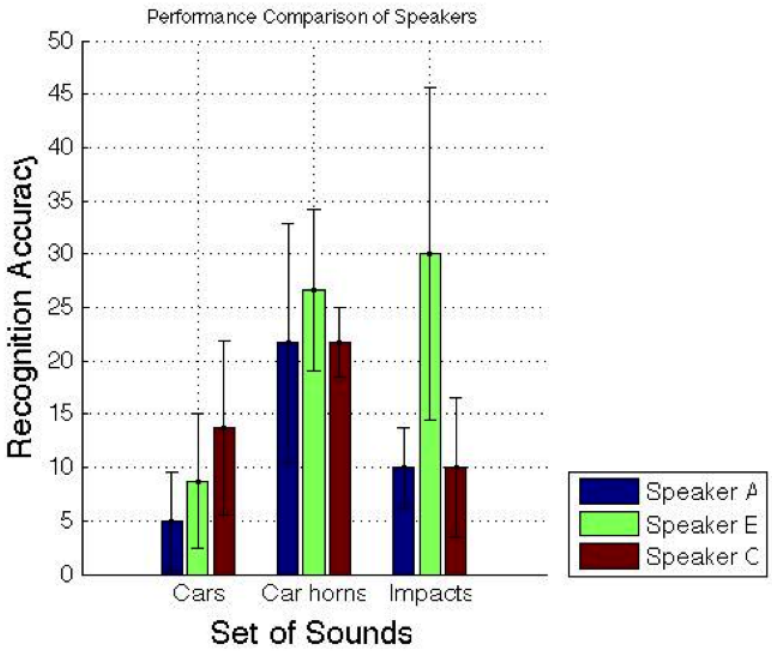

FIGURE 3.3 – Taux de reconnaissance correcte de chaque ensemble de sons pour chaque locuteur

Le recouvrement des intervalles de confusion montre qu'il n'y a pas de grosse différence entre les performances des locuteurs. Néanmoins, nous pouvons constater que les imitations des impacts issues du locuteur B ont été reconnues avec plus de succès que celles des autres (voir le recouvrement faible) et que le locuteur A a reçu des taux relativement faibles pour tous les trois ensembles. Malgré tout, même avec les estimations les plus optimistes, les taux de reconnaissance correcte n'atteignent jamais à 50 % pour aucun locuteur.

Les matrices de confusion pour chaque ensemble et chaque locuteur rendent plus de détails sur la performance des locuteurs qui valide l'analyse décrite ci-dessus. Par exemple dans la matrice de confusion de locuteur B, une vague diagonale est visible.

3.1.3 Discussion

L'analyse des résultats de l'expérience que nous avons menée nous montre que l'imitation et/ou la reconnaissance des imitations des sons choisis sont très difficiles pour les sujets de notre expérience parce que les sons se ressemblent beaucoup. Nous avons donc décidé de créer un nouveau corpus selon deux principes :

— Créer des sons très distincts.
— Recruter un expert pour répondre à la question des limites et la capacité humaine à imiter des sons.

3.2 Expérience : jugement de dissemblances pour obtenir l'espace du timbre des sons de référence

Le principe de la méthode MDS est que l'on utilise un ensemble de sons restreint et homogène, variant sur peu de dimensions. Mais du coup cela veut dire, par conséquence, que les sons se ressemblent beaucoup. Nous avons décidé de créer un corpus homogène variant sur une dimension temporelle (fluctuation) et une dimension spectrale (acuité). L'acuité est une mesure des contenus à hautes fréquences et la force de fluctuation est une mesure de la perception des variations de l'enveloppe temporelle du son. Ensuite, nous avons fait un test d'évaluation de dissemblance et à partir des données de ce test, nous avons obtenu l'espace du timbre des sons à l'aide de la technique d'analyse multidimensionnelle de proximités (MDS).

3.2.1 Création du corpus de sons

Pour avoir un échantillonnage régulier selon les deux dimensions perceptives, la force de fluctuation et l'acuité, nous avons créé plusieurs corpus de sons en utilisant des techniques de base de la synthèse sonore. Tous ces corpus sont composés des bruits filtrés par un filtre passe-bande et modulés en amplitude et pour chaque corpus, nous avons configuré les paramètres du filtrage et de la modulation. Nous avons mené des tests pilotes dans l'équipe et nous avons examiné le taux de reconnaissance des imitations pour différents corpus (différentes configurations). Finalement, nous avons adopté celui qui nous semblait le plus facile à imiter et à reconnaître.

L'acuité et la force de dluctuation sont des paramètres perceptifs. Nous utilisons des paramètres acoustiques simplifiés qui permettent de controller, à priori et pour nos sons, ces attributs perceptifs. Ainsi, les sons de ce corpus varient selon deux paramètres : la fréquence centrale du filtre passe-bande et la fréquence de modulation en amplitude. Nous avons généré des bruits blancs gaussiens et nous y avons imposé un filtre passe-bande. La largeur de bande est d'une bande critique (dépendant de la fréquence centrale). Nous avons ensuite modulé ces signaux en amplitude avec le maximum de la profondeur de modulation. Nous y avons ensuite

appliqué une enveloppe temporelle dont le temps d'attaque est égal à 10 ms pour l'ensemble des sons (voir annexe B pour plus de détails).

Pour que les sons soient faciles à imiter, nous avons mené plusieurs essais pour déterminer la forme des filtres et la gamme de fréquences centrales et celle de fréquences de modulation. Dans les gammes choisies, les modèles de Zwicker proposent un rapport linéaire entre la fréquence centrale et l'acuité et un rapport linéaire entre le logarithme de base 2 de la fréquence de modulation et la force de fluctuation. Ainsi, ce modèle suggère que dans cette gamme de fréquences de modulation, le rythme est fortement corrélé avec la force de fluctuation.

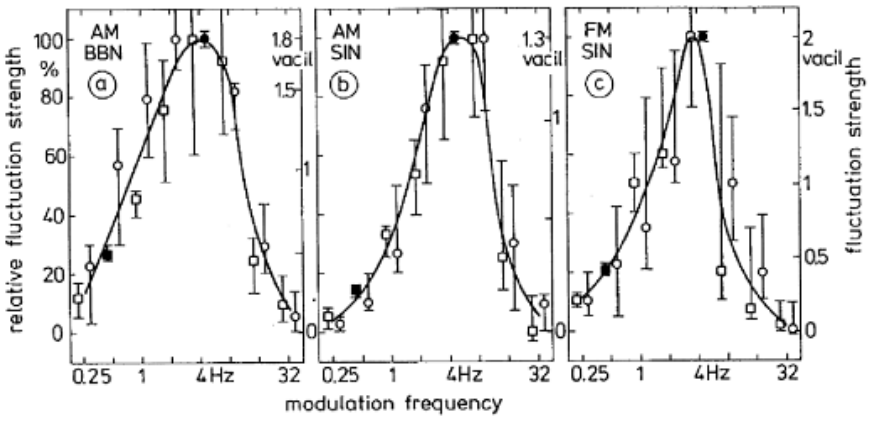

FIGURE 3.4 – La relation logarithmique entre la force de fluctuation et la fréquence de modulation pour les bruits large-bande modulés en amplitude

Nous avons donc créé un grand nombre de sons en variant la fréquence centrale du filtrage et la fréquence de modulation en amplitude et nous en avons échantillonné les 16 sons finaux de manière à avoir un espace homogène en deux dimensions perceptives : la force de fluctuation et l'acuité (figure 3.5).

Ce corpus a été conçu sous MATLAB avec une fréquence d'échantillonnage de 44,1kHz et a été égalisé en sonie.

3.2.2 Expérience de dissemblance

Nous savons que dans l'évaluation de la dissemblance par les sujets humains, certaines propriétés acoustiques dominent perceptivement les autres. Par exemple, les sujets jugent deux sons principalement différents en sonie à base de cette dimension sans prendre en compte la contribution des autres dimensions. Afin d'éliminer

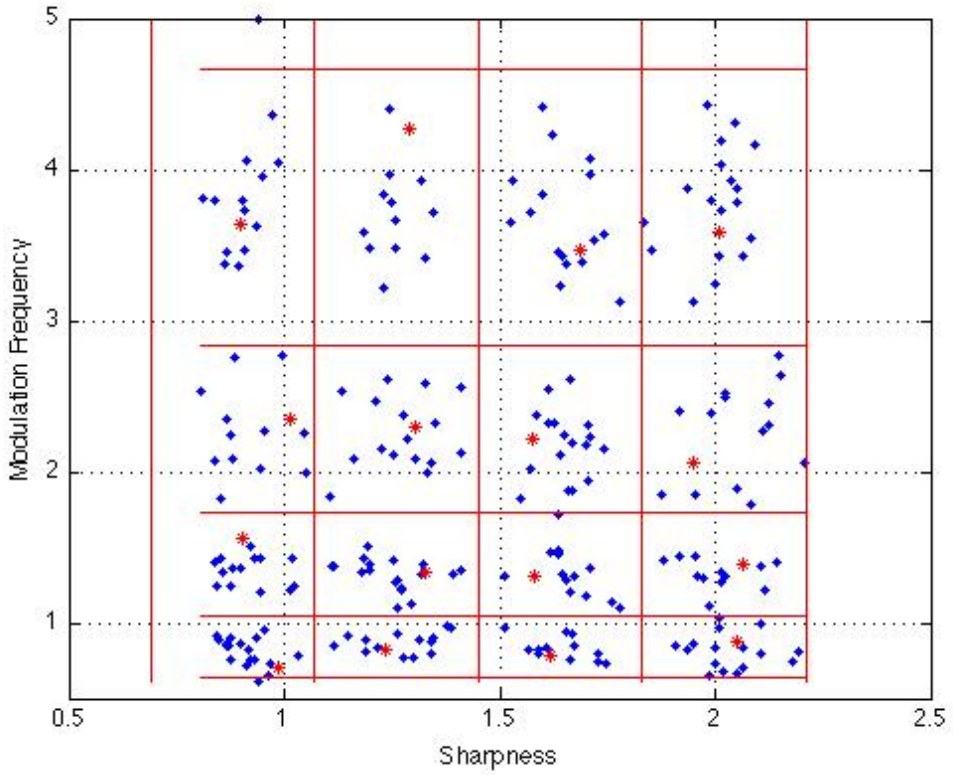

FIGURE 3.5 – Nous avons créé un grand nombre de sons (les points bleus) en variant la fréquence centrale du filtrage et la fréquence de modulation en amplitude et nous en avons échantillonné 16 sons du corpus (les étoiles rouges)

cet effet, nous avons égalisé le corpus de sons en sonie. Aussi, les durées des sons sont identiques et égales à 3 secondes.

En suivant la méthode décrite ci-dessous, nous avons demandé aux sujets de l'expérience d'évaluer la dissemblance parmi les sons du corpus synthétisé et les sons d'impacts. Pour l'instant nous nous intéressons à l'analyse des résultats du corpus synthétisé.

Sélection de sujets

Nous avons recruté 24 sujets (11 hommes 13 femmes) âgés de 18 ans à 34 ans (médiane 24 ans). Les sujets n'avaient aucun problème d'audition et aucune expertise en musique. Nous avons choisi des sujets naïfs pour nous assurer qu'ils ne suivent aucune stratégie d'écoute d'expert dans leur évaluation.

Stimuli et dispositifs

L'expérience s'est déroulée dans une cabine d'audiométrie IAC. La diffusion des stimuli a été programmée sur un Apple Mac Pro à travers d'une carte du son RME Fire-face 800 et des enceintes Yamaha MSP5.

Procédure

Au départ, le sujet a écouté l'ensemble de 16 sons du corpus afin d'avoir une impression de la variété des sons. Ensuite, chacune des 120 paires possibles des 16 sons (sans compter la rotation des paires AB, BA) a été présentée au sujet. Chaque paire a été composée de deux sons tirés au hasard, séparés avec une seconde de silence entre eux. Le sujet a évalué la dissemblance entre les deux sons en glissant le curseur graphique sur une échelle dont les deux extrémités ont été libellées « très semblable » et « très différant ». Le sujet a eu la possibilité de réécouter chaque paire de sons à plusieurs fois. Il passait automatiquement à la paire suivante en enregistrant son jugement.

3.2.3 Analyse multidimensionnelle des proximités

Les données utilisées dans une analyse MDS sont constituées des jugements de dissemblance pour chaque paire possible des stimuli considérés. C'est-à-dire $N(N-1)/2$ jugements pour N stimuli pour chaque sujet. Les valeurs de jugements de dissemblance sont enregistrées dans une demi matrice sur une échelle de 0 (très semblables) à 1 (très différents). Nous avons calculé la corrélation entre les sujets pour vérifier s'il existe des cas aberrants parmi les sujets (Figure 3.6). Nous avons exclu un sujet aberrant.

Nous avons assumé que le test de dissemblance est métrique. Les données d'un test sont considérées comme métrique si :

— La distance entre un objet et lui-même est zéro.
— La distance entre l'objet A et l'objet B est égale à la distance entre l'objet B et l'objet A.
— La distance entre l'objet A et l'objet C est inférieure où égale à la somme des distances AB et BC.

Nous avons choisi le modèle INDSCAL car ça nous permet de comparer les stratégies des sujets. Nous avons utilisé l'algorithme SMACOF[30] dans le logiciel *R*.

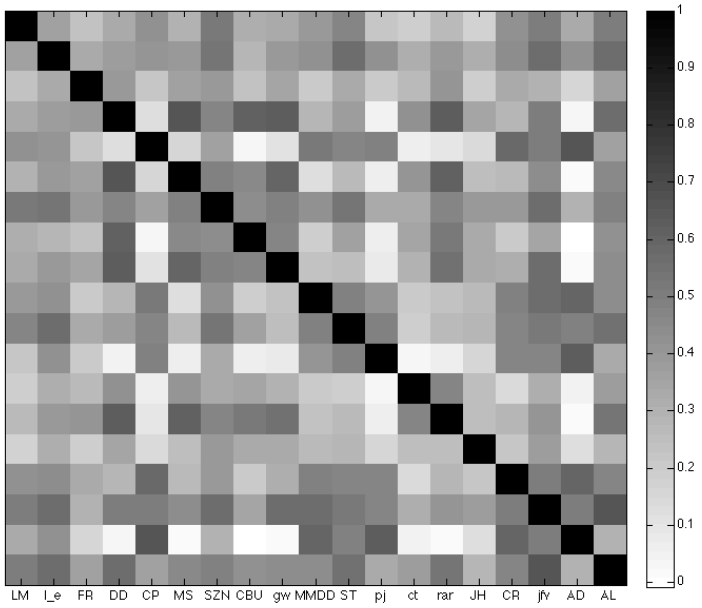

FIGURE 3.6 – La corrélation entre les sujets de l'expérience de dissemblance du corpus synthétique

En observant l'espace MDS obtenu et l'espace de synthèse (figure 3.7), nous pouvons constater que la même configuration géométrique et homogénéité est respecté.

Dimensions de l'espace perceptif

A partir de l'espace MDS obtenu, nous voulons trouver les corrélations acoustiques des dimensions perceptives. Pour le calcul des descripteurs nous avons utilisé les algorithmes et modèles suivantes :

Dimension perceptive	Algorithme utilisé
Acuité	méthode de Fastl[31]
Force de fluctuation	modèle de Holdrich[32]
Sonie	modèle de Zwicker[33]
F_0	YIN[34]
autres	IrcamDescriptors 0.43[35]

TABLE 3.1 – Les descripteurs utilisé pour le calcul des dimensions perceptives

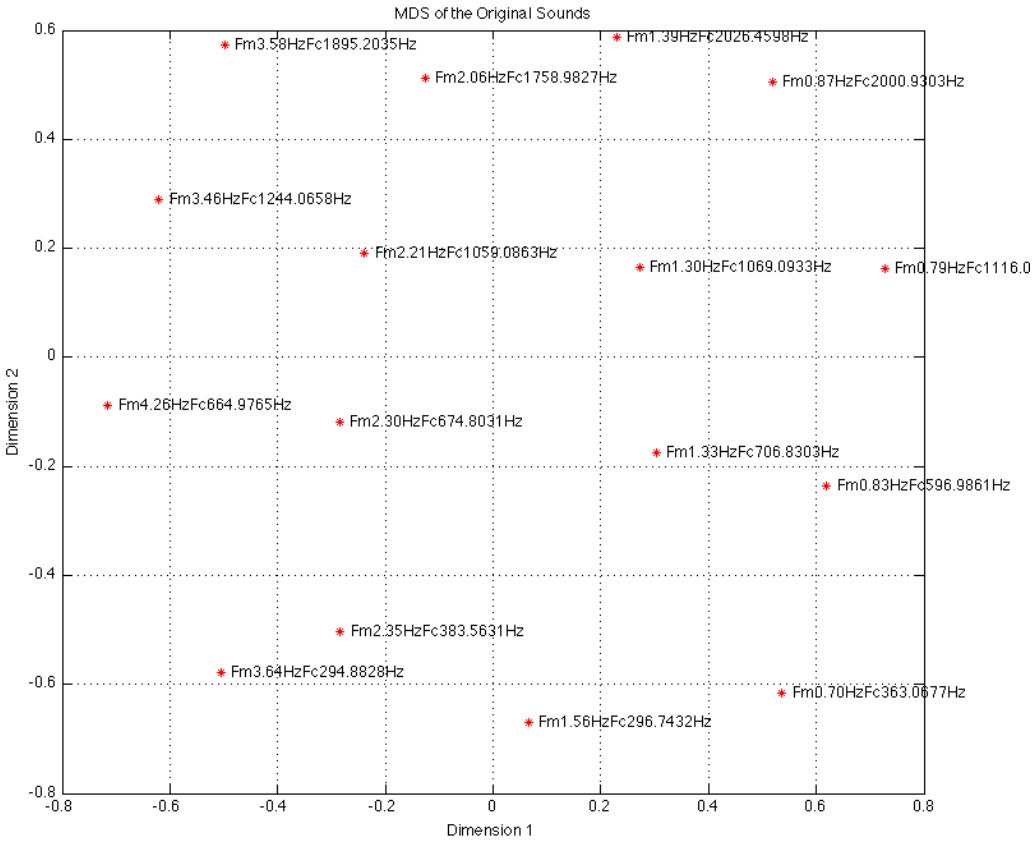

FIGURE 3.7 – L'espace MDS du corpus synthétique calculé par le modèle INDSCAL. Nous avons utilisé l'algorithme SMACOF du *R*

Nous avons ensuite calculé la corrélation entre les dimensions de MDS et les valeurs de ces descripteurs pour chaque son. Les résultats obtenus montre une forte corrélation entre la dimension 1 de l'espace perceptif et la force de fluctuation, aussi entre la dimension 2 et l'acuité :

	Dimension 1	Dimension 2
Acuité	0,07	**0,99**
Force de fluctuation	**0,97**	0,10
Sonie	0,09	0,56
F_0	0,02	0,86

TABLE 3.2 – Correlations entre les dimensions de l'espace obtenu et les grandeurs psychoacoustiques

3.3 Expérience de dissemblance sur les imitations

Nous pouvons donc conclure que le corpus conçu est homogène et variable selon deux dimension perceptives : l'acuité et la force de fluctuation. Nous avons validé notre choix de corpus et selon notre hypothèse nous attendons un espace similaire pour les imitations de ce corpus. Ainsi, dans la prochaine étape, nous allons enregistrer des imitations de ce corpus.

3.3.1 Enregistrement d'imitations

Afin de nous assurer de la qualité d'imitations, nous avons recruté deux locuteurs, un homme et une femme, avec des expertises en musique et en reproduction vocale. Les sujets n'ont reporté aucun problème d'audition et ils étaient de la langue maternelle française.

Sylvie Levesque comédien, chanteuse, professeur d'Art dramatique au CRD de St Quentin.

Richard Dubleski acteur, musicien (percussion), compositeur et metteur en scène.

Dans les deux cas, ce sont des spécialiste du répertoire contemporain, et familiers des techniques vocales étendues propres à ce répertoire.

Stimuli et dispositifs

Pendant des séances de quatre heures, les sujets ont enregistré leur imitation des sons des quatre corpus sélectionnés. Ces corpus sont composés de 16 sons de voitures, 15 sons de klaxons, 15 sons d'impacts et 16 sons synthétiques (voir section 3.2.1 pour plus amples détails sur la procédure de la synthèse). Pour l'instant, nous ne nous intéressons qu'aux imitations du corpus synthétique et les autres nous serviront plus tard.

L'enregistrement s'est déroulé dans une cabine d'audiométrie IAC par intermédiaire de la même interface graphique MAX/MSP que nous avons utilisée dans l'expérience pilote. La diffusion des stimuli a été programmée sur un Apple Mac Pro à travers d'une carte du son RME Fire-face 800 et des enceintes Yamaha MSP5.Tous les stimuli ont été égalisés en sonie à 76 phones avec un algorithme de MATLAB basé sur l'implémentation des algorithmes de sonie de Zwicker[33] [1].

1. Disponible sur le site de Genesis : www.genesis-acoustics.com

Procédure

Le locuteur reste seul dans la cabine pendant l'expérience. Au départ, il écoute l'ensemble des sons du corpus afin d'avoir une impression de la variété existante parmi les sons. Ensuite, chaque son du corpus se présente dans un ordre aléatoire. Il n'y a pas de limite pour réécouter les sons ou le nombre d'essai d'enregistrement. En revanche, une fois que le sujet valide son enregistrement, ce n'est plus possible de revenir au stimulus précédent. De même façon, le sujet enregistre ses imitations des quatre corpus.

Ensuite, nous avons nettoyé les enregistrements en supprimant les souffles, clic de souris, etc. Puis, nous avons ajouté des fade-in et fade-out aux deux bouts des sons pour éliminer les artefacts. Nous avons égalisé la sonie des sons à 76 phones avec le même algorithme.

Test de reconnaissance des imitations

Nous avons conduit le même test de reconnaissance que nous avons mené dans l'expérience pilote sur les imitations enregistrées par les locuteurs experts. Les résultats de ce test montre un taux de reconnaissance autours de 70 % . Le taux de reconnaissance correcte pour les deux locuteurs sont comparés dans la figure 3.8.

Malgré les taux de reconnaissance assez proches pour les deux locuteurs, nous pouvons observer une différence dans les matrices de confusion (figure 3.9). En plus, en écoutant les enregistrements nous avons découvert des imitations similaires pour des sons différents. Il y en a plusieurs dans les enregistrements du locuteur B.

3.3.2 Expérience de dissemblance d'imitations

Afin d'obtenir l'espace perceptif des imitations et de pouvoir le comparer avec celui des sons originaux, nous avons mené une expérience d'évaluation de dissemblance similaire à celle effectuée sur les sons originaux. Nous avons fait deux tests d'évaluation de dissemblance séparés : le premier avec tous les imitations d'un seul locuteur et le deuxième avec 8 sons imiter par les deux locuteurs.

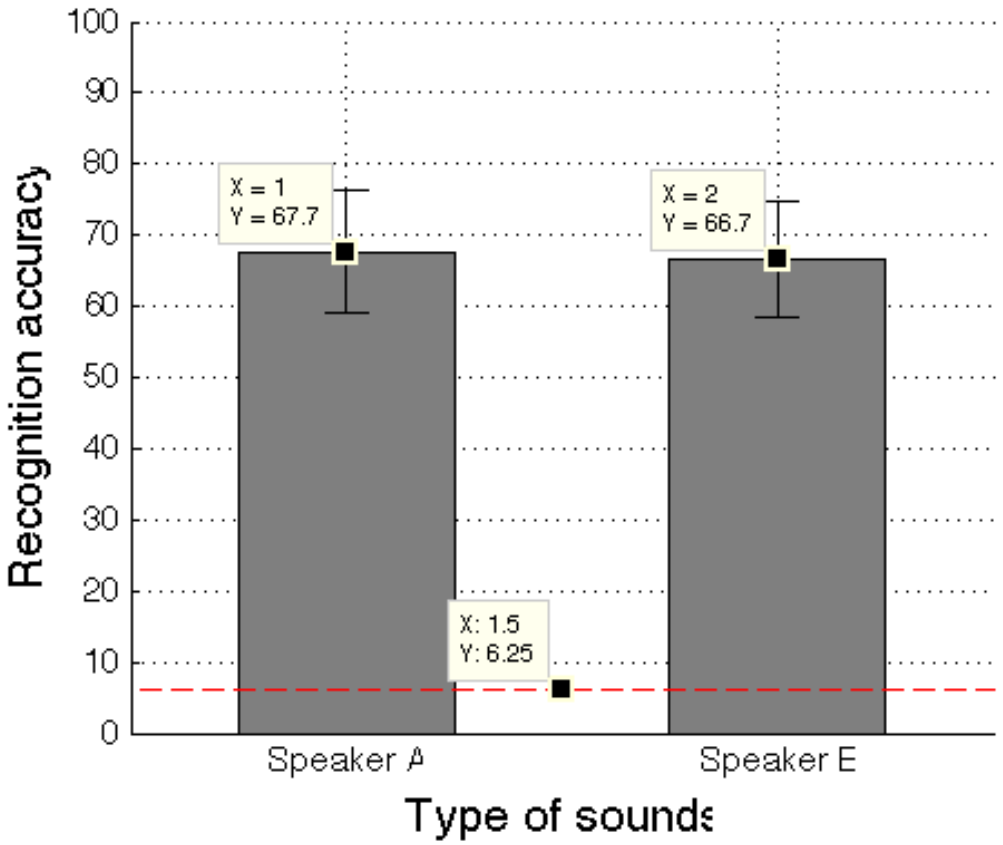

FIGURE 3.8 – Les taux de reconnaissance d'imitations des deux locuteurs. La ligne pointillée indique le niveau du hasard

Sélection de sujets

Nous avons recruté 23 sujets (7 hommes 16 femmes) âgés de 18 ans à 43 ans (médian 26 ans). Les sujets n'avaient aucun problème d'audition et aucune expertise en musique. Nous avons choisi des sujets naïfs pour nous assurer qu'ils ne suivent aucune stratégie d'écoute d'expert dans leur évaluation. Pour éviter l'influence du répertoire phonologique propre à chaque langue, tous les sujets sélectionnés ont été de la langue maternelle française. Ils ont été indemnisés pour leur participation.

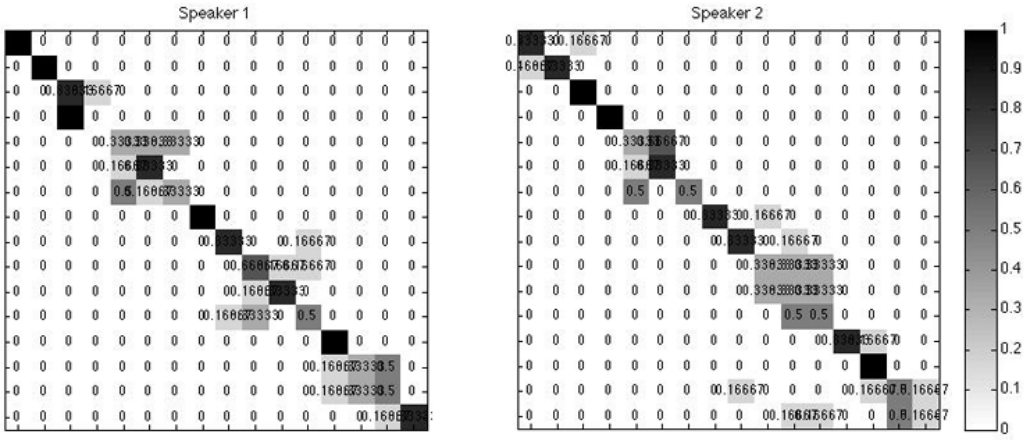

FIGURE 3.9 – Les matrices de confusion des deux locuteurs pour le corpus synthétisé

Stimuli et dispositifs

Le test de reconnaissance est composé de deux parties avec différents stimuli. Dans la première partie, nous avons mené le test de dissemblance sur l'ensemble des imitations du locuteur A. Les 16 stimuli de la deuxième partie consistent en des imitations produites par les deux locuteurs pour 8 sons sélectionnés d'une façon homogène. Pour cela nous avons sélectionné les 4 sons au centre et les 4 sons aux coins de l'espace perceptif du corpus.

Nous avons utilisé les mêmes appareils que la dernière expérience d'évaluation de dissemblance (voir la section 3.2.2 pour les détails).

Procédure

Dans chaque partie, les 16 imitations enregistrées sont présentées à sujets. La tâche des sujets est de juger le degré de dissemblance entre les deux sons de chaque paire possible (120 paires au total dans chaque partie) à travers de l'interface graphique conçue sous PsiExp. Les réponses de chaque sujet sont enregistrées sous la forme d'une demi matrice.

3.3.3 Analyse des résultats

D'abord nous avons calculé la corrélation entre les sujets de l'expérience. Les zones plus claires dans la figure 3.10 signifient les corrélations fortes entre les sujets.

L'arbre hiérarchique de dissemblance des sujets suggère la présence de quelques sujets aberrants (figure 3.10).

Nous avons calculé dans MATLAB les valeurs des deux descripteurs acoustiques qui nous intéressent, l'acuité et le rythme, pour les imitations du locuteur A et nous les avons comparé avec celles des sons originaux. Les résultats sont présentés dans la figure 3.11 :

Nous pouvons observer que pour le rythme, le rapport entre les sons originaux et les imitations est pseudo-linéaire et le rythme est conservé dans l'imitation. Par contre, ce rapport n'est plus respecté dans le cas de l'acuité où on voit grossièrement deux zones d'acuité : la première autours de 1,6 acum et la deuxième autours de 1,8 acum. Le locuteurs n'ont pas réussi à imiter les quatre gammes de l'acuité présentes dans les sons originaux mais ils ont différencié deux niveau de l'acuité (graves/aigus).

Ensuite, nous avons calculé les corrélations entre les valeurs des descripteurs acoustiques des sons originaux du corpus et des imitations du locuteur A. Cela montre que les descripteurs du rythme tels que tempo et l'estimation de la fréquence de modulation des imitations sont fortement (à $r(N = 16) = 0,99$) corrélé avec ceux du corpus original. Mais dans le cas de l'acuité cette corrélation est très faible. Nous pouvons donc conclure que l'acuité n'est pas incorporée directement dans les imitations mais le rythme est communiqué à travers des imitations.

Ensuite, nous avons calculé l'espace MDS pour les données du test de dissemblance en utilisant les modèles classique et INDSCAL à diverses dimensionnalités. Les digrammes de stress pour les modèles MDS classique et INDSCAL suggèrent que le taux de stress est tolérable pour les modèle à 2 dimensions et plus (voir annexe C pour les résultats MDS obtenus et l'estimation du stress).

Les descripteurs que nous avons utilisés jusqu'à présent sont conçus pour les sons musicaux. Pour calculer les descripteurs vocaux corrélés aux dimensions de l'espace perceptif des imitations tant que les formants, nous avons employé le logiciel *Praat*[36]. Nous avons ensuite calculé les corrélations entre tous les descripteurs et le modèle MDS classique à une dimension et les modèles INDSCAL à 2, 3, 4 et 5 dimensions (voir annexe D).

Dans toutes les solutions, il y a (au moins) une dimension qui est fortement corrélée avec l'aspect rythmique des imitations. La corrélation de l'acuité avec les dimensions des modèls MDS est très faible mais d'autres propriétés spectrales sont aussi corrélées. Ce sont parmi d'autres l'étalement spectral, les coefficients du tristimulus, les MFCC (Mel-Frequency ceptral coefficients) et la largeur du deuxième formant de la parole.

Dans la meilleure solution avec les corrélations les plus élevées, modèle INDSCAL à deux dimensions, la deuxième dimension est bien corrélée avec un coefficient du tristimulus. Nous pouvons constater que la deuxième dimension de l'espace des imitations est corrélée à la répartition de l'énergie du spectre en hautes fréquences.

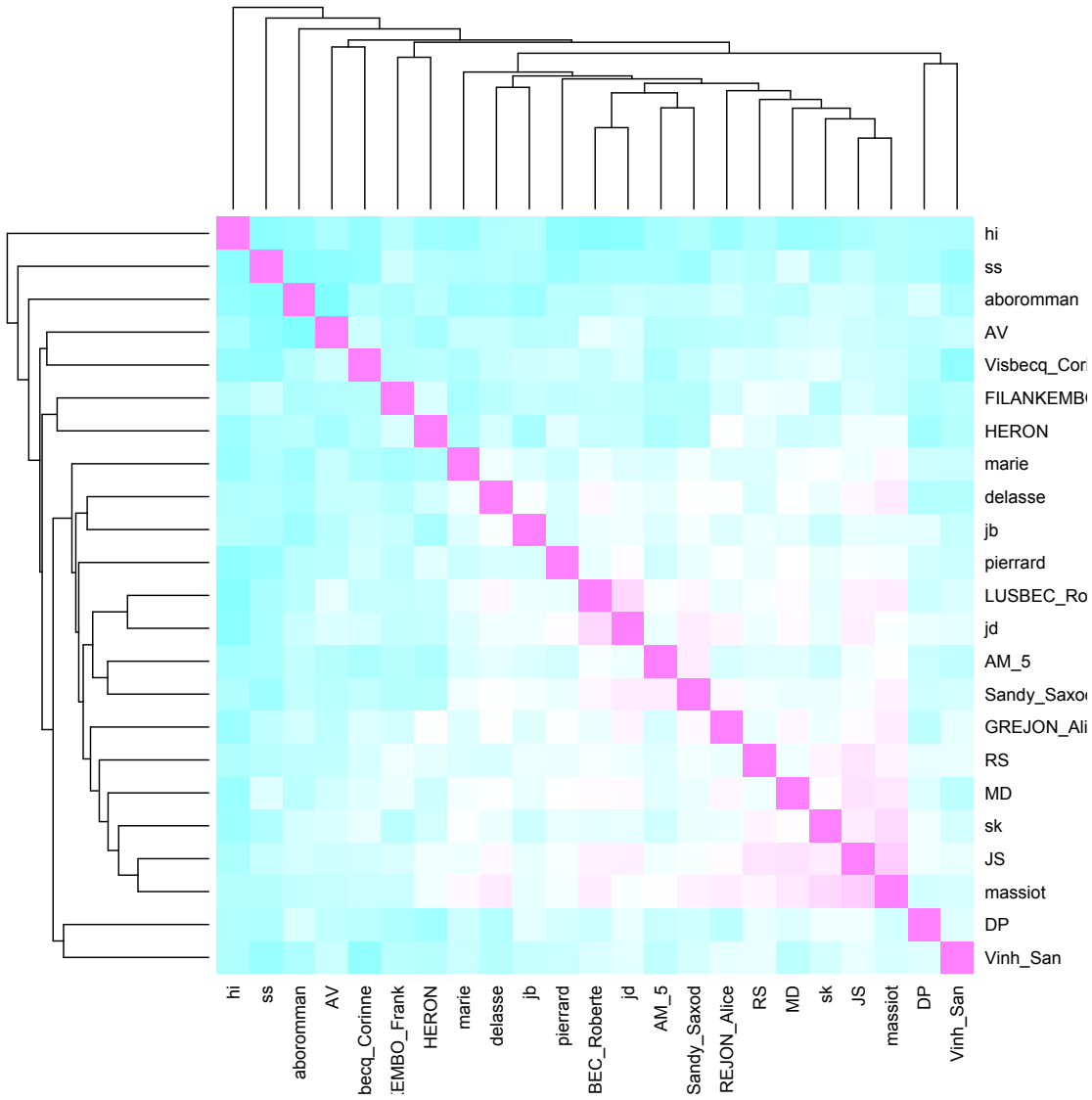

FIGURE 3.10 – Les corrélations réorganisées entre les sujets de l'expérience de dissemblance des imitations. L'arbre hiérarchique de dissemblance des sujets suggère la présence de quelques sujets aberrants.

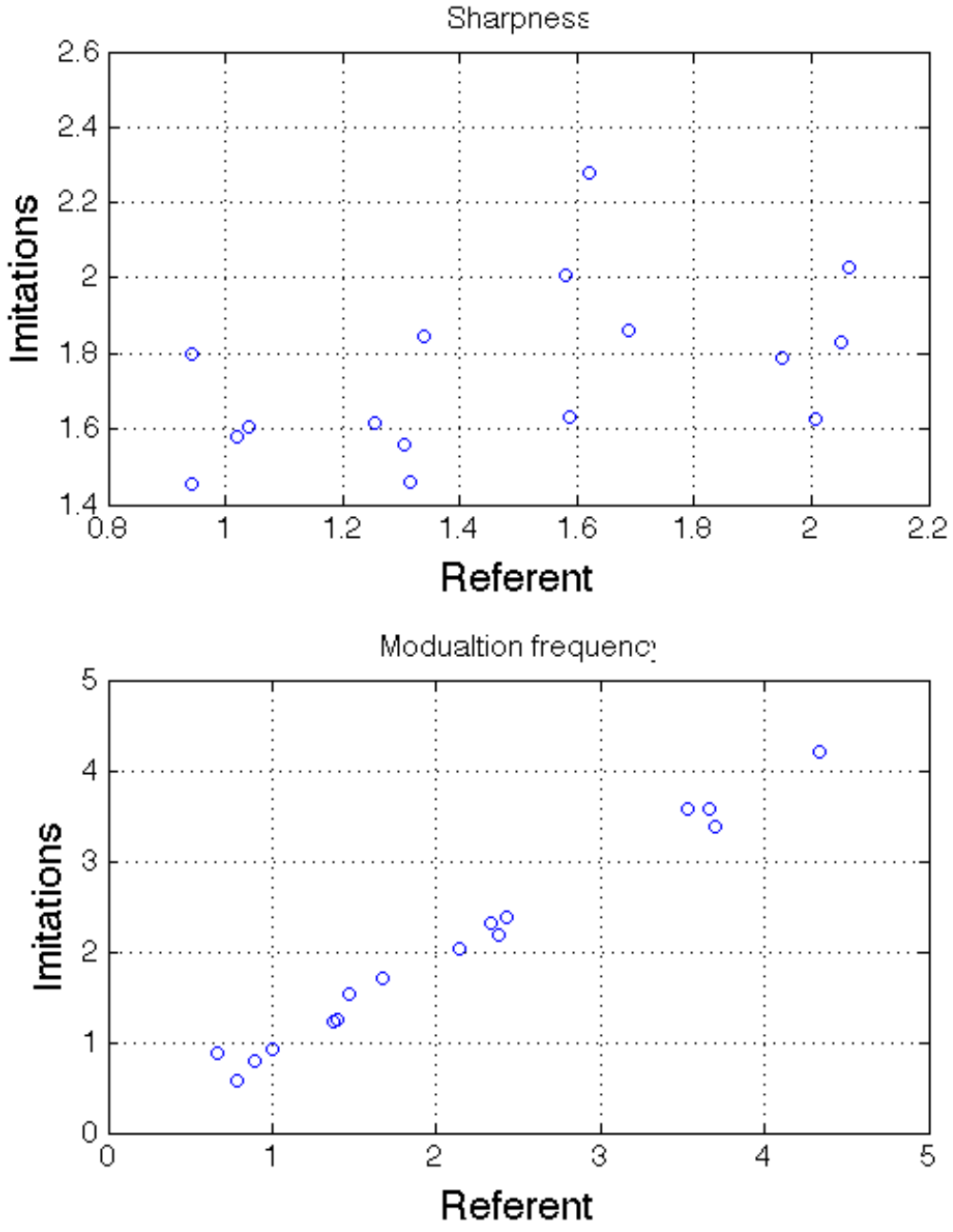

FIGURE 3.11 – La relation entre les sons originaux et les imitations au niveau de (l'estimation de) la fréquence de modulation et l'acuité.

Chapitre 4

Conclusion et Perspectives

Jusqu'à présent, nous avons créé un corpus homogène à deux dimensions perceptives, l'acuité et la force de fluctuation. Nous avons ensuite vérifié notre choix de corpus par une expérience de dissemblance analysée à l'aide d'une technique d'analyse multidimensionnelle de proximités et nous avons observé que l'espace perceptif issu de cette expérience correspond à l'espace de synthèse. Dans l'étape suivante, nous avons enregistré les imitations de deux experts en performances vocales en espérant y retrouver les mêmes dimensions perceptives. Pour cela, nous avons calculé les corrélations entre les enregistrements et les descripteurs acoustiques. Les résultats nous ont montré que l'aspect rythmique des sons a été transmis mais la dimension spectrale (l'acuité) n'a pas été directement incorporée dans les imitations. Il est en effet possible qu'une ou plusieurs dimensions du timbre des imitations correspondent à l'acuité. Si l'acuité n'était pas du tout communiquée par les imitations, les taux de reconnaissance devraient être autour de 50 %. Considérant les taux de reconnaissance que nous avons obtenus (supérieur à 50 %), nous pouvons dire qu'une partie de l'acuité est présente dans les imitations. L'analyse des corrélations des dimensions de l'espace MDS et les descripteurs acoustiques suggèrent que l'acuité est imitée par moyen d'autres dimensions spectrales. Il nous apparait que la dimension de l'acuité n'est pas transmise par imitation mais elle est rattrapée par une mesure de balance d'énergie de spectre qui différencie les sons plus graves des son plus aigus.

Pour la suite de ce stage, nous allons répondre à cette question, ≪Pourquoi les locuteur n'ont pas réussi à imiter l'acuité ? bien qu'il apparaisse relativement aisé , pour de tels experts, de produire des résonances du conduit vocal à différentes fréquences≫. Nous avons plusieurs hypothèses : ça pourrait être le rythme qui écrase d'autre dimension. Pour vérifier, nous pouvons fixer le rythme et refaire l'expérience. Aussi, ça pourrait être la protocole de l'expérience qui ne permet

pas aux locuteurs de comparer leurs imitations. Nous pouvons également nous intéresser qu'aux sujets bien corrélés entre eux (la zone claire de la figure 3.10). Pour cela, il nous fera de supprimer les sujets hors de cette zone et de mener l'expérience avec plus de sujet pour avoir assez de sujets dans la même "groupe".

Annexes

Annexe A

Interfaces graphiques des expériences

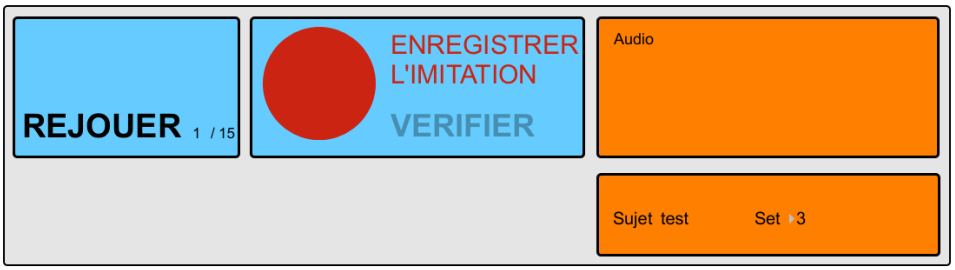

FIGURE A.1 – L'interface graphique d'enregistrement d'imitations, conçue sous MAX/MSP.

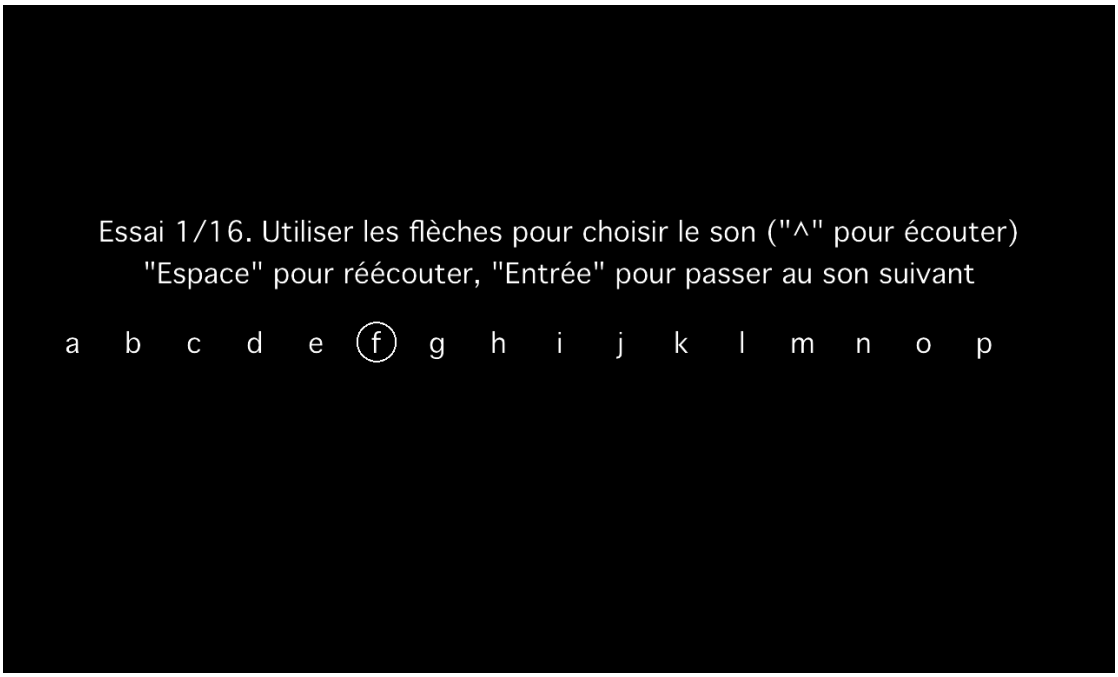

FIGURE A.2 – L'interface graphique d'enregistrement d'imitations, conçue sous Psychotoolbox de MATLAB.

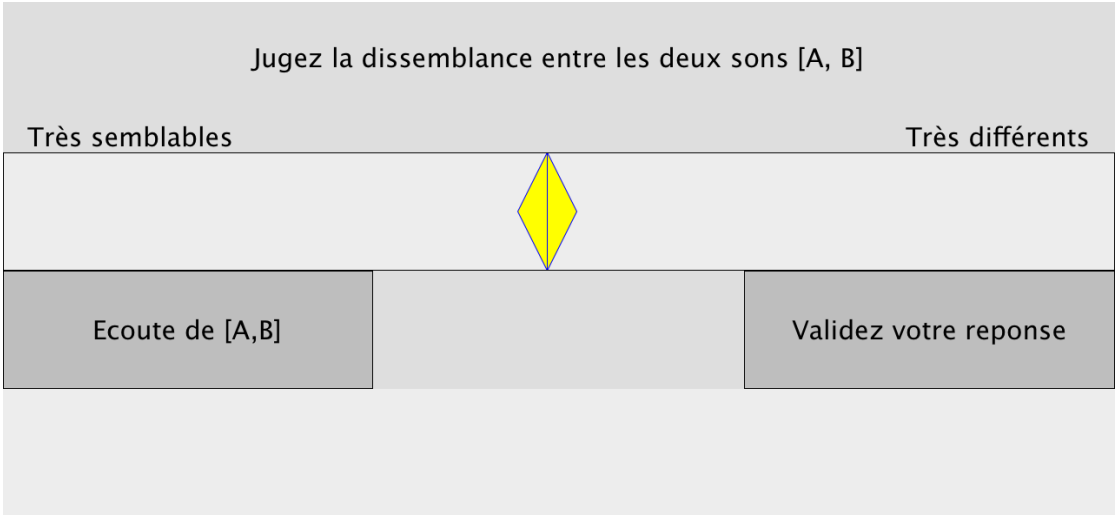

FIGURE A.3 – L'interface graphique du test de dissemblance, conçue sous PsiExp.

Annexe B

Paramètres de la synthèse du corpus

Type du filtre	passe-bande Butterworth
Ordre du filtre	2
Profondeur de modulation	1
Durée d'attaque	10 ms
Durée de déclin	10 ms
Niveau de maintien	1
Durée de relâchement	0 ms

TABLE B.1 – Paramètres des sons du corpus synthétisé

#	Fréquence de modulation	Fréquence centrale	Largeur de bande
1	0,70	363,07	100
2	0,79	1116,07	223,21
3	0,83	596,99	119,40
4	0,87	2000,93	400,19
5	1,30	1069,09	213,82
6	1,33	706,83	141,37
7	1,39	2026,46	405,29
8	1,56	296,74	100
9	2,06	1758,98	351,80
10	2,21	1059,09	211,82
11	2,30	674,80	134,96
12	2,35	383,56	100
13	3,46	1244,06	248,81
14	3,58	1895,20	379,04
15	3,64	294,88	100
16	4,26	664,98	133

TABLE B.2 – Paramètres des sons du corpus synthétisé

Annexe C

Espaces MDS des imitation du locuteur A

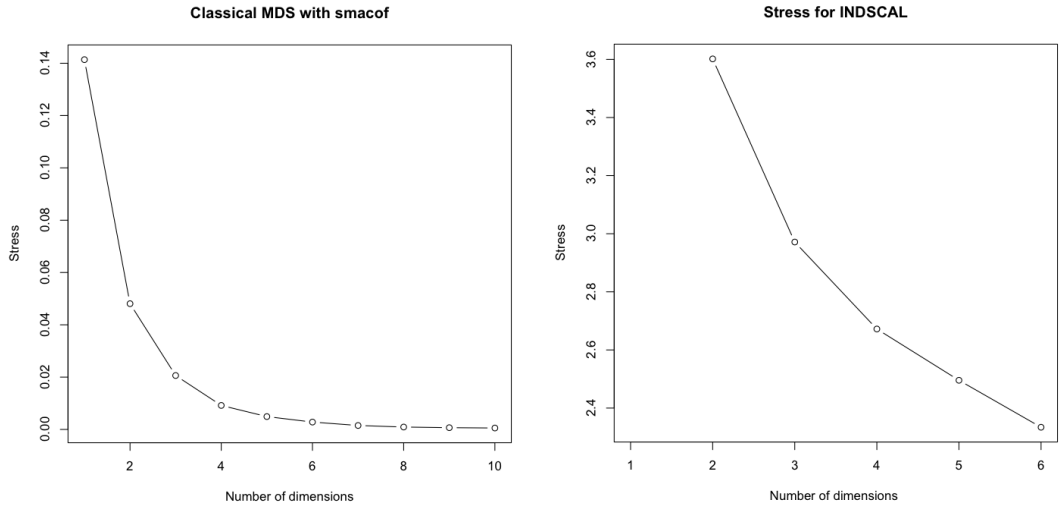

FIGURE C.1 – Le rapport entre le stress et le nombre de dimensions pour le modèle MDS classique (avec algorithme smacof de *R*) et pour le modèle INDSCAL.

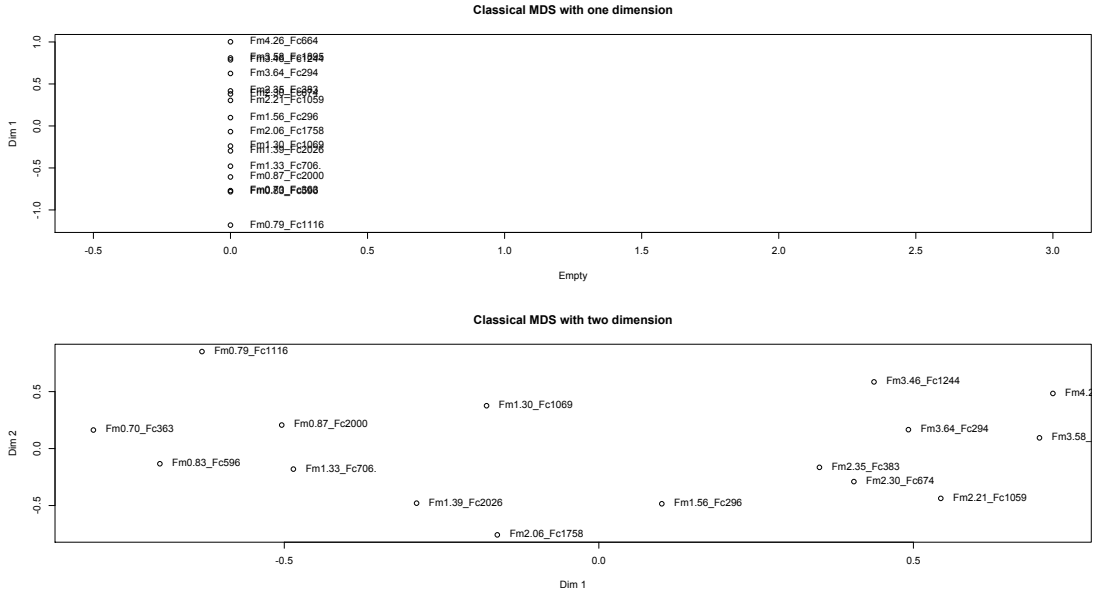

FIGURE C.2 – Les espaces des imitations du locuteur A en utilisant le modèle classique de MDS à 1 et 2 dimensions.

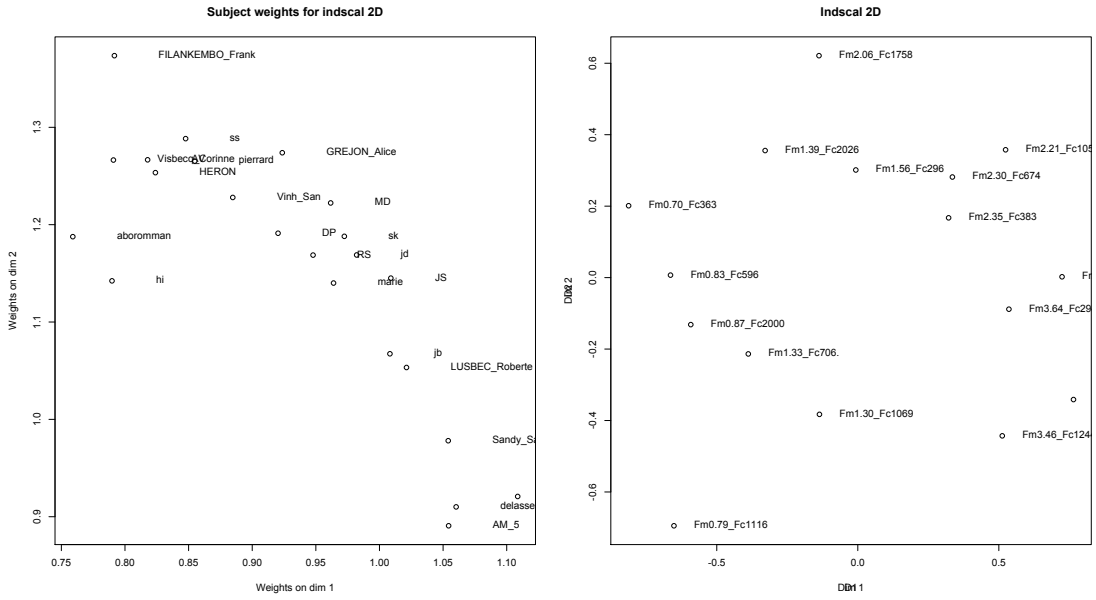

FIGURE C.3 – A gauche : la ponderation des sujets de l'expérience MDS des imitations du locuteur A. A droite : l'espace MDS obtenu en utilisant INDSCAL à 2 dimensions.

Annexe D

Corrélations des dimensions des espaces MDS

Modèle MDS	Dimension	Descripteur	Correlation
MDS classique 1D	Dim. 1	Estimation of modulation frequency	0,95
		MIR tempo	0,94
INDSCAL 2D	Dim. 1	MIR tempo	0,95
		Estimation of modulation frequency	0,94
	Dim. 2	Perceptual Tristimulus Loudness Weighted Mean var7	0,88
		Perceptual Tristimulus Loudness Weighted Mean var9	0,87
INDSCAL 3D	Dim. 1	MIR tempo	0,95
		Estimation of modulation frequency	0,94
	Dim. 2	Spectral Spread Loudness Weighted Mean var6	0,83
		DDMFCC Loudness Weighted Mean var5	0,83
	Dim. 3	MFCC Loudness Weighted Mean var4	0,84
		Signal AutoCorrelation Loudness Weighted Mean var6	0,82
INDSCAL 4D	Dim. 1	MIR tempo	0,97
		Estimation of modulation frequency	0,97
	Dim. 2	Spectral Spread Loudness Weighted Mean var6	0,84
		DDMFCC Loudness Weighted Mean var5	0,80
	Dim. 3	MIR mfcc1	0,76
		Signal AutoCorrelation LoudnessWeighted Standard Deviation var1	0,75
	Dim. 4	Perceptual Tristimulus Loudness Weighted Mean var4	0,70
		Perceptual Tristimulus Loudness Weighted Mean var1	0,68
INDSCAL 5D	Dim. 1	MIR tempo	0,98
		Estimation of modulation frequency	0,97
	Dim. 2	DDMFCC Loudness Weighted Mean var5	0,86
		Spectral Spread Loudness Weighted Mean var6	0,84
	Dim. 3	Perceptual Tristimulus Loudness Weighted Mean var4	0,74
		Perceptual Tristimulus Loudness Weighted Mean var1	0,72
	Dim. 4	Spectral Centroid Loudness Weighted Standard Deviation var2	0,77
		Signal AutoCorrelation LoudnessWeighted Standard Deviation var1	0,77
	Dim. 5	Spectral Crest Loudness Weighted Mean var3	0,72
		Perceptual Spectral Skewness Loudness Weighted Standard Deviation var2	0,65

TABLE D.1 – Correlations entre les dimensions des espaces MDS obtenu et les descripteurs psychoacoustiques

Bibliographie

[1] Vanderveer, N. J., *Ecological acoustics : human perception of environmental sounds*. in Structure and Perception of electroacoustic Sound and Music Elsevier, Amsterdam, 1989.

[2] Krumhansl, C. L., *Why is musical timbre so hard to understand*. Cornell University, 1989.

[3] Hajda, J. M., *Methodilogical issues in timbre research*. Perception and cognition of Music, 1997.

[4] Kendall, R. A., et Carterette, E. C., *Verbal attributes of simultaneous wind instruments timbres* . Music percept. 10, 1994.

[5] Grey, J. M., *Multidimensional perceptual scaling of musical timbres*. Standford University, 1977.

[6] Plomp, R., *Timbre as a multidimensional attribute of complex tones*. in frequency Analysis and Periodicity Detection in Hearing, 1970.

[7] Torgerson, W. S., *Theory and Methods of Scaling*. Wiley, 1958.

[8] Gower, J. C., *Some Distance Properties of Latent Root and Vector Methods using Multivariate Analysis*. Biometrika,53, 1966.

[9] Carroll, J. D., Chang J. J., *Analysis of Individual Differences in Multidimensional Scaling via an n-way generalization of Eckart-Young Decomposition* . Psychometrika,35, 1970.

[10] Winsberg, S., De Soete, G., *A latent Class approach to fitting the weighted Euclidean Model, CLASCAL*. Psychometrika,58, 1993.

[11] Hope, A. C., *A simplified Monte Carlo significance test procedure*. Journal of the Royal Statistical Society, 1968.

[12] Aitken, M. et al., *Statistical Model of Data in Teaching Styles*. Journal of the Royal Statistical Society, 1981.

[13] Miller, J. R., Carterette, E. C., *Perceptual Space for musical Structures*. Journal of the Acoustical Society of America, 1975.

[14] McAdams, S. et al., *Perceptual Scaling of Synthesized Musical Timbres*. Psychological Research 58, 1995.

[15] Marozeau, J. et al., *The dependency of timbre on fundamental frequency*. The Journal of the Acoustical Society of America vol. 114, no. 5, 1995.

[16] Caclin, A. et al., *Acoustic correlates of timbre space dimensions : A confirmatory study using synthetic tones* . The Journal of the Acoustical Society of America vol. 118, 2005.

[17] Lemaitre, G., Houix, O., Misdariis, N., et Susini, P., *Listener expertise and sound identification influence the categorization of environmental sounds* . J. Exp. Psychol., 2010.

[18] Hashimoto, T. et al. *The neural mechanism associated with the processing of onomatopoeic sounds*. Neuroimage 31, 2006.

[19] Iwasaki, N. et al., *What do English speakers know about gera-gera and yota-yota ? A cross linguistic investigation of mimetic words for laughing and walking*. Jpn. Lang. Educ. Globe 17, 2007.

[20] Oswalt, R. L., *inanimate imitatives*. in Sound Symbolism, Cambridge University Press, 1994.

[21] Gygi, B. et al., *Spectral-temporal factors in the identification of environmental sounds*. The Journal of the Acoustical Society of America vol. 115, 2004.

[22] Strange, W., Shafer, V., *Speach perception in second language learners : The reeducation of selective perception*. in Phonology and second language acquisition, 2008.

[23] Lemaitre, G. et Rocchesso, D., *On the effectiveness of vocal imitation and verbal descriptions of sounds*. The Journal of the Acoustical Society of America vol. 135, 2013.

[24] Lemaitre, G. et Hellar, L. M., *Auditory perception of material is fragile, while action is strikingly robust*. The Journal of the Acoustical Society of America vol. 131, 2012.

[25] Lemaitre, G., Susini, P., et al., *The sound quality of car horns : A psychoacoustical study of timbre*. Acustica, vol. 93, 2007.

[26] Lakatos, S. et al., *The representation of auditory source characteristics : Simple geometric form*. Perception and Psychophysics 59, 1997.

[27] Susini, P. et al., *A multidimensional technoque for Sound Quality Assessment*. Acustica vol. 85, 1999.

[28] Misdariis, N. et al., *Environmental sound perception : Metadescription and modeling based on independent primary studies*. EURASIP Journal on Audio, Speech and Music Processing, 2010.

[29] Brainard, D. H., *The Psychophysics Toolbox*. Spatial Vision 10 :433-436, 1997.

[30] Leeuw, J. & Mair, P., *Multidimensional scaling using majorization : The R package smacof*. Journal of Statistical Software, 31(3), 1-30, 2009. http ://www.jstatsoft.org/v31/i03/

[31] Claire Churchill, *Expression for weighting function obtained by fitting an equation to data given in 'Psychoacoustics : Facts and Models'*. 1991.

[32] Alois Sontacchi, *Entwicklung eines Modulkonzeptes für die psychoakustische Geräuschenalayse unter MatLab*. Diplomarbeit, Institut für Elektronische Musik der Kunstuniversität Graz, Graz, Austria, 1999. implémenté dans le logiciel LEA de GENESIS (www.genesis-acoustics.com).

[33] Zwicker, E., Fastl, H., *Program for calculating loudness according to DIN 45 631 (ISO 532B)*. The Journal of the Acoustical Society of America, 1991.

[34] Alain de Cheveigné et Hideki Kawahara, *YIN, a fundamental frequency estimator for speech and music*. The Journal of the Acoustical Society of America vol. 111, 2002.

[35] Geoffroy Peeters et al., *The Timbre Toolbox : Extracting audio descriptors from musical signals*. The Journal of the Acoustical Society of America vol. 130, 2011.

[36] Paul Boersma et David Weenink, *Praat : doing phonetics by computer [Computer program]*. Version 5.3.51, retrieved 2 June 2013 from http ://www.praat.org/, 2013.