Université Claude Bernard Lyon 1 – Université de Lyon

École doctorale Neurosciences et Cognition (ED NSCo)

Thèse

en vue de l'obtention du titre de

Docteur en Neurosciences

Identification des indices acoustiques utilisés lors de la compréhension de la parole dégradée

Présentée et soutenue publiquement le 18/11/2015 par

Léo VARNET

Dirigée par

Michel HOEN et Fanny MEUNIER

Jury:

Willy SERNICLAES
Christian LORENZI
Jean-Luc SCHWARTZ
Kenneth KNOBLAUCH
Éric TRUY
Fanny MEUNIER

Thèse préparée au sein du Laboratoire sur le Langage, le Cerveau et la Cognition (L2C2),

Institut des Sciences Cognitives (ISC), 67 Bd Pinel - 69675 - BRON

À mes parents

Résumé / abstract

Identification des indices acoustiques utilisés lors de la compréhension de la parole dégradée

Bien qu'il existe un large consensus de la communauté scientifique quant au rôle des indices acoustiques dans la compréhension de la parole, les mécanismes exacts permettant la transformation d'un flux acoustique continu en unités linguistiques élémentaires demeurent aujourd'hui largement méconnus. Ceci est en partie dû à l'absence d'une méthodologie efficace pour l'identification et la caractérisation des primitives auditives de la parole. Depuis les premières études de l'interface acoustico-phonétique par les Haskins Laboratories dans les années 50, différentes approches ont été proposées ; cependant, toutes sont fondamentalement limitées par l'artificialité des stimuli utilisés, les contraintes du protocole expérimental et le poids des connaissances a priori nécessaires. Le présent travail de thèse s'est intéressé à la mise en œuvre d'une nouvelle méthode tirant parti de la situation de compréhension de parole dégradée pour mettre en évidence les indices acoustiques utilisés par l'auditeur.

Dans un premier temps, nous nous sommes appuyés sur la littérature dans le domaine visuel en adaptant la méthode des Images de Classification à une tâche auditive de catégorisation de phonèmes dans le bruit. En reliant la réponse de l'auditeur à chaque essai à la configuration précise du bruit lors de cet essai, au moyen d'un Modèle Linéaire Généralisé, il est possible d'estimer le poids des différentes régions temps-fréquence dans la décision. Nous avons illustré l'efficacité de notre méthode, appelée Image de Classification Auditive, à travers deux exemples : une catégorisation /aba/-/ada/, et une catégorisation /da/-/ga/ en contexte /al/ ou /aʁ/. Notre analyse a confirmé l'implication des attaques des formants F2 et F3, déjà suggérée par de précédentes études, mais a également permis de révéler des indices inattendus. Dans un second temps, nous avons employé cette technique pour comparer les résultats de participants musiciens experts (N=19) ou dyslexiques (N=18) avec ceux de participants contrôles. Ceci nous a permis d'étudier les spécificités des stratégies d'écoute de ces différents groupes.

L'ensemble des résultats suggèrent que les Images de Classification Auditives pourraient constituer une nouvelle approche, plus précise et plus naturelle, pour explorer et décrire les mécanismes à l'œuvre au niveau de l'interface acoustico-phonétique.

Mots-clefs : Images de Classification (CI) ; Images de Classification Auditives (ACI) ; Modèle Linéaire Généralisé (GLM) ; Catégorisation de phonème ; Indices acoustiques ; Parole dans le bruit

Identification of acoustic cues involved in degraded speech comprehension

There is today a broad consensus in the scientific community regarding the involvement of acoustic cues in speech perception. Up to now, however, the precise mechanisms underlying the transformation from continuous acoustic stream into discrete linguistic units remain largely undetermined. This is partly due to the lack of an effective method for identifying and characterizing the auditory primitives of speech. Since the earliest studies on the acoustic–phonetic interface by the Haskins Laboratories in the 50's, a number of approaches have been proposed; they are nevertheless inherently limited by the non-naturalness of the stimuli used, the constraints of the experimental apparatus, and the a priori knowledge needed. The present thesis aimed at introducing a new method capitalizing on the speech-in-noise situation for revealing the acoustic cues used by the listeners.

As a first step, we adapted the Classification Image technique, developed in the visual domain, to a phoneme categorization task in noise. The technique relies on a Generalized Linear Model to link each participant's response to the specific configuration of noise, on a trial-by-trail basis, thereby estimating the perceptual weighting of the different time-frequency regions for the decision. We illustrated the effectiveness of our Auditory Classification Image method through 2 examples: a /aba/-/ada/ categorization and a /da/-/ga/ categorization in context /al/ or /aʁ/. Our analysis confirmed that the F2 and F3 onsets were crucial for the tasks, as suggested in previous studies, but also revealed unexpected cues. In a second step, we relied on this new method to compare the results of musical experts (N=19) or dyslexics participants (N=18) to those of controls. This enabled us to explore the specificities of each group's listening strategies.

All the results taken together show that the Auditory Classification Image method may be a more precise and more straightforward approach to investigate the mechanisms at work at the acoustic-phonetic interface.

Keywords: Classification images (CI); Auditory Classification Images (ACI); Generalized Linear Model (GLM); Phoneme categorization; Acoustic cues; Speech-in-noise

Remerciements

Bien que la préparation d'une thèse soit réputée un exercice solitaire, ce travail n'aurait pas pu voir le jour sans la collaboration, l'aide et le soutien de nombreuses personnes que je tiens à remercier ici en premier lieu :

Merci tout d'abord à Fanny et Michel qui m'ont fait confiance dès le début, lorsque j'ai toqué à leur porte pour un stage, il y a 5 ans, qui m'ont ensuite offert la possibilité de rester dans l'équipe et qui m'ont finalement encadré tout au long de ces 3 années de thèse. Vous m'avez fait découvrir non seulement les rouages complexes de la perception de la parole mais aussi (surtout ?) les rouages complexes de la Recherche. Vous avez su être présents dans les difficultés – scientifiques, humaines ou administratives – de la thèse tout en me laissant mener ce projet comme je l'entendais, et partager votre inégalable enthousiasme et votre goût communicatif de la recherche (particulièrement bienvenus dans mes moments de doute).

Merci aux membres du jury pour l'intérêt qu'ils ont porté à mon travail, et pour le temps qu'ils lui ont consacré et, particulièrement, à Christian Lorenzi et Willy Serniclaes pour avoir accepté d'être rapporteurs de cette thèse. Merci également à Ken Knoblauch et Willy Serniclaes pour votre collaboration directe aux articles, en tant que co-auteurs, et pour les discussions statistiques et linguistiques qui ont beaucoup contribué à ma réflexion.

Merci aux participants des 4 études, et des nombreux tests préliminaires, qui ont ~~subi~~ écouté des heures durant mes stimuli de parole bruités. Merci, notamment, à mes proches qui, au péril de leur santé mentale, ont su sacrifier leur temps et leurs oreilles – parfois même à plusieurs reprises – pour le progrès de la psycholinguistique : Mégane, Charlotte, Gwen, Sam, Chloé, Michel, Mimi, Migou et Marie (prix spécial « Wonderwoman »).

Un merci encore plus grand à trois collaboratrices d'exception, Gwen, Chloé et Tian Yun, qui ont eu la tâche encore plus pénible et répétitive d'organiser et d'encadrer une bonne partie de ces heures de passations et qui se sont néanmoins acquittées de ce travail avec une indéfectible bonne humeur.

Merci également à tous mes collègues du L2C2 et de l'ISC, pour leur aide, et tout particulièrement à Tatiana et Yves pour nous avoir « hébergés » dans leur labo avant que nous n'en devenions officiellement membres. Merci à tous les doctorants et masters qui sont passés ici durant ces trois ans (que je n'énumèrerai pas ici de peur d'oublier des noms) et à tous les doctorants de notre groupe Etu-CRNL (et spécialement au « noyau dur » de l'équipe). Toutes ces rencontres ont été extrêmement enrichissantes scientifiquement et humainement.

Merci aux étudiants de Polytech Lyon que j'ai eu le plaisir d'encadrer durant ces trois années pour m'avoir donné l'occasion de faire mes premières armes dans l'enseignement et pour m'avoir obligé à trouver les mots pour leur transmettre les bases de la programmation, des statistiques et du traitement du signal.

Un merci tout particulier à Marie, ma co-bureau et amie, alliée dans l'adversité de la thèse et soutien indispensable quand tout va mal, qui a su m'obliger à prendre des pauses-café, supporter l'omniprésence de Matlab, partager son amour de la chanson québécoise (et autres styles de musiques moins avouables), décoder les signes sociaux, et maintenir une folle ambiance dans le bureau n°260).

Merci à mes parents pour leur soutien et leur amour, pour m'avoir donné le goût des sciences mais aussi – et surtout – du reste, et pour la relecture de la thèse dans la dernière ligne droite.

Merci à tous mes amis, amours, cousins, de Lyon et d'ailleurs, toujours prêts à aller boire une grosse bière ou à jouer de la musique pour me faire perdre mon précieux temps de travail ou à me demander « Alors, ça avance cette thèse ? »… Et, surtout, pour m'avoir supporté (dans tous les sens du terme) au quotidien.

Et enfin, merci chaton.

Table des matières

Index

Acronymes :

2AFC (2 Alternatives Forced Choice) : Choix forcé à 2 alternatives

3DDS (3-Dimensional Deep Search)

ACI (Auditory Classification Image) : Image de classification auditive

AI (Articulation Index) : Indice d'articulation

BF (Best Frequency) : Fréquence préférée d'un champ récepteur

BR (Best Rate) : Taux préféré d'un champ récepteur

BS (Best Scale) : Modulation de fréquence préférée d'un champ récepteur

CI (Classification Image) : Image de classification

CV : Consonne + voyelle

f_0 : Fréquence fondamentale

Fx : Formant x

FDR (False Discovery Rate) : Taux de faux positifs

FWER (Family-Wise Error Rate) : Risque global d'erreur de type 1

GLM (Generalized Linear Model) : Modèle linéaire généralisé

LAFF (Lexical Access From Features)

LAFS (Lexical Access From Spectra)

MAP : Maximum A Posteriori

MSE (Mean Squared Error) : Erreur quadratique moyenne

SDT (Signal Detection Theory) : Théorie de la détection du signal

STMI (Spectro-Temporal Modulation Index) : Indice de modulation spectro-temporel

STRF (Spectro-Temporal Receptive Field) : Champ récepteur spectro-temporel

SNR (Signal to Noise Ratio) : Rapport signal sur bruit

VOT (Voice Onset Time) : Délai d'établissement du voisement

Conventions utilisées :

Phonétique

/·/ : transcription phonologique

[·] : transcription phonétique

Calcul matriciel

$\underline{x}$: vecteur de composantes x_i

$\underline{x}^T$: transposée du vecteur $\underline{x}$

$\underline{\underline{x}}$: matrice de composantes $x_{i,j}$

$\underline{\underline{x}} \cdot \underline{\underline{y}}$: produit matriciel

$\underline{\underline{x}} * \underline{\underline{y}}$: produit scalaire

Statistique

Pour X une variable aléatoire

E[X] : espérance mathématique de X

$V(X) = \sigma_X^2$: variance de X

$\hat{X}$: estimateur de X

Avant-propos

« On ne parle jamais autant de communication que dans une société qui ne sait plus communiquer avec elle-même ».

Lucien Sfez, *La Communication*, Presses Universitaires Françaises, 2010

L'évolution des centres d'intérêt de la Recherche est le reflet des changements de la société, sur le plan économique[1], politique[2] et, bien sûr, technologique. Ce constat est particulièrement vrai pour les Sciences de la parole dont l'essor coïncide avec le développement de la téléphonie (les premières découvertes dans le domaine sont dues aux chercheurs des *Bell Telephone Laboratories*[3]) et de certaines problématiques militaires de l'époque, notamment la cryptographie et la radiocommunication. L'essor du marché de la communication et l'implication financière ou technique d'acteurs de l'Industrie ou de la Défense[4] ont clairement orienté ces recherches vers les applications pratiques – que ce soit pour la reconnaissance ou la synthèse de la parole, la téléphonie mobile ou, encore, les prothèses auditives. C'est justement l'échec des tentatives d'automatisation[5] dans les années 50 à 70 qui constituera la rupture épistémologique nécessaire à une remise en question des premiers modèles devenus trop simplistes. L'héritage de cette origine reste néanmoins une conception « mécaniste » de la parole, fondement de la psycholinguistique : schémas émetteur-récepteur, entrée-sortie, stimulus-réponse… Ainsi la Théorie de l'Information de Shannon et la Théorie de la Détection du Signal, nées de préoccupations techniques sur la transmission optimale d'un message, invitèrent-elles à étendre l'analyse à tout système de signes. Conçue en ces termes, la perception de la parole est comparable à une conversion d'un signal acoustique en un flux de données linguistiques (soit une compression de 40000 bits/s à 40 bits/s environ)[6] au moyen d'un système de décodage binaire[7] très élaboré et structuré comme un programme informatique[8]. Dans ce cadre, l'activité du chercheur, démontant pièce par pièce une mécanique opérationnelle pour tenter d'en comprendre le fonctionnement, se rapproche de celle de l'ingénieur (*reverse-engineering*).

Ce lien entre chercheurs et ingénieurs est renforcé par la dépendance des premiers à un outillage sophistiqué, nécessitant souvent le recrutement de spécialistes, électroniciens, informaticiens ou acousticiens. Ainsi, d'après M. Grossetti et L.-J. Boë, la communauté de l'Étude de la parole est née de la rencontre entre les Sciences Humaines et la Recherche instrumentale[9], ce qui est perceptible dans la composition des équipes de recherche[10] et l'interdisciplinarité des conférences organisées[11]. L'innovation technique a conditionné les principales avancées théoriques dans le domaine, que ce soit au niveau de la visualisation du signal acoustique[12] (p.ex. le spectrographe qui a facilité l'exploration du contenu des productions de parole), de la synthèse vocale[13] (p.ex. le *Pattern Playback* qui a permis une analyse plus systématique de la perception), ou encore des méthodes d'imagerie (p.ex. l'électromyographie pour l'étude du

fonctionnement de l'appareil phonatoire). De plus, de manière générale, les Sciences Cognitives sont dépendantes des méthodes de mesure (chronométrie, occulométrie, potentiels évoqués…), qui ont défini les contours des différentes disciplines[14], et des méthodes statistiques employées pour analyser ces données[15].

Le présent manuscrit s'inscrit dans cette démarche mêlant l'approche de l'ingénieur et celle du chercheur. Pour répondre à une question fondamentale concernant la perception de la parole, l'accent est porté ici sur la mise au point d'un nouvel outil permettant de révéler les traitements effectués par un auditeur. D'après T. Shinn, l'un des principes de la recherche instrumentale est la « généricité », c'est-à-dire la possibilité de transposition depuis un domaine vers un autre[16]. Ainsi, dans le cadre de ce travail de thèse, nous adapterons une technique déjà répandue dans la communauté visuelle à une problématique propre à la modalité auditive. La méthode des Images de Classification, rencontrée au détour d'un article sur la perception des contours illusoires[17], a constitué notre point de départ et notre fil conducteur pour aborder la question des indices acoustiques impliqués dans la compréhension de la parole dégradée.

[1] (Clément et al., 2013)
[2] (Godement, 1992)
[3] Parmi lesquels notamment Harvey Fletcher (Allen, 1996). Son article « The nature of speech and its interpretation » constitue un bel exemple de travail mêlant phonétique, électronique et psychoacoustique (Fletcher, 1922).
[4] IBM, la DGA ou le *United States Department of Defense*, le CNET, Microsoft ou Google, pour ne citer que les plus influents.
[5] Échecs de la machine à lire (Cooper et al., 1984; Shankweiler & Fowler, 2015), de la synthèse par concaténation (Boë, 1997) et du projet ARPA SUR pour la reconnaissance automatique de la parole continue (Boë & Liénard, 1988)
[6] (Liberman et al., 1972)
[7] Fondé sur la présence ou l'absence de certains indices acoustiques (Jakobson, 1961; Nève, 2002)
[8] (James E. Cutting, 1978)
[9] (Grossetti & Boë, 2008). « Les sciences phonétiques sont ainsi un reflet fidèle et presque synchrone des mutations technologiques qui ont fait passer l'instrumentation de la mécanique à l'électricité, à l'électronique et à l'informatique » (Boë, 1997)
[10] Pour citer deux exemples célèbres, le trio Moris Halle (phonologue) - Roman Jacobson (linguiste) - Gunnar Fant (électronicien) au MIT, et le trio Pierre Delattre (phonéticien) - Alvin Liberman (psychologue) - Francis Cooper (électronicien) aux laboratoires Haskins.
[11] En particulier, les Journée d'Étude de la Parole (JEP) et Interspeech, deux colloques majeurs dans la construction de la communauté, font se rencontrer des chercheurs d'horizons très divers, depuis les neurosciences jusqu'à la reconnaissance automatique en passant par l'analyse de scènes multimodales.
[12] (Chafcouloff, 2004)
[13] (Peterfalvi, 1966)
[14] (Chamak, 2011; Plas, 2011)
[15] Ainsi, d'après G. Tiberghien et M. Jeannerod, « identifier un phénomène cognitif c'est inventer un indicateur de ce phénomène » (Tiberghien & Jeannerod, 1995)
[16] (Shinn, 2000)
[17] (Gold et al., 2000)

Partie théorique

« Je sais que la langue est à l'intérieur du monde et que, simultanément, le monde est dans la langue. Je sais que nous sommes à la lisière de la langue et du monde. »

Patrick Dubost, *Œuvres Poétiques (tome 2)*, éditions La Rumeur Libre, 2013

1. Introduction

Le son de parole possède un statut particulier parmi l'ensemble des sons qui parviennent à l'oreille humaine : il est produit dans le but d'être perçu. Alors que les bruits de l'environnement nous renseignent indirectement sur le monde qui nous entoure, le langage parlé, lui, naît d'une démarche active. Dans la quasi-totalité des cas, il est émis à l'intention d'une personne ou d'un ensemble de personnes, que ce soit pour transmettre une information, pour agir sur l'auditeur ou, même, dans une fonction purement sociale.

Dans une conversation face-à-face, la communication parlée pourrait être envisagée comme un système primitif de télécommunication (étymologiquement : « communication à distance ») à très courte portée. Ici, la transmission depuis la bouche du locuteur jusqu'au tympan de l'auditeur s'opère par l'intermédiaire des vibrations de l'air ambiant. Comme en radiotéléphonie, l'émetteur doit traduire le message sur un support adapté (par exemple les ondes radioélectriques pour la voie hertzienne) afin qu'il soit ensuite transmis puis décrypté par le récepteur. La parole est donc un code destiné à véhiculer l'information de manière efficace à travers l'espace. Elle offre un support physique sensible aux représentations mentales et permet ainsi leur diffusion. Cette transmission suppose nécessairement un partage de certaines connaissances entre les interlocuteurs : une entente tacite sur le sens des concepts mis en jeu, un lexique et une grammaire communs et, plus fondamentalement, un même système de codage-décodage associant un mot à sa forme acoustique.

Au niveau de l'auditeur, la reconnaissance d'un mot prononcé dépend de la reconnaissance des sons qui le composent, un processus réalisé extrêmement rapidement et sans effort par le système auditif. Bien que les mécanismes précis mis en jeu ne soient pas encore pleinement compris, plusieurs étapes ont d'ores et déjà été identifiées (voir schéma Figure 1). Le signal acoustique continu doit tout d'abord être découpé en petites unités discrètes, la syllabe ou le phonème (phase de segmentation). Dans un deuxième temps, ces unités sont décodées à partir de leurs caractéristiques acoustiques (phase de catégorisation). Enfin, ces briques linguistiques élémentaires sont combinées pour former des mots qui activeront in fine les représentations mentales correspondantes dans la mémoire à long terme (phase d'accès lexical). Ce modèle est relativement rudimentaire mais il a néanmoins l'avantage de donner un aperçu réaliste du fonctionnement global du système.

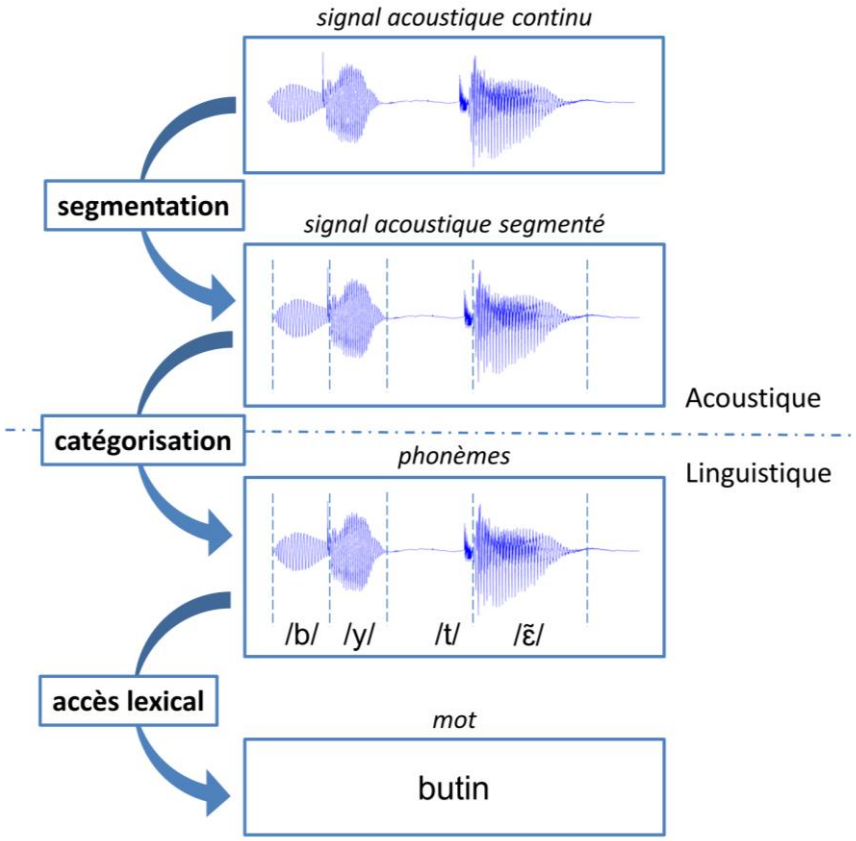


Figure 1 – Schéma des étapes de la reconnaissance d'un son de parole. Les noms des différentes étapes de traitement sont indiqués en gras et la nature des éléments manipulés en italique. Le trait discontinu marque la position de la frontière acoustico-phonétique.

Dans cette thèse nous nous intéresserons spécifiquement à la deuxième étape de ce traitement. Cette phase de catégorisation est critique puisque c'est à ce niveau que s'effectue la conversion du signal acoustique en une information linguistique d'où son nom d' « interface acoustico-phonétique ». Le problème central qui se pose à ce niveau est celui des primitives perceptuelles de la parole : sur quelles informations acoustiques (ou « indices acoustiques ») contenues dans le signal de parole l'auditeur s'appuie-t-il pour reconnaître les phonèmes, les syllabes ou les mots ? Selon quel ensemble de règles l'auditeur associe-t-il une représentation physique de surface, extrêmement variable, à une représentation phonologique sous-jacente, abstraite et invariante ? Malgré la simplicité apparente de la question, l'identification des indices acoustiques est un problème ardu, comme en témoigne la longévité de ce champ de recherche entamé dans les années 50. À l'heure actuelle, les mécanismes à l'œuvre au niveau de l'interface acoustico-phonétique demeurent en grande partie inconnus.

Dans une situation d'écoute réelle, le son de parole reçu par l'auditeur reflète non seulement le message transmis mais également l'environnement acoustique au sein duquel ce signal transite. Souvent, la communication est perturbée par l'addition de bruit (p.ex. fond sonore ambiant, « friture » lors d'une communication téléphonique,

conversations alentours, etc…) ou déformée (réverbérations et résonances de la pièce, bande passante du téléphone…). Pourtant, le système auditif assure l'intelligibilité du message avec une robustesse et une fiabilité qui restent jusqu'à présent inégalées par les systèmes de reconnaissance vocale. Notre thèse portera un intérêt tout particulier à cette situation de parole dégradée non seulement en raison de son caractère naturel mais, aussi, comme un moyen d'appréhender les mécanismes à l'œuvre lors de la compréhension de la parole. En effet, la mise en difficulté du système auditif permet d'examiner ses erreurs récurrentes qui sont le reflet des traitements effectués.

1.1. Rappels anatomiques sur la production de la parole

Comme dans un instrument à vent, la production d'un son de parole est permise par l'excitation d'une colonne d'air couplée avec un résonateur. Il n'existe pas d' « organe de la phonation » unique mais un ensemble d'organes qui coopèrent pour former des sons (voir Figure 2).

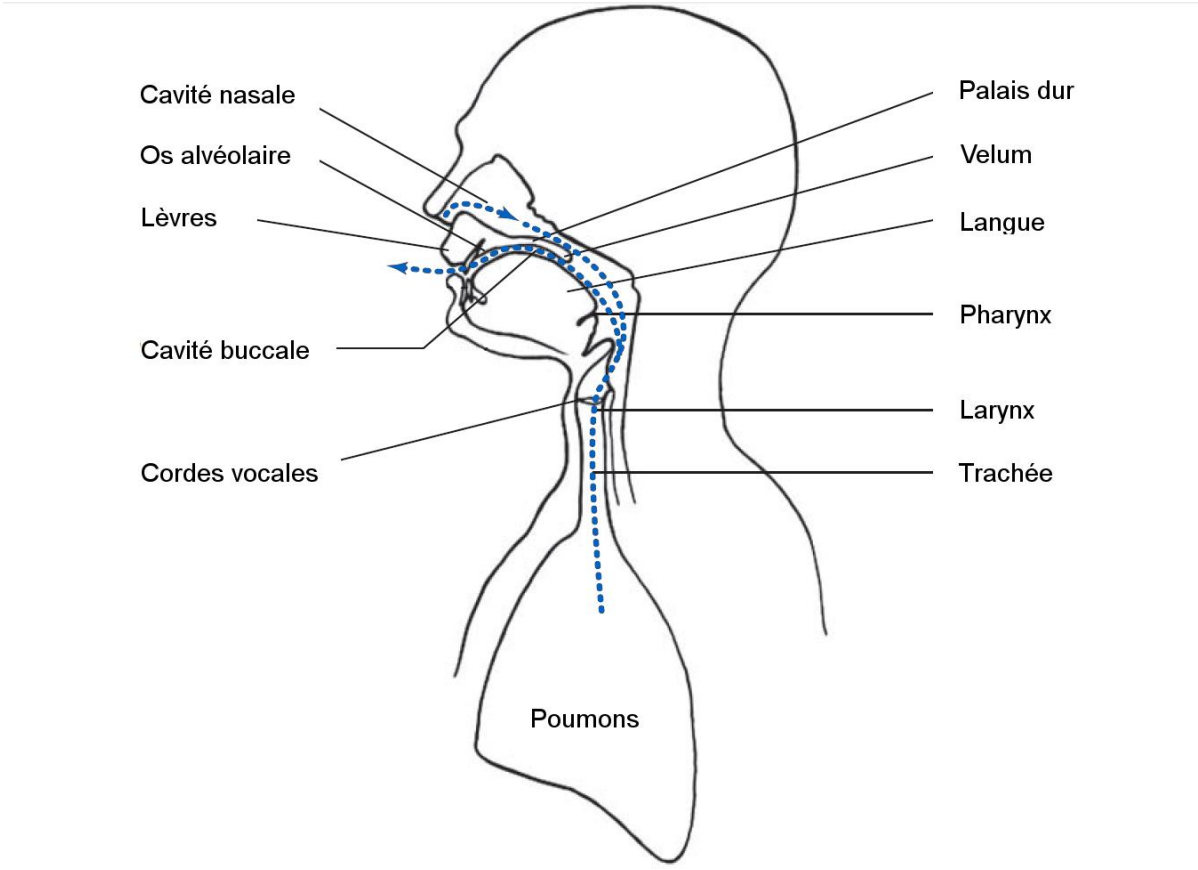


Figure 2 – Illustration de l'appareil phonatoire et de ses différents éléments. Le trajet de l'air est indiqué en bleu. Adapté de (Singh & Singh, 2005).

L'appareil phonatoire se divise fonctionnellement en 3 groupes d'organes :

- L'ensemble pulmonaire (poumons, diaphragme, trachée…) qui assure la création d'un flux d'air, contrôlé par l'expiration.

- L'ensemble laryngé (cordes vocales, musculature laryngée…) transformant l'énergie du souffle en provenance des poumons en un son audible. Les cordes vocales entrent en vibration au passage de l'air, à une fréquence qui définit la fréquence fondamentale de la voix (f_0). Celle-ci dépend de la taille des cordes vocales (et donc du sexe et de la physiologie du locuteur), mais elle est également modulable au moyen des muscles du larynx.

- L'ensemble supra-laryngé (pharynx, langue, palais, lèvres…) dont l'action permet de générer les caractéristiques particulières des sons du langage. Les cavités (orale et parfois nasale) jouent ici le rôle de résonateur, enrichissant harmoniquement la vibration primitive issue du larynx. Leur configuration particulière à un instant donné détermine les résonnances du son, c'est-à-dire son timbre. Les muscles articulateurs (langue, palais mou ou velum, lèvres, mâchoire) se coordonnent pour contrôler précisément la forme du conduit vocal et ainsi les fréquences de résonnance (ou formants) du signal. L'obstruction totale ou partielle du flux d'air au moyen de la langue ou des lèvres permet également de produire d'autres catégories de sons correspondant aux consonnes.

1.2. Rappels acoustiques sur le signal de parole

Le mode de production particulier de la parole décrit ci-dessus a des conséquences sur la structure et le contenu du signal acoustique résultant. La Figure 3A. est une représentation temporelle, ou forme d'onde, de la phrase « chers auditeurs, bonsoir » prononcée par un locuteur masculin. On constate (Figure 3B.) que ce signal est composé de différents segments pseudopériodiques de forte énergie correspondant approximativement aux voyelles et séparés ou non par des segments non périodiques de faible énergie (bruit, silence ou explosions).

Cette représentation temporelle n'étant pas assez détaillée pour l'étude du signal de parole, on se base le plus souvent sur une représentation temps-fréquence, ou spectrogramme (Figure 3B.). Il apparaît ici que les segments énergétiques possèdent un riche contenu spectral. Ils sont constitués notamment d'une structure harmonique (fines raies parallèles horizontales sur le spectrogramme) avec une fréquence fondamentale f_0 (raie de plus basse fréquence). Cette dernière correspond à la fréquence de vibration des cordes vocales, c'est-à-dire à la hauteur de la voix, et véhicule des informations sur la morphologie (sexe, âge, etc…) et la psychologie (émotions) du locuteur. Outre la

présence d'harmoniques, la vibration des cordes vocales, ou voisement, peut être repérée sur le spectrogramme sous la forme d'une barre de voisement, concentration d'énergie dans les basses fréquences (observable, par exemple, entre 0.15 s et 0.55 s sur la Figure 3B).

Par ailleurs, les segments énergétiques sont caractérisés par la présence de 4 ou 5 bandes foncées horizontales plus larges, appelées formants (F1, F2, F3, F4, et F5). Elles correspondent aux fréquences de résonnance, déterminées par la forme du conduit vocal. Lors de la production d'une voyelle, la position fixe des muscles articulateurs se traduit par des fréquences de formants relativement constantes qui définissent l'identité de la voyelle prononcée. Au contraire, les phases de transition des formants entre les voyelles caractérisent les changements de position des muscles articulateurs lors de la production de certaines consonnes.

Finalement, les segments de faible énergie marquent la présence d'un silence ou d'une consonne fricative (/ʃ/ et /s/ dans l'exemple Figure 3).

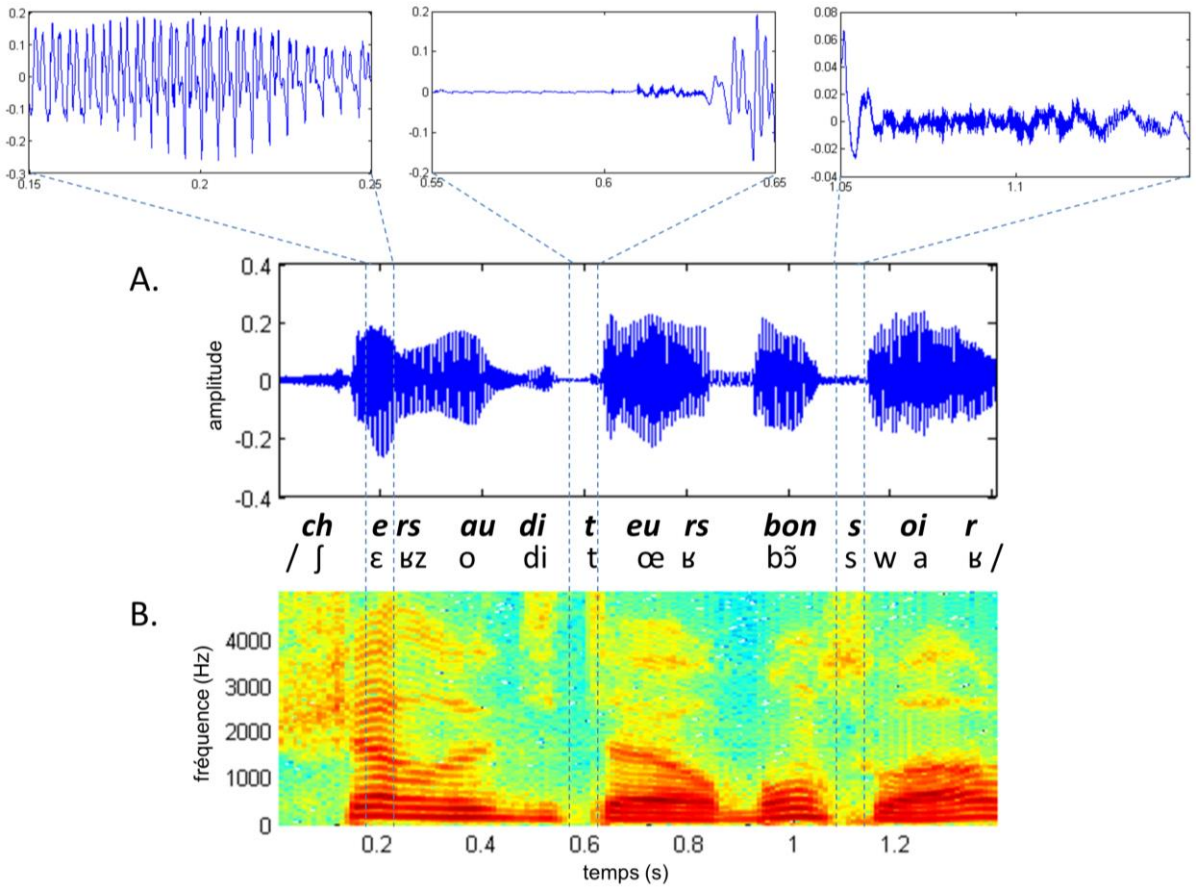


Figure 3 – Représentations de la phrase « Chers auditeurs, bonsoir. ». A. Représentation temporelle du signal, ou forme d'onde, et agrandissements de certaines portions du signal correspondant aux phonèmes /ɛ/, /t/ et /s/. B. Représentation temps-fréquence, ou spectrogramme, du même signal.

1.3. L'exemple des occlusives voisées

Nous avons choisi, pour illustrer les différentes méthodes et théories que nous décrirons dans la suite de cette partie, de présenter principalement (mais non exclusivement) les résultats obtenus concernant la compréhension d'un type particulier de consonnes : les occlusives voisées (/b/, /d/ et /g/). Ce choix est guidé par le fait que les 4 articles présentés dans cette thèse portent tous sur des tâches de catégorisation /b/-/d/ ou /d/-/g/. Nous verrons comment les différents groupes de recherche abordèrent le problème des indices acoustiques utilisés dans la différenciation de ces phonèmes.

En termes articulatoires, la production des consonnes occlusives implique un blocage momentané du flux d'air, au moyen des lèvres (/b/), de la langue placée au contact des alvéoles (/d/) ou au contact du palais mou (/g/). Le relâchement soudain de cette occlusion du conduit vocal, sous la pression de l'air contenu, produit une « explosion » acoustique d'amplitude importante, caractéristique de ce groupe de consonnes. Dans le cas des occlusives voisées, le voisement précède la désocclusion, c'est-à-dire que les cordes vocales sont déjà en vibration avant l'explosion. Ce trait phonétique les différencie des occlusives non voisées (/p/, /t/ et /k/) pour lesquelles les cordes vocales n'interviennent pas, ou seulement après le relâchement de l'occlusion. Ce mode de production particulier du son se répercute sur son spectrogramme. Ainsi pour une occlusive voisée la barre de voisement est présente avant l'explosion. Les spectrogrammes de trois productions de /aba/, /ada/ et /aga/ sont présentées Figure 4.

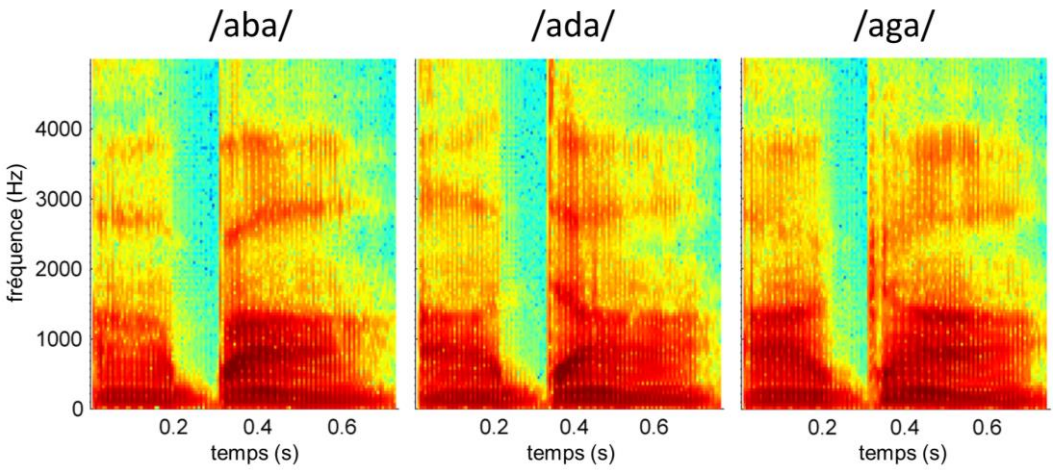


Figure 4 – Spectrogrammes des consonnes occlusives voisées dans un contexte vocalique /aCa/ (/aba/, /ada/ ou /aga/). Source des sons : Oldenburg Logatome Speech Corpus (Wesker et al., 2005).[1]

[1] Les sons peuvent être téléchargés à l'adresse : http://medi.uni-oldenburg.de/ollo/html/download.html

La barre de voisement apparaît entre les deux syllabes (de 200 à 300 ms). L'explosion correspond à la fine barre verticale de forte énergie au début de la deuxième syllabe, aux alentours de 300 ms. Du fait du mouvement transitoire des muscles articulateurs depuis leur position lors de l'occlusion pour atteindre la position correspondant à la voyelle (/a/), les formants passent par une phase de transition. La position initiale de ce mouvement est différente pour les trois consonnes et les trajectoires des formants sont donc différentes. Il est évident visuellement que le trait distinctif entre /b/, /d/ et /g/ (le point d'articulation) possède plusieurs corrélats acoustiques : notamment, la composition spectrale de l'explosion, la position initiale des formants F2 et F3 au début de la syllabe et la direction des transitions de ces formants.

Cette première description articulatoire nous permet donc d'identifier un certain nombre d'informations acoustiques disponibles dans le son de parole lors de sa réception par l'auditeur. Elle ne nous dit cependant pas lesquelles sont effectivement utilisées par le système auditif lors de la compréhension, ni sur quel « support » acoustique est fondée la perception des traits distinctifs. Pour répondre à ces questions, il est nécessaire de compléter cet examen de la production de la parole et de la composition du spectrogramme par des résultats en perception de la parole.

1.4. Organisation de la partie théorique

La question de la relation entre un son de parole et le percept phonémique correspondant traverse la recherche en psycholinguistique. L'accent s'est souvent porté sur l'identification des indices acoustiques, éléments présents dans le signal acoustiques et investis d'une fonction discriminante dans le processus de reconnaissance de la parole. Le chapitre 2 résume les premières recherches effectuées sur le sujet et les obstacles auxquels elles se sont trouvées confrontées. À partir de ce point, nous retracerons le chemin suivi par les chercheurs en psychoacoustique de la parole depuis plus d'un demi-siècle, avec une démarche aussi chronologique que possible, la connaissance scientifique dans le domaine étant intimement liée au progrès des techniques utilisées.

Les observations rassemblées au chapitre 2 conduisirent les chercheurs à réviser leur conception intuitive de la perception de la parole et à formuler de nouveaux modèles de l'interface acoustico-phonétique, dont certains seront présentés au chapitre 3.

Sur la base de ces modèles, nous verrons au chapitre 4 comment plusieurs groupes de recherche mirent en œuvre des approches très différentes pour parvenir à identifier les primitives acoustiques de la parole. Les avantages et contraintes relatives en seront évalués, afin de faire apparaître les aspects de la perception de la parole qui échappent encore aux techniques actuelles.

Les mécanismes décrits dans les premiers chapitres ne sont ni immuables ni universels. Au contraire, ils évoluent au cours de l'existence et peuvent se révéler déficitaires dans certains cas. L'aspect « dynamique » de la question sera envisagé au chapitre 5, à travers la description de deux formes de plasticité du système auditif, sur deux échelles de temps différentes : l'adaptation de l'extraction des indices lors d'une écoute en présence de bruit et les bénéfices de l'entraînement musical sur la perception de la parole. Finalement, nous verrons dans une 3ème partie un exemple de trouble neuro-développemental se traduisant notamment par un déficit phonologique : la dyslexie.

Le chapitre 6 s'intéressera ensuite aux bases neuronales de l'encodage de la parole, à travers la description des champs récepteurs des neurones auditifs. Un bref aperçu des résultats neurobiologiques obtenus par l'analyse des enregistrements d'électrodes implantées dans le cortex auditif nous permettra de corroborer les modèles déduits des observations comportementales.

La recherche des primitives auditives utilisées par le système pour identifier les phonèmes s'inscrit dans le champ d'étude plus général des primitives perceptuelles visant à répondre à deux questions majeures : quels sont les paramètres du stimulus qui influencent la perception de l'observateur ? et comment l'observateur combine-t-il ces indices pour donner naissance à un percept ? Nous verrons au chapitre 7 comment ce problème fut abordé pour la modalité visuelle par la méthode des Images de Classification. Nous décrirons en détail l'évolution de la technique et de son cadre conceptuel, depuis sa formulation par Ahumada en 1971 jusqu'à ses perfectionnements théoriques les plus récents. Nous pourrons alors examiner la possibilité de nous inspirer de ces travaux pour attaquer le problème des primitives perceptuelles de la parole (chapitre 8).

1.5. Organisation de la partie expérimentale

La partie théorique sera suivie d'une partie expérimentale décrivant les travaux réalisés au cours de cette thèse. Chaque chapitre correspond ici à un article publié ou en cours de publication dans une revue internationale à comité de lecture et, donc, rédigé en anglais. Néanmoins, nous introduirons et conclurons chaque article par un paragraphe de résumé en français.

Le premier article (chapitre 9) examine la possibilité de transférer la méthode des Images de Classification depuis la modalité visuelle vers la modalité auditive. Nous décrirons d'abord les contraintes pratiques et théoriques puis nous illustrerons le bon fonctionnement de notre algorithme à travers l'exemple d'une catégorisation /aba/-/ada/ dans le bruit. Cette étude nous permettra d'identifier les indices acoustiques employés par les auditeurs lors de cette tâche.

Le second article (chapitre 10) concerne l'application de cette nouvelle méthode à un groupe de participants, en vue du test d'hypothèses sur la catégorisation de phonèmes dans la population normo-entendante. La tâche utilisée est inspirée de l'expérience de Mann (Mann, 1980) et requiert la distinction des syllabes /da/ et /ga/ indépendamment du contexte dans lequel elles ont été prononcées (/al/ ou /aʁ/).

Les deux articles suivant visent à étudier les stratégies d'écoute de différents groupes d'auditeurs. La méthode développée sera ici appliquée à des questions théoriques actuelles dans le domaine de la psycholinguistique. Le troisième article (chapitre 11) porte sur les capacités accrues des musiciens experts en termes de compréhension de la parole dans le bruit. Enfin, le dernier article (chapitre 12) explore la question des stratégies d'écoute développées par les auditeurs dyslexiques pour compenser leur déficit phonologique.

Les résultats obtenus dans ces quatre études seront récapitulés au chapitre 13. Pour terminer, le dernier chapitre de cette thèse (chapitre 14) propose une synthèse de ces résultats en vue d'une interprétation plus générale dans le contexte de la littérature. Nous discuterons notamment des limites et de possibles améliorations de notre méthode, de la portée des observations réalisées pour la compréhension de l'interface acoustico-phonétique ainsi que celle de son lien avec la neurobiologie.

2. Une approche psychoacoustique de la compréhension de la parole

L'intérêt des chercheurs pour la perception de la parole précède largement le développement des premières techniques de neuro-imagerie. La psychoacoustique, l'étude de la perception des sons sur la seule base de l'analyse du comportement du participant, fut donc longtemps la seule approche possible pour élucider les questions posées dans ce domaine. Il s'agissait de révéler la structure cachée de la cognition par la mesure des actions : ainsi un grand nombre de modèles psycholinguistiques encore utilisés aujourd'hui, comme TRACE (McClelland & Elman, 1986) ou la Théorie Motrice de Liberman (Liberman & Mattingly, 1985), sont fondés quasi-exclusivement sur l'observation des réponses d'un participant à la présentation de stimuli auditifs (mesurées en termes de temps de réaction, de taux de compréhension correcte, de seuil d'intelligibilité…). En outre, cette méthode d'investigation purement comportementale conserve toute sa pertinence, même après l'avènement de l'imagerie médicale parmi les outils d'étude mis à disposition des chercheurs. En effet, cette méthodologie demeure un indicateur cognitif plus direct (et par suite, moins sujet à interprétation) et moins coûteux pour tester les hypothèses concernant les traitements effectués par le cerveau et former de nouveaux modèles cognitifs.[2]

Dans ce chapitre nous décrirons les premières expériences psychoacoustiques concernant la perception des phonèmes (partie 2.1) et fondant un modèle naïf, appelé modèle alphabétique, de la compréhension de la parole. Dans une deuxième partie (2.2) nous nous appuierons sur les résultats d'expériences postérieures pour critiquer ce modèle selon cinq axes principaux.

2.1. La perception catégorielle de la parole

2.1.1. L'expérience du continuum synthétique

L'une des contributions fondamentales de la psychoacoustique pour la compréhension des mécanismes à la base de la reconnaissance de la parole fut la mise en évidence de la perception catégorielle des phonèmes. Dans les années 50, Alvin Liberman et ses collaborateurs des Haskins Laboratories mirent au point une expérience qui allait donner matière à étude pour les décennies à venir (Liberman et al., 1957). Le principe en est extrêmement simple : 14 stimuli de parole synthétique sont répartis à intervalles égaux sur un continuum acoustique. Ils constituent des versions « simplifiées » d'un son de parole composé de deux formants uniquement. Le premier

[2] Voir (Tiberghien, 2007) pour une discussion sur les rapports entre neurosciences et psychologie cognitive.

est fixé à 360 Hz avec une attaque à 0 Hz. Le second possède une partie stable, à 2160 Hz, identique pour tous les stimuli, mais avec une fréquence d'attaque variable entre 1320 Hz et 2880 Hz (par pas de 120 Hz). Les 14 stimuli obtenus sont représentés Figure **5**. D'un point de vue perceptuel, ces sons de paroles vont de /ba/ (stimulus #1) à /ga/ (stimulus #14) en passant par /da/ (autour du stimulus #7).

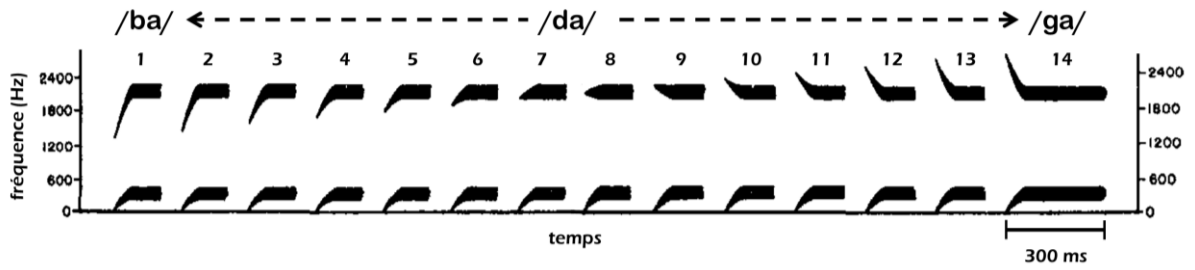


Figure 5 – Représentation schématique des 14 stimuli de parole synthétique utilisés dans l'expérience de perception catégorielle. Tous les stimuli ont une durée de 300 ms. Adapté de (Liberman et al., 1957).

Lorsqu'ils demandèrent à des participants d'identifier ces syllabes, ils observèrent que le long de ces variations uniformes du signal acoustique les proportions de chaque réponse (/ba/, /da/ ou /ga/) n'évoluent pas de manière régulière, mais semblent au contraire marquer des frontières brutales (voir Figure 6A). En d'autres termes, le continuum physique est divisé en plages où les stimuli, bien que différents, sont systématiquement regroupés sous une même étiquette. Ces zones sont séparées par des points d'inflexion très marqués aux alentours des stimuli #4 et #9. Par conséquent peu de stimuli sont perçus de manière ambigüe. Cette observation constitue le premier aspect de la perception catégorielle de la parole : elle agit comme une interface robuste entre l'espace continu des stimuli acoustiques et l'espace discret des perceptions phonémiques. La frontière phonémique entre deux réponses est ainsi définie comme le point où le stimulus a des probabilités égales d'être catégorisé comme un phonème ou un autre (Repp & Liberman, 1984).

Une observation complémentaire de l'équipe de Liberman concerne les capacités de discriminations entre les sons proches. Celles-ci sont évaluées classiquement en mesurant la capacité de l'auditeur à différencier un stimulus (i) appartenant au continuum du stimulus qui le suit (i+1). Le pourcentage de discriminations correctes pour chacune des 13 comparaisons possibles est présenté Figure 6B (trait continu). On constate qu'il présente deux maxima, indiquant que le participant distingue très bien le stimulus #4 du #5 et le stimulus #8 du #9. Ces deux points du continuum correspondent approximativement aux frontières mises en évidence dans l'expérience précédente. Au contraire, les paires de stimuli séparés par une différence acoustique équivalente mais situés au centre d'une catégorie phonémique

(par exemple les stimuli #6 et #7) possèdent un niveau de discrimination sensiblement égal au hasard (50%), ce qui signifie qu'ils apparaissent identiques à l'auditeur.

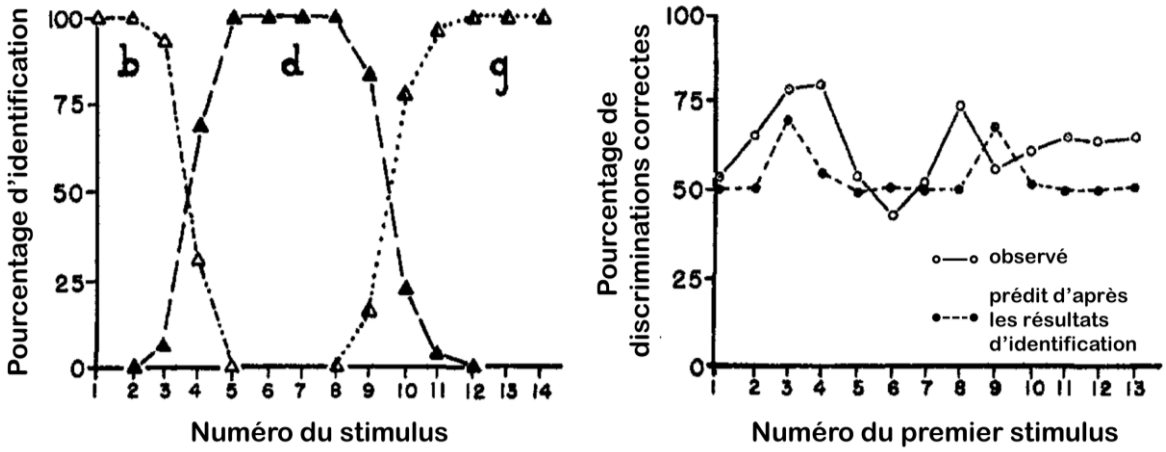


Figure 6 – Résultats de l'expérience de perception catégorielle. A. Pourcentage d'identification de /ba/, /da/ et /ga/ le long du continuum de stimuli. B. Pourcentage de réponses correctes dans la tâche de discrimination entre deux stimuli adjacents du continuum (trait continu : observé ; trait discontinu : prédit d'après les pourcentages d'identification). Adapté de (Liberman et al., 1957).

En résumé, non seulement la perception catégorielle délimite des régions phonétiques discrètes dans l'ensemble des sons de parole mais elle déforme aussi l'espace perceptuel en fonction de ces catégories. La sensibilité du système à une petite variation physique du signal de parole n'est pas constante tout le long du continuum : elle est perçue très nettement lorsqu'elle chevauche une frontière mais elle est estompée lorsque les deux stimuli sont catégorisés de la même manière. Tout se passe comme si l'auditeur faisait abstraction des variations acoustiques au sein des catégories phonémiques et se focalisait seulement sur les différences acoustiques entre les catégories, pertinentes pour la transmission du message linguistique.[3]

[3] Lorsqu'ils mirent en évidence ce phénomène, Liberman et ses collaborateurs émirent l'hypothèse que la perception catégorielle était spécifique à la perception de la parole (Liberman et al., 1957). En effet, une expérience en tout point similaire à la précédente, mais effectuée sur des stimuli non paroliers, ne permet pas de reproduire les résultats. Cependant, il fut démontré par la suite que ce phénomène n'était en fait qu'un cas particulier d'un mécanisme beaucoup plus large (Holt & Lotto, 2010). En effet, le même type de comportement a été observé pour la modalité visuelle (Hoonhorst et al., 2011; Maddox et al., 2002), notamment pour la perception des expressions des visages ou des couleurs. Par ailleurs, la perception catégorielle des sons en fonction de leur signification existe également chez les primates, indépendamment de tout contact humain (Russ et al., 2007).

2.1.2. Variabilité du signal de parole naturel

Le mécanisme de perception catégorielle permet de résoudre, au moins partiellement, le problème fondamental que représente la variabilité intrinsèque du signal pour tous les systèmes de reconnaissance automatique de la parole. Cette variabilité a de nombreuses origines : des différences environnementales (p.ex. acoustique de la pièce, distance du locuteur), des différences interindividuelles de morphologie (p.ex. longueurs du conduit vocal et des cordes vocales, liées notamment à l'âge, au sexe, et à la taille du locuteur), d'accent (p.ex. le /r/ français qui sera, selon les régions, roulé, grasseyé ou guttural) ou de style d'élocution (p.ex. parole exagérément articulée à l'intention d'un enfant ou lors d'une discussion dans un bruit de fond important). Même deux productions d'un phonème donné, par un même locuteur et dans les mêmes conditions, ne sont jamais identiques (Kuhl, 2004; Repp & Liberman, 1984). Les facteurs responsables de cette variabilité intra-individuelle englobent, notamment, l'influence du niveau sémantique (prosodie, accent d'insistance), la position du phonème dans le mot ou encore la coarticulation avec les phonèmes précédents et suivants sur laquelle nous reviendrons en détail par la suite (voir le paragraphe 2.2.2).

Il résulte de ces différentes sources de variabilité que des phonèmes possédant des caractéristiques acoustiques très différentes peuvent être perceptuellement équivalents. Autrement dit, ces variations ne sont pas pertinentes phonétiquement puisqu'elles ne permettent pas d'identifier le contenu du message. Ainsi, si l'on positionne différents exemplaires de voyelles sur un plan en fonction des fréquences de leurs formants F1 et F2 [représentation en Triangle de Delattre (Delattre, 1948; Delattre et al., 1952)], on constate une dispersion importante des points correspondant à une même voyelle (Figure 7). Des régions assez vastes du plan F1-F2 sont ainsi associées à un même élément. Ceci complexifie le décodage du son de parole perçu.

En définissant des régions de l'espace sonore associées systématiquement à une même étiquette phonétique la perception catégorielle semble fournir une solution au problème de la variabilité car elle permet de faire abstraction des variations non pertinentes du son de parole. Par ce mécanisme, deux exemplaires de la voyelle /i/, même s'ils ne sont pas identiques acoustiquement (i.e. situés à des positions différentes sur le Triangle de Delattre), engendrent le même percept chez l'auditeur, assurant de ce fait un décodage robuste de l'information. Nous verrons cependant par la suite que cette explication simple est remise en question par l'existence d'une variabilité inter-catégorielle (cf. paragraphe 2.2.2).

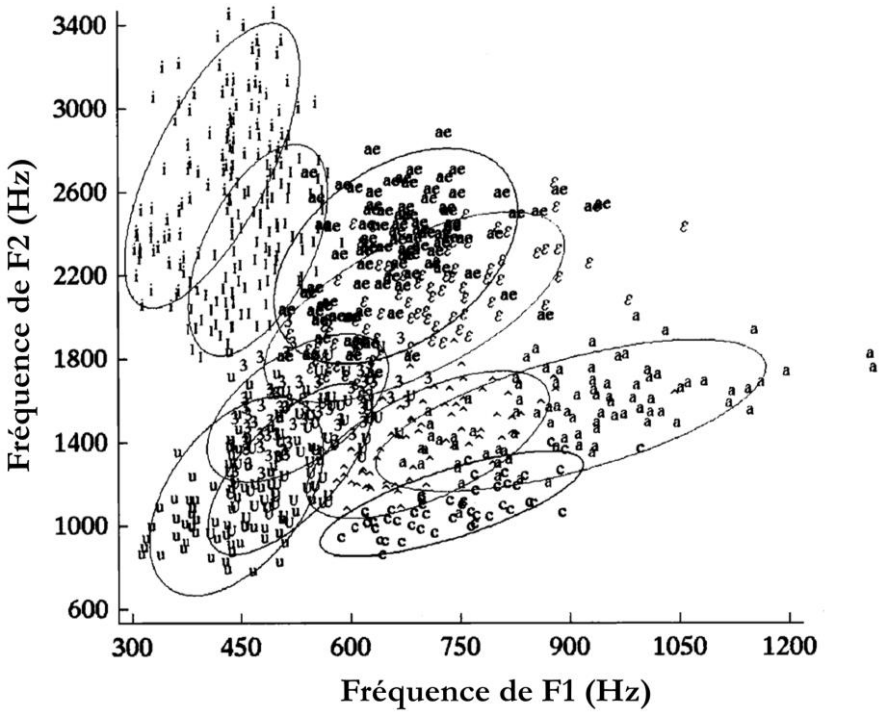


Figure 7 – Représentation du triangle de Delattre. Positions sur le plan F1-F2 des exemplaires de 10 voyelles anglaises, produites par 140 locuteurs différents. Adapté de (Hillenbrand et al., 1995).

2.1.3. Frontières phonémiques

La perception des phonèmes apparaît donc à ce stade comme une forme de code projetant l'espace multidimensionnel et continu des signaux de parole dans un ensemble discret et fini de catégories phonétiques. Pour « casser » le code utilisé par notre cerveau, il suffirait donc simplement d'identifier les dimensions du signal qui déterminent ces catégories (par exemple, l'attaque du F2 pour le continuum /ba/-/da/-/ga/) ainsi que les frontières phonémiques qui marquent le passage d'une catégorie à une autre (situées ici à 1680 Hz et 2280 Hz) (Liberman et al., 1967).

La technique du continuum synthétique exposée ci-dessus, bien que présentant de sérieuses limitations (cf. paragraphe 4.1), a été utilisée de façon méthodique par les chercheurs des Haskins Laboratories avec différents continuums acoustiques composés d'occlusives non voisées /p/-/t/-/k/ (Liberman et al., 1952) ou voisées /b/-/d/-/g/ (Liberman et al., 1957), de nasales /m/-/n/-/ŋ/ (Liberman et al., 1954), de fricatives voisées /v/-/ð/ et /z/-/ʒ/ (Delattre et al., 1964) et, enfin, des consonnes continues /w/-/r/-/l/-/j/ (O'Connor et al., 1957). La consonne était systématiquement présentée dans un complexe CV, avec différentes voyelles. Ce travail laborieux permit aux chercheurs de tracer des « cartes » relativement précises des frontières phonémiques de consonnes

pour un large panel de syllabes.[4] Ils démontrèrent ainsi que les transitions de formants, qui pouvaient apparaître comme de simples épiphénomènes de la production de parole dus à l'inertie des articulateurs, étaient en réalité déterminants pour la perception : leurs attaques (*onsets*), caractérisées par leur fréquence initiale et leur direction (montante ou descendante), jouent un rôle capital dans l'identification des consonnes.

2.2. Le modèle alphabétique et ses limites

La série d'observations présentée dans le paragraphe précédent nous conduit à un premier modèle relativement simple de la compréhension de la parole : les sons paroliers, bien que très variables en apparence, présentent des caractéristiques acoustiques qui codent les phonèmes selon des règles invariantes, connues de tous les locuteurs de la langue. Ainsi l'auditeur n'a-t-il pas accès à toute la richesse du signal de parole car il perçoit uniquement les événements acoustiques porteurs de sens au niveau linguistique. Ces événements sont caractérisés au niveau perceptif par des frontières abruptes qui délimitent précisément les différents phonèmes de la langue. À l'inverse, une variation du signal à l'intérieur de la région correspondant à un phonème est imperceptible pour l'auditeur. Le décodage individuel et successif des phonèmes produits par le locuteur permettrait ainsi à l'auditeur de reconstituer les mots, puis le message transmis. Cette conception « intuitive » de la compréhension de la parole, construite par analogie avec la lecture, était celle des chercheurs des Haskins Laboratories au début des années 50. Par la suite, nous lui donnerons le nom de modèle « alphabétique ».

Un certain nombre d'observations ultérieures les obligèrent cependant à reconsidérer cette conception trop élémentaire de l'interface acoustico-phonétique. Nous les regroupons ici selon cinq axes de critique principaux. Ces observations conduiront ensuite à raffiner le modèle et poseront les contraintes pour la recherche des indices acoustiques de la parole.

[4] La mise en évidence de ces frontières phonémiques pose naturellement la question de leur origine : sont-elles placées arbitrairement dans l'espace des signaux acoustiques ou coïncident-elles avec des discontinuités perceptives innées ? Libermann soutint d'abord la première hypothèse, en argumentant que toutes les langues ne possèdent pas les même frontières pour des phonèmes similaires (Liberman et al., 1957). Ce point de vue fut néanmoins contredit par des découvertes ultérieures - notamment la capacité des nouveau-nés (Kuhl, 2004), et de certains mammifères non humains (Serniclaes, 2000), à percevoir les frontières phonémiques de langues auxquelles ils n'ont jamais été exposés – ainsi que par la généralisation de certaines frontières phonémiques à des tâches non parolières (Serniclaes, 2011). Ceci amena les chercheurs à formuler l'hypothèse qu'il existe des prédispositions perceptives favorisant certains contours pour les catégories phonétiques mais que ceux-ci sont ensuite modulés par des adaptations linguistiques au cours du développement (Serniclaes, 2000; Serniclaes et al., 2004).

2.2.1. Intrication des informations phonétiques

Historiquement le premier point d'achoppement du modèle alphabétique de la compréhension de la parole fut l'impossibilité de concevoir une machine à lire fonctionnelle destinée aux aveugles (Cooper et al., 1984; Galantucci et al., 2006; Liberman et al., 1967; Shankweiler & Fowler, 2015). Le principe de base en était simple : un capteur optique balayait le texte écrit à vitesse donnée en convertissant directement le signal lumineux recueilli en signal sonore (plusieurs règles de conversion furent expérimentées dans les différentes versions de la machine). Ce dispositif permettait en pratique de traduire lettre par lettre l'alphabet visuel en un alphabet auditif. Après une phase d'apprentissage plus ou moins longue, l'auditeur aveugle, à l'écoute de la version transcrite par la machine, était capable de restituer, au moins partiellement, le texte écrit. Néanmoins, la vitesse à laquelle les utilisateurs parvenaient à décoder efficacement le signal émis était extrêmement lente (de l'ordre de 4 à 10 mots par minute selon la règle de conversion utilisée, contre 200 à 300 mots par minute dans le cas de la lecture visuelle). Cette borne supérieure des performances de décodage, quel que soit l'alphabet sonore utilisé et la durée de la période d'apprentissage, est due à une limite inhérente à la résolution temporelle de l'oreille humaine. En effet, celle-ci ne peut distinguer qu'un maximum de 12 à 15 sons successifs par seconde. Au-delà de ce débit, les « lettres sonores » ne sont plus perçues comme des entités distinctes, d'où l'impossibilité pratique d'augmenter la vitesse de lecture.

Suite à l'échec de leurs efforts pour améliorer les performances de la machine à lire, les chercheurs se demandèrent ce qui différencie fondamentalement leur système et le codage de la parole naturelle. Comment le son de parole parvient-il à transmettre environ 30 phonèmes par seconde, un débit qui dépasse la limite de la résolution du système auditif évoquée ci-dessus (Liberman et al., 1967) ? Ils réalisèrent alors que la faiblesse de l'alphabet sonore réside dans son caractère séquentiel (i.e. les lettres sont transmises les unes après les autres). Au contraire, dans un son de parole, les segments correspondants aux phonèmes successifs se superposent dans une large mesure[5]. Les informations phonétiques sont ainsi transmises de manière simultanée plutôt que par une séquence de structures sonores discrètes. Chaque phonème est si inextricablement intriqué avec le reste du son de parole que ce dernier s'avère en grande partie indécomposable. Par exemple à partir d'un enregistrement de la syllabe /di/ il est impossible d'isoler par troncature un segment du signal correspondant au phonème /d/ seul.

La cooccurrence des mouvements articulatoires correspondant à la production des phonèmes successifs lors de l'émission entraîne une superposition des indices acoustiques qui les définissent. Autrement dit, à l'intérieur d'un segment donné du son de parole, sont simultanément présents des indices relatifs à l'identité des phonèmes

[5] Les sons ne sont donc pas des « perles enfilées sur l'axe du temps », contrairement à ce que suggère la formule « *beads on a string* » couramment utilisée jusque dans les années 1950. (Boë, 1997)

précédents et suivants. Naturellement, cette observation remet en cause le caractère discret du décodage dans le modèle alphabétique évoqué ci-dessus mais elle a également d'autres implications.

2.2.2. Coarticulation et variabilité des frontières phonémiques

Comme nous venons de l'évoquer, les mouvements articulatoires liés à la production de deux phonèmes successifs se superposent dans une large mesure. Ils sont de plus fortement interdépendants. En effet, la déformation du conduit vocal permettant de passer d'un phonème au suivant s'effectue de manière continue ; la position des articulateurs au début de la production d'un phonème, ainsi que la position visée à la fin du mouvement articulatoire, influent sur le décours exact de l'articulation et sur les commandes motrices réalisées. En outre, dans la plupart des cas, l'articulation s'ajuste en fonction du voisinage phonémique afin de minimiser l'effort articulatoire. Ainsi par exemple, en français, le son /k/ est réalisé comme une occlusive palatale [c] (i.e. articulée avec la langue contre l'avant du palais) lorsqu'il précède une voyelle antérieure comme dans /ki/, mais il est réalisé comme une occlusive vélaire [k] (i.e. articulée avec la langue contre l'arrière du palais) lorsqu'il précède une voyelle postérieure comme dans /ku/. La coarticulation peut être observée non seulement entre une consonne et une voyelle comme dans l'exemple ci-dessus mais aussi entre deux consonnes (Mann, 1980; Mann & Repp, 1981) ou entre deux voyelles (Carré et al., 2001).

Ce phénomène est omniprésent dans le langage parlé, à tel point qu'un signal de parole synthétisé sans prise en compte de la coarticulation nous est incompréhensible. Il participe de la variabilité du signal de parole naturelle mentionnée au paragraphe 2.1.2 : les différents exemplaires d'un même phonème sont dispersés au sein de la région de l'espace acoustique correspondant à leur catégorie. Cependant l'exemple de la coarticulation permet de mettre en évidence un degré de variabilité supplémentaire : les régions elles-mêmes sont flexibles, leurs frontières pouvant varier selon les conditions de la perception, notamment le contexte phonétique. Ce phénomène, appelé compensation pour la coarticulation, est le pendant perceptif de l'effet de coarticulation qui accompagne la production.

Deux cas de figure sont particulièrement révélateurs de cette interaction entre le contexte phonétique et la position des frontières phonémiques (Liberman & Mattingly, 1985) :

(1) Deux signaux très différents peuvent correspondre à un même phonème, prononcé dans différents contextes. Par exemple, dans les syllabes /di/ et /du/ dont les spectrogrammes schématiques sont présentés Figure 8, des transitions de formants très différentes (montante ou descendante) donnent lieu au même percept /d/ en contexte /i/ ou /u/ (Fowler & Rosenblum, 2014; Galantucci et al., 2006).

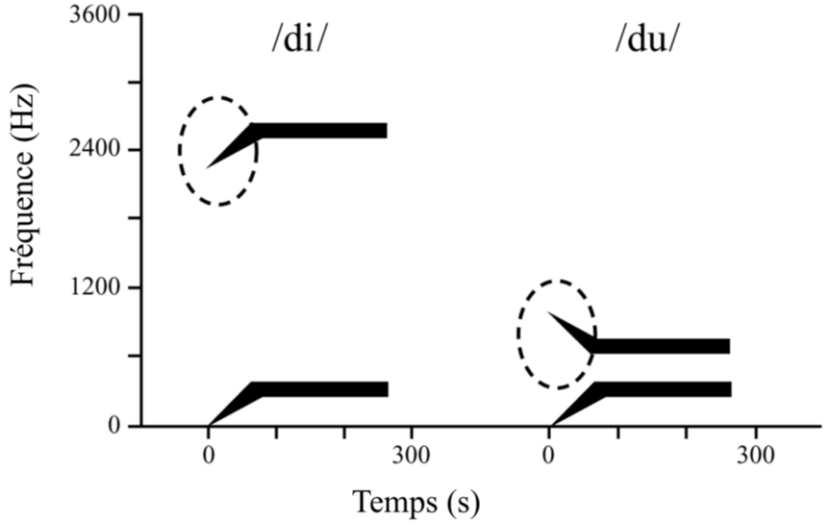


Figure 8 – Représentation simplifiée des spectrogrammes des syllabes /di/ et /du/. Les ovales indiquent les transitions de F2 donnant lieu au percept /d/ dans les deux cas. Adapté de (Galantucci et al., 2006).

(2) De manière complémentaire, un même signal acoustique peut parfois être interprété comme un phonème ou un autre selon les sons qui le précèdent ou le suivent. Le premier exemple de ce phénomène fut donné par Liberman et ses collaborateurs. Un bruit bref situé autour de 1600 Hz était placé avant une voyelle synthétique /i/ ou /u/, conduisant aux percepts /pi/ ou /pu/. Pourtant, le même bruit placé devant une voyelle synthétique /a/ était perçu comme /ka/ (Liberman et al., 1952). Ce bruit seul ne correspond donc ni à un /p/ ni à un /k/ et son interprétation ne dépend donc pas uniquement de son contenu acoustique. De même, la représentation des taux de réponse dans une tâche de catégorisation de consonnes /b/, /d/ ou /g/, dans des syllabes synthétiques CV indique qu'une même transition de formant peut être interprétée de manière différente selon le contexte vocalique (Liberman et al., 1954).

Ces deux observations remettent en question le projet de tracer une « carte » pour localiser de manière absolue les catégories phonétiques de manière absolue dans l'espace des signaux de parole puisque, d'après le point 1 (sons différents + contextes différents → percepts identiques), un même phonème est susceptible d'être représenté par des sons extrêmement différents selon le contexte et que, d'après le point 2 (sons identiques + contextes différents → percepts différents), leurs frontières ne sont pas situées à des positions fixes mais, au contraire, largement dépendantes de nombreux facteurs parmi lesquels le contexte. L'analyse de la catégorisation d'un phonème dans une syllabe donnée ne peut être généralisée directement à la catégorisation du même phonème dans un autre contexte. Il n'existerait donc pas de correspondance biunivoque entre une propriété physique d'un son de parole et sa perception par l'auditeur.

Bien qu'un grand intérêt ait été porté à la coarticulation (voir notamment l'exemple de l'étude de Mann mentionné au paragraphe 3.1.3), il ne s'agit pas de l'unique cause de déplacement des frontières phonémiques. D'autres sources importantes de variabilité peuvent être trouvées dans les caractéristiques individuelles des locuteurs [p.ex. déformation du triangle vocalique en fonction de l'âge et du sexe du locuteur (Hillenbrand et al., 1995)], la réduction de geste articulatoire dans un débit d'élocution rapide (Carré & Divenyi, 2000), le contenu lexical du signal (Ganong, 1980), l'identité perçue des phonèmes adjacents (Serniclaes & Wajskop, 1992) ou, encore, des effets d'adaptation perceptuelle à la tâche (Repp & Liberman, 1984). L'idée d'une étude exhaustive des frontières des catégories phonémiques dans tous les contextes apparaît donc impossible.

2.2.3. Le phonème n'est pas l'unité de base de la compréhension de la parole

L'étude de Liberman est fondée sur le postulat selon lequel le phonème possède une réalité cognitive ; elle ne se place donc pas en position de déterminer si les unités phonémiques jouent un véritable rôle lors de la perception. Pourtant, la compréhension du langage oral ne se limite pas à l'identification d'une chaîne de phonèmes.

L'étude des lésions anatomiques ou artificielles (par stimulation magnétique transcrânienne) d'aires cérébrales impliquées dans le traitement de la parole, semble indiquer que les deux processus ne sont pas directement dépendants. En effet, dans certains cas, la capacité à identifier les mots est préservée tandis que l'accès aux phonèmes qui le composent est perturbé, voire impossible (Holt & Lotto, 2010; Krieger-Redwood et al., 2013). Ceci implique que le phonème n'est pas l'unité fondamentale utilisée pour le décodage mais seulement un objet perceptuel secondaire. Dans la même ligne, Segui et ses collaborateurs démontrèrent que le temps de catégorisation d'un phonème est supérieur au temps de catégorisation de la syllabe à laquelle il appartient et que ces deux mesures sont fortement corrélées (Segui et al., 1981). Ceci indique que l'identification du phonème dépend de l'identification de la syllabe, et non l'inverse. Enfin, comme nous l'avons noté plus haut les indices acoustiques permettant l'identification de deux phonèmes successifs peuvent être superposés dans le son de parole. De ce point de vue, une unité de langage un peu plus grande, comme la syllabe, offre plus de régularité (Delattre et al., 1952; Massaro & Chen, 2008). Tous ces résultats pointent donc vers la syllabe – et non le phonème – comme unité de base de la perception de la parole.

La question de savoir si les syllabes priment sur les phonèmes dans le traitement de la parole est importante mais dépasse néanmoins le cadre de cette thèse. En effet, les modèles prenant pour unité perceptuelle la syllabe ou le phonème ne sont pas différentiables dans des tâches de catégorisation phonémiques telles que celles étudiées ici.

2.2.4. Catégorisation plutôt que perception catégorielle

Une autre objection concerne la nature même du percept phonémique. Il est clair que les phonèmes forment des catégories discrètes ; néanmoins, deux exemplaires du même phonème ne sont pas toujours perçus de manière équivalente. En d'autres termes, pour les résultats présentés Figure 6, le taux de discriminations correctes n'est pas parfaitement prédictible par le seul taux d'identifications correctes, car la comparaison des instances de deux phonèmes ne consiste pas seulement à examiner si les étiquettes phonémiques qui leurs sont associées sont différentes ou non. Par ailleurs, des études calquées sur le protocole de l'expérience de Liberman ont montré que la perception des voyelles et des consonnes fricatives et liquides n'est pas aussi catégorielle que celle des consonnes occlusives (Fry et al., 1962).

Mieux, au sein d'une même catégorie, les auditeurs perçoivent certains exemplaires comme étant « plus représentatifs » que d'autres du phonème auquel ils correspondent. De plus, pour un espacement donné sur le plan acoustique, un couple de productions proche de ces prototypes est plus difficile à discriminer qu'un couple de productions qui en est éloigné (Holt & Lotto, 2010). Le fait que les catégories possèdent une structure interne remet en cause la validité perceptuelle de la notion de frontière phonémique. L'espace acoustique serait plutôt déformé à proximité de représentations prototypiques des différents phonèmes, stockées en mémoire et agissant comme des aimants perceptifs (Johnson, 2011; Kuhl, 1991; Serniclaes, 2011). Au final, les sons de parole sont donc catégorisés mais leur perception ne paraît pas exclusivement catégorielle.

À la suite de ces résultats, les chercheurs ont progressivement abandonné la notion de frontière phonémique dans leurs modèles psycholinguistiques pour porter l'accent sur les représentations mentales des phonèmes. Ces dernières structurent l'espace perceptuel, l'auditeur analysant un signal de parole reçu en le comparant à ces exemplaires idéaux. La question de la nature de ces représentations sera abordée au paragraphe 3.1.

2.2.5. Redondance des indices acoustiques

Le modèle alphabétique exposé précédemment suppose que la catégorisation d'un son de parole s'effectue en fonction d'un indice acoustique unique. Ceci est effectivement vrai dans le cadre de l'expérience de Liberman puisqu'une seule propriété du son (l'attaque du F2) est variable le long du continuum synthétique. En revanche, dans le signal de parole naturel, la covariance des caractéristiques acoustiques entraîne la multiplicité des informations disponibles pour sa compréhension.

En effet, les actions des organes à l'origine de la parole n'ont pas une traduction biunivoque en termes acoustiques (Delattre, 1958). Par exemple, la production du trait

de voisement est le plus souvent caractérisée au moyen du Délai d'Établissement du Voisement (VOT, *Voice Onset Time*), c'est-à-dire l'intervalle de temps entre le relâchement du blocage de l'écoulement de l'air et le début de la vibration laryngale. Serniclaes et Arrouas soulignent cependant que la production du trait de voisement se manifeste à la fois par un VOT positif, une f_0 initiale grave et une durée plus importante des transitions formantiques (Serniclaes & Arrouas, 1995). Ces trois caractéristiques acoustiques peuvent donc être utilisées alternativement, ou même conjointement, lors de l'identification du trait.

Cette redondance de l'information dans le signal de parole le dote d'une grande robustesse face à tous les types de dégradation, aucun indice acoustique n'étant à lui seul absolument nécessaire à sa compréhension. Ceci est attesté par plusieurs expériences complémentaires portant sur la parole dégradée. Ainsi, des enregistrements de phrases ou de mots pour lesquels l'information spectrale a été fortement altérée peuvent-ils demeurer globalement intelligibles (Apoux & Healy, 2009; Loizou et al., 1999; Shannon et al., 1995). De même, dans la dimension temporelle, il faut appliquer des dégradations relativement importantes pour que leurs effets se fassent ressentir sur la compréhension (Xu et al., 2005). Même la suppression de segments de signal à des intervalles réguliers (Miller & Licklider, 1950) ou leur inversion temporelle (Saberi & Perrott, 1999) ne réduit que faiblement l'intelligibilité. Ainsi, malgré une dégradation assez sévère du signal selon l'une de ses deux dimensions (temporelle ou spectrale), l'auditeur reste capable de récupérer l'information manquante dans l'autre. En effet, l'information sur l'identité du phonème n'est pas localisée sur une seule propriété du son mais peut au contraire être déduite de différents indices. Logiquement, ceci impose de réviser le modèle – devenu trop simpliste – selon lequel chaque trait phonétique correspondrait à un indice acoustique unique.

3. Théories et modèles de l'interface acoustico-phonétique

Les premières expériences psychoacoustiques de catégorisation de phonèmes conduisirent logiquement les chercheurs à supposer un fonctionnement de la compréhension de la parole calqué sur celui de la lecture, qui consisterait en un décodage individuel et successif des phonèmes par une détection de propriétés acoustiques les définissant de manière univoque. Ce modèle alphabétique a cependant été invalidé par plusieurs observations postérieures, parmi lesquelles :

- La redondance de l'information contenue dans le signal de parole : un grand nombre d'indices portant sur l'identité des phonèmes sont disponibles au système auditif mais seul un petit nombre d'entre eux sont effectivement utilisés, i.e. déterminent directement la décision de l'auditeur.

- Le manque d'invariance des productions : un même phonème peut être produit de manière très différente selon la situation d'énonciation, le locuteur, le contexte phonémique, etc… Le mécanisme en jeu lors de la perception doit donc permettre d'associer tous ces signaux à un seul et même percept phonémique.

- La variabilité des catégories elles-mêmes : Les règles associant un signal physique à un phonème varient selon les situations, puisqu'un même son peut être identifié comme un phonème ou un autre lorsqu'il est présenté dans des contextes différents.

Face à ces limites du modèle alphabétique, les chercheurs ont dû élaborer de nouvelles théories de la perception de la parole dont le principe visait à conserver le mécanisme simple de comparaison du signal entrant à un gabarit en mémoire (*pattern-matching*), mais en déplaçant le niveau auquel cette comparaison est effectuée. Le système auditif opèrerait une transformation des sons de parole dans un espace différent de celui des amplitudes acoustiques, puis les confronterait avec les prototypes des phonèmes pour finalement sélectionner le phonème le plus proche dans cet espace. Ces questionnements sur la nature des représentations phonémiques seront abordés dans la partie 3.1. Néanmoins, comme nous le verrons, aucune des représentations envisagées ne fournit une explication totalement cohérente du phénomène de compensation pour la coarticulation. D'autres modèles de l'interface acoustico-phonétique, plus complexes, seront donc envisagés dans la partie 3.2.

3.1. Nature des représentations des phonèmes

Jusqu'ici nous avons employé les termes de « représentation phonémique » ou de « prototype » pour évoquer les informations phonologiques stockées dans la mémoire à long terme, sans détailler précisément ce qu'ils recouvrent. Une question se pose néanmoins concernant la nature de ces représentations : s'agit-il simplement d' « enregistrements » des phonèmes ou prennent-elles des formes plus complexes ? Deux grandes familles de théories se dégagent selon que les représentations des catégories phonémiques stockées en mémoire sont décrites dans un espace acoustique (théories auditives) ou articulatoire (théories motrices).

3.1.1. Théories motrices

Suite à leurs observations les chercheurs des Haskins Laboratories (et principalement Alvin Liberman et Ignatius Mattingly) mirent au point une théorie appelée Théorie Motrice de la Perception de la Parole (*Motor Theory of Speech Perception*). Constatant qu'ils ne parvenaient pas à identifier d'indices acoustiques corrélant de manière systématique avec la perception des phonèmes, ils entreprirent de chercher cette invariance au niveau articulatoire (Repp & Liberman, 1984). Par exemple, les sons /di/ et /du/, dont les spectrogrammes schématiques sont représentés Figure 8, sont tous les deux produits en bloquant momentanément le flux d'air au moyen de la pointe de la langue, positionnée au niveau des dents, puis en relâchant l'occlusion. Ainsi, si les signaux acoustiques résultants présentent de larges différences, les articulations qui leur donnent naissance sont identiques. Cette observation conduisit Liberman et Mattingly à formuler une proposition radicale à la base de la Théorie Motrice de la Perception de la Parole : nous ne percevons pas les sons de parole eux-mêmes mais seulement les gestes articulatoires qui les engendrent (Liberman et al., 1967).[6]

Bien que le son de parole soit évidemment transmis par le média acoustique (les vibrations de l'air), la théorie motrice de la perception de la parole suppose que l'objet distal perçu et manipulé par l'auditeur est en réalité l'ensemble des mouvements du conduit vocal du locuteur qui donnent naissance à ces signaux acoustiques (de la même manière qu'un observateur ne perçoit pas la lumière captée par sa rétine mais les objets sur lesquels cette lumière a été réfléchie) (Liberman & Whalen, 2000). Autrement dit, la parole pourrait être envisagée comme une suite de mouvements rendus audibles par

[6] Un autre argument avancé en faveur de la perception de l'articulation est que, lors d'une conversation face à face, certaines informations phonétiques sont transmises par des voies non acoustiques. L'effet McGurk en est un exemple célèbre : voir un interlocuteur produire une syllabe tout en écoutant l'enregistrement d'une syllabe différente peut affecter la perception de la syllabe entendue (Fowler & Rosenblum, 2014; Johnson, 2011; Olasagasti et al., 2015). L'aspect acoustique seul ne suffit donc pas à décrire la perception de la parole. Une explication plus satisfaisante serait donc qu'une intégration des informations audiovisuelles donne naissance à des représentations articulatoires.

l'addition d'un flux d'air. L'auditeur étant aussi un locuteur, Liberman émet l'hypothèse qu'il peut faire appel à sa propre expérience de la production de la parole pour reconnaître les mouvements articulatoires à l'origine du signal acoustique reçu. Cette opération serait effectuée par un module spécialisé du système moteur de l'auditeur qui simulerait les mouvements articulatoires du locuteur et comparerait le résultat prédit avec le signal effectivement perçu (analyse par synthèse) (Fowler & Rosenblum, 2014; Kuhl et al., 2014; Liberman et al., 1967). Ce recrutement des représentations du conduit vocal de l'auditeur pour la compréhension du son émis par le locuteur présuppose naturellement que le traitement de la parole est « spécial », impliquant des processus distincts des traitements acoustiques généraux (Fowler & Rosenblum, 2014; Liberman & Mattingly, 1985).

En 1985, après avoir constaté que les effets de coarticulation affectaient aussi les muscles articulateurs, remettant ainsi en cause l'invariabilité des gestes articulatoires, Liberman et Mattingly ont été amenés à réviser leur théorie. Ils proposèrent alors de considérer que l'objet de la perception n'est pas le geste articulatoire lui-même mais plutôt la commande neuromotrice à l'origine de ce geste articulatoire (Liberman & Mattingly, 1985).

Enfin, en 1986, une élève de Liberman, Carol Fowler, proposa une version moins contraignante de la Théorie Motrice, qualifiée de directe-réaliste (Fowler, 1986; Fowler & Rosenblum, 2014). Bien que partant également du principe que l'objet de la perception est le mouvement articulatoire et non le signal acoustique, cette explication ne suppose pas un accès à un module spécialisé du système moteur. En lieu et place du processus d'analyse par synthèse proposé par Liberman, Fowler fait plutôt appel à un mécanisme très général d'apprentissage perceptuel des régularités statistiques du signal de parole pour expliquer l'association entre le son perçu et les mouvements articulatoires qui lui ont donné naissance. Cette théorie est donc « directe » en cela que certaines informations acoustiques contenues dans le signal de parole activent directement des représentations motrices. Bien que le principe d'analyse par synthèse ait été largement battu en brèche par Fowler, il connut un regain d'intérêt à partir des années 90 avec la découverte des neurones miroir qui pourraient offrir un support cérébral idéal pour cette fonction (Hickok, 2014; Massaro & Chen, 2008).

3.1.2. Théories auditives

Parallèlement aux théories motrices décrites ci-dessus se développèrent des théories dites « auditives », refusant de faire appel au concept de représentation articulatoire, inessentiel selon elles pour expliquer la perception de la parole. Les chercheurs à l'origine ces théorie partaient de la conviction que certains invariants acoustiques existaient mais qu'ils n'étaient tout simplement pas directement visibles sur la représentation arbitraire du son qu'est le spectrogramme. Leur approche consista donc à rechercher une variable acoustique secondaire permettant de prédire la

perception d'un son de parole. Contrairement aux théories articulatoires qui proposent une hypothèse sur le fonctionnement global du système de perception de la parole, les théories auditives consistent – pour la plupart – en un ensemble de règles mathématiques parfois relativement abstraites, avec un domaine d'application limité (p.ex. invariant acoustique valable uniquement dans le cadre d'une catégorisation de consonnes occlusives), à l'exception peut-être de la Théorie Radiale mentionnée plus loin, qui relie perception des consonnes et des voyelles.

La théorie du locus, l'une des premières formalisations d'une théorie auditive, offre un bon exemple de cette approche. Pour reprendre le problème de la variabilité des spectrogrammes de /di/ et /du/, illustrée Figure 8, une première tentative d'explication (antérieure à celle de la Théorie Motrice exposée ci-dessus) fut formulée par Delattre et ses collaborateurs (Delattre et al., 1955). Si l'on considère les spectrogrammes des syllabes de forme /dV/ (/di/, /de/, /dɛ/, /da/, /də/, /do/, /du/) présentés Figure 9, on constate un point commun : toutes les transitions de F2 semblent provenir d'un même point imaginaire sur l'axe des fréquences situé approximativement à 1800 Hz. Cette origine commune est appelée le locus, et elle est supposée être invariante pour une consonne donnée, quelle que soit la syllabe qui la suit (on peut donc représenter la hauteur du locus comme une fonction constante de la fréquence du F2 de la voyelle, Figure 10). Si tel était le cas, le locus constituerait un indice acoustique fiable pour l'identification des consonnes. Les spectrogrammes des syllabes de forme /bV/ et /gV/ indiquent cependant que la caractérisation du locus n'est pas toujours aussi probante.

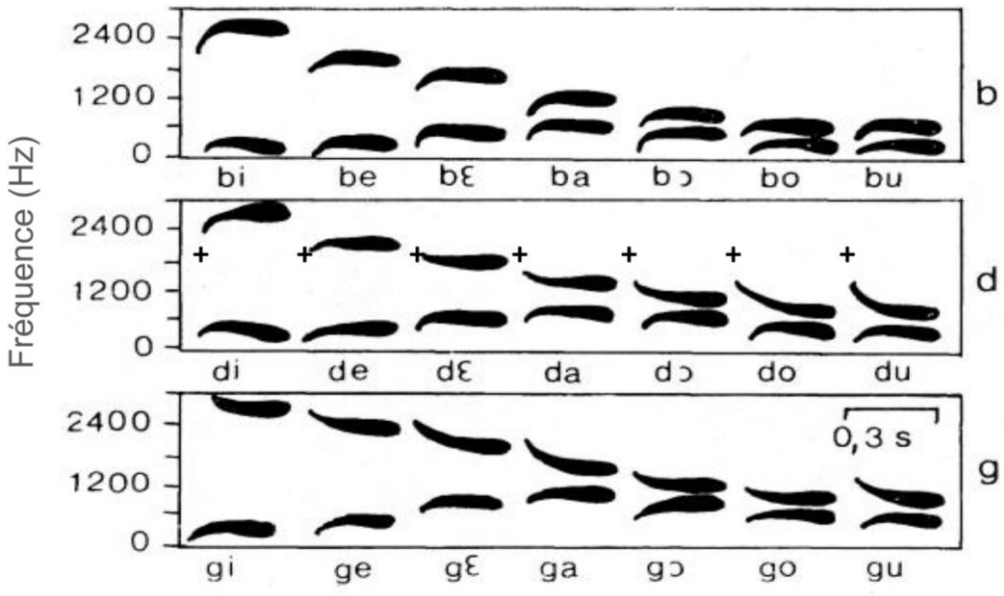


Figure 9 – Spectrogrammes schématiques des mouvements de F1 et F2 pour différentes syllabes CV. La croix marque la position supposée du locus pour chaque syllabe de forme /dV/. Adapté de (Peterfalvi, 1966).

D'autres formes de relations, correspondant à différentes équations du locus, ont donc été proposées pour tenter d'englober tous les cas de figure (Figure 10) : le modèle linéaire (Sussman et al., 1991), pour lequel la position du locus est en relation affine avec la hauteur du F2 de la voyelle, ou bien le modèle additif composé de segments de droites affines. L'invariant acoustique n'est alors plus le locus lui-même mais la relation linéaire entre l'attaque du F2 et sa partie stable. La théorie auditive la plus complète à notre connaissance est le Modèle Radial proposé par Serniclaes et Carré en 2002. Dans ce cas, les frontières phonémiques dans les différents contextes vocaliques convergent en un point du plan (partie stable de F2 / attaque de F2). Ce point correspond à la catégorisation pour la voyelle neutre /ə/. De manière intéressante ce dernier modèle établit un pont entre la perception des consonnes et celle des voyelles et, également, entre les théories articulatoires et auditives (Serniclaes, 2011).

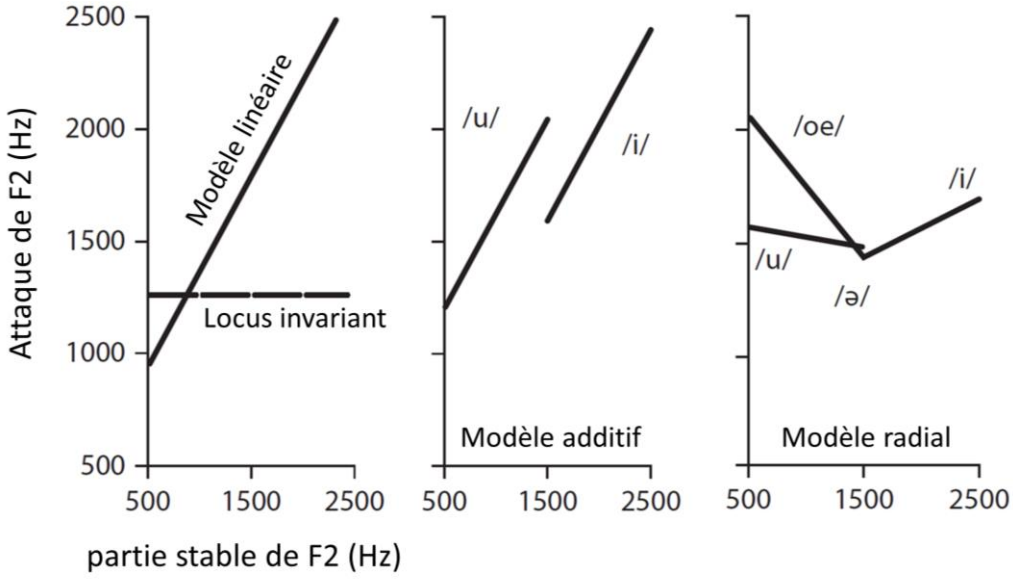


Figure 10 – Tracés de l'équation du locus pour différents modèles de perception de la place d'articulation. Représentations dans le plan (fréquence d'attaque de F2 / fréquence de la partie stable de F2). Adapté de (Serniclaes, 2011)**.**

3.1.3. L'exemple de la compensation pour la coarticulation

Un point de conflit majeur entre les théories motrices et auditives réside dans l'explication de la compensation pour la coarticulation. Depuis la mise en évidence de ce phénomène par Virginia Mann en 1980 (Mann, 1980), un vif échange d'arguments s'est engagé entre les théoriciens du direct-réalisme, Carol Fowler en tête, et les partisans d'une explication purement auditive par le mécanisme de contraste spectral. Un débat théorique s'ensuivit, qui n'est pas encore clos à ce jour (Fowler, 2006; Holt, 2005, 2006a, 2006b; Holt & Lotto, 2002; Kingston et al., 2011; Lotto, 2004; Lotto & Holt, 2006; Lotto &

Kluender, 1998; Lotto et al., 1997, 2003; Stephens & Holt, 2003; Viswanathan et al., 2009, 2010). Nous résumons brièvement ici les principaux arguments en faveur de l'un et l'autre de ces points de vue.

L'effet de compensation pour la coarticulation fut observé dès 1952 par Liberman et ses collaborateurs (Liberman et al., 1952). Néanmoins, la première étude approfondie du phénomène fut menée par Virginia Mann, à travers l'exemple de l'influence d'une consonne liquide sur la perception d'une consonne occlusive qui est devenu par la suite un cas d'école repris par grand nombre de chercheurs (Mann, 1980). L'étude se concentre sur la perception de 4 non-mots : « Alda », « Alga », « Arda » et « Arga ». Mann avait en effet noté que cet enchaînement de deux syllabes donnait lieu à une coarticulation importante en parole naturelle, l'articulation de la consonne continue ([l] ou [r]) se superposant à celle de la consonne occlusive ([d] ou [g]).

D'un point de vue articulatoire, le [d] est une consonne occlusive alvéolaire (réalisée avec la langue vers l'avant, au contact des alvéoles à la base des dents) tandis que le [g] est une consonne occlusive vélaire (l'occlusion étant située plus en arrière de la cavité buccale, avec la langue au contact du palais mou). De la même manière, le [l] (consonne alvéolaire) est produit avec la pointe de la langue vers l'avant tandis que le [r] anglais (consonne post-alvéolaire) est situé comparativement plus en arrière dans la cavité. L'enchaînement des consonnes /ld/ et /rg/ correspond donc à des mouvements cohérents des articulateurs. Au contraire, la production de /lg/ et /rd/ implique une transition rapide de la langue d'une position avant à une position arrière, ou inversement. La conséquence physique de ces lieux d'articulation antagonistes est un rapprochement des phonèmes de manière à réduire l'effort articulatoire : La consonne occlusive est réalisée avec la langue plus en avant lorsqu'elle est précédée de [l] que lorsqu'elle est précédée de [r]. Ainsi un /ga/, lorsqu'il est prononcé après /al/, a une place d'articulation plus antérieure (i.e. proche du /da/). Réciproquement un /da/ prononcé après /ar/ a une place d'articulation plus postérieure (i.e. proche du /ga/).

Pour mettre en lumière le déplacement des frontières phonétiques en fonction du contexte, Mann s'appuie elle aussi sur le paradigme du continuum de parole synthétique que nous avons déjà évoqué plus haut (paragraphe 2.1). Un ensemble de 7 stimuli synthétiques allant perceptuellement de /da/ à /ga/ est généré en variant la fréquence d'attaque du F3 par pas réguliers entre 2690 Hz et 2104 Hz (Figure 11A.). On désigne les éléments du continuum da-ga par le symbole /Ca/ (Consonne + /a/). 12 exemplaires de chacun de ces stimuli étaient présentés dans un ordre aléatoire. Les participants avaient pour consigne de catégoriser la consonne comme un /d/ ou un /g/.

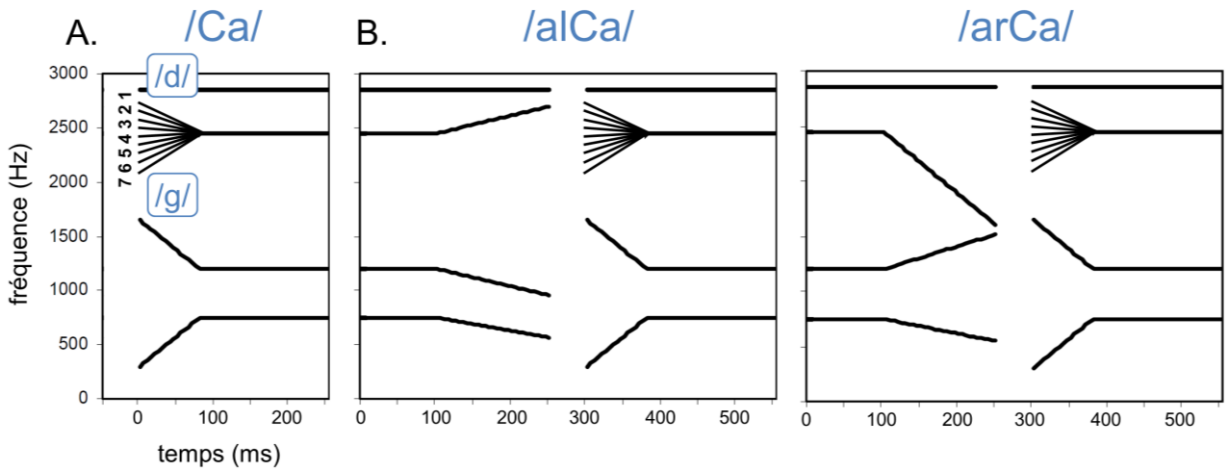


Figure 11 – Stimuli utilisés dans l'expérience de Mann. A. Continuum synthétique /da/-/ga/, avec les numéros des stimuli. B. Représentation schématique des stimuli /AlCa/ et /ArCa/. Adapté de (Holt & Lotto, 2002).

Comme pour l'expérience de Liberman, la représentation du pourcentage d'identification de « ga » le long du continuum fait clairement apparaître une frontière abrupte entre les deux catégories phonémiques située entre les stimuli #4 et #5 (trait pointillé sur la Figure 12).

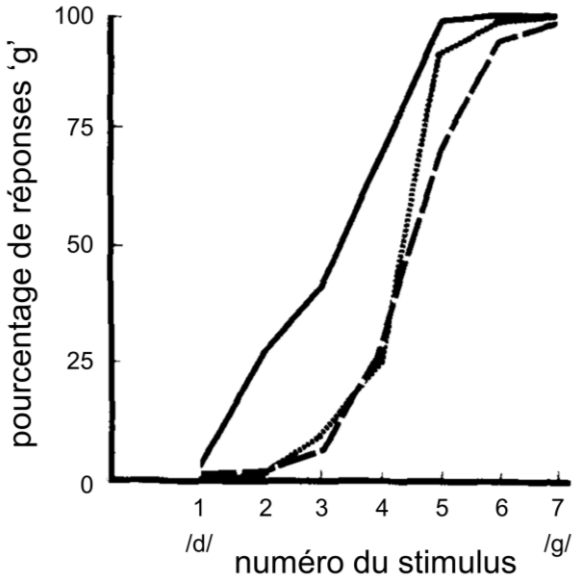


Figure 12 – Résultats de l'expérience de Mann. Pourcentage de réponses « ga » le long du continuum pour un stimulus synthétique seul (trait pointillé), ou précédé de /al/ (trait continu) ou de /ar/ (trait discontinu). Adapté de (Mann, 1980).

Le but de Mann étant de mettre en évidence un déplacement de cette frontière selon le contexte phonémique, elle reproduisit l'expérience en faisant cette fois précéder le stimulus synthétique de différentes productions naturelles de « Al » et de « Ar » (stimuli schématisés Figure 11B). À nouveau la visualisation de la proportion de réponses « ga » en fonction de la position du stimulus sur le continuum lui permit de localiser la frontière phonémique (Figure 12, traits continu et discontinu). Elle constata ainsi que le point d'inflexion des réponses des participants est globalement déplacé vers « da » en contexte « Al » (frontière située entre les stimuli #3 et #4) et vers « ga » en contexte « Ar » (frontière située entre les stimuli #4 et #5). Par conséquent le stimulus #4 est interprété majoritairement comme un « da » après « Ar », mais comme un « ga » après « Al ». Cette expérience apporte donc une preuve du déplacement de la frontière phonémique en fonction du contexte de la perception.

Par ailleurs, il s'agit bien d'une compensation perceptive pour le phénomène de coarticulation. En effet, comme nous l'avons évoqué plus haut, un /ga/ prononcé après /al/ possède une place d'articulation plus proche du /da/ et, réciproquement, un /da/ prononcé après /ar/ possède une place d'articulation plus proche du /ga/. Lors de la perception, on observe l'appariement opposé : en moyenne, les participants répondent majoritairement « ga » lorsqu'un stimulus débutant par /al/ leur est présenté, et majoritairement « da » pour un stimulus débutant par /ar/. Ce biais perceptif de l'auditeur en sens inverse du biais produit par le locuteur « compense » donc effectivement l'effet de coarticulation.

Dès ce premier article, Mann envisage deux explications également plausibles pour le phénomène qu'elle décrit.

(1) L'interprétation de la compensation pour la coarticulation dans le cadre des théories motrices est relativement directe (Fowler, 2006). Bien que la réalisation acoustique d'une même consonne diffère selon son contexte phonémique, les gestes articulatoires qui l'engendrent restent fondamentalement identiques (langue à l'arrière pour /g/, et à l'avant pour /d/). Dans le contexte des théories motrices, la présence d'une coarticulation, comme le recul du /d/ prononcé à la suite d'un /r/, ne constituera pas une source de confusion pour la perception du /d/ mais sera au contraire interprétée comme un indice supplémentaire pour la perception du /r/. Grâce aux représentations articulatoires partagées par le locuteur et l'auditeur, la perception s'opère en prenant en compte les contraintes articulatoires et dynamiques de la production de la parole.

(2) Une explication alternative peut être trouvée dans le contexte des théories auditives. Elle se fonde sur le principe du contraste spectral : la sensibilité du système auditif à une fréquence particulière est atténuée par la présence d'un son à cette même fréquence immédiatement avant. Par exemple, dans les stimuli de Mann présentés Figure 11B, la fin du F3 de « Al » coïncide avec l'attaque du F3 au début de la 2ème syllabe pour les sons proches de « da » dans le continuum. Au

contraire, la fin du F3 de « Ar » coïncide approximativement avec l'attaque du F3 au début de la 2ème syllabe pour les sons proches de « ga » dans le continuum. Ainsi, l'identification de « da » est-elle plus difficile en contexte /al/, tandis que l'identification de « ga » est plus difficile en contexte /ar/. Ce phénomène de masquage auditif explique donc lui aussi l'effet observé.

Dans les décennies qui suivirent, l'expérience de Mann fut reproduite à de nombreuses reprises avec diverses variantes du protocole expérimental. Il apparut alors qu'il était impossible de trancher en faveur de l'une ou l'autre interprétation, chacune étant invalidée par un certain nombre d'observations[7] :

- Le phénomène de compensation pour la coarticulation est également observé, à un degré moindre, pour des enfants nouveaux nés (Fowler et al., 1990) et des cailles du Japon (Lotto et al., 1997). Ceci est contraire à la prédiction des théories motrices, le comportement de ces sujets qui ne possèdent pas ou peu d'expérience des sons de parole ne pouvant être expliqué ni par un apprentissage des régularités statistiques dans le signal de parole, ni par une connaissance des mouvements des articulateurs dans le conduit vocal humain.

- Le résultat de Mann peut être reproduit en remplaçant la première syllabe par certains sons non paroliers [ton pur (Lotto, 2004), ton modulé en fréquence (Lotto & Kluender, 1998), série de tons purs formant une mélodie (Holt, 2005, 2006b), ou « négatif » acoustique d'un son (Coady et al., 2003)]. Réciproquement, la présence d'un contexte /al/ ou /ar/ module également la perception d'une cible non parolière (Stephens & Holt, 2003). Ces phénomènes ne trouvent pas d'explication dans le cadre de la théorie motrice puisque celle-ci ne s'applique qu'à des stimuli articulés.

- Il existe des effets rétroactifs de compensation pour la coarticulation : la présentation de la deuxième syllabe /da/ ou /ga/ produite en contexte /l/ ou /r/ biaise la catégorisation de la première syllabe en tant que /al/ ou /ar/ (Fowler, 2006). Au contraire, le contraste spectral soutenu par les partisans des théories auditives implique nécessairement un masquage dirigé vers l'avant.

- Enfin il est possible de moduler l'effet de compensation sans aucune variation des stimuli acoustiques. Lorsqu'un contexte ambigu, situé perceptuellement entre /al/ et /ar/, est utilisé, la présentation simultanée d'une vidéo du locuteur prononçant l'un ou l'autre contexte entraîne un biais correspondant en faveur de /da/ ou /ga/ dans l'identification de la 2ème syllabe (Fowler et al., 2000). Ce

[7] « Nous avons donc des affirmations absolument incompatibles de la part des partisans des théories motrice et auditive. […] Dans les deux cas, je suis sincèrement impressionné de la qualité des recherches mises en œuvre pour invalider la position opposée. Chacune des deux parties m'a convaincu que l'autre a tort. » (Nearey, 1997)

phénomène ne peut donc pas être intégré au cadre des théories auditives, purement acoustiques.[8]

3.2. Modèles de l'interface acoustico-phonétique

Les théories de la perception résumées ci-dessus représentent une amorce d'explication aux phénomènes observés dans les expériences de psychoacoustique, tels que la compensation pour la coarticulation. Elles sont sous-tendues par un modèle simple : après la réception du son de parole, une série de traitements lui sont appliqués pour en fournir une représentation dans un espace pouvant être soit articulatoire (théories motrices) soit également acoustique (théories auditives). La distance relativement aux représentations « idéales » des phonèmes dans cet espace est évaluée et le phonème apparaissant le plus proche est sélectionné. Ce déplacement du *pattern-matching* dans un niveau de représentation supérieur permet de rendre partiellement compte des phénomènes d'intrication, de coarticulation et de manque d'invariance lors de la perception du son de parole. Ces théories apparaissent cependant limitées au vu de l'exemple de la compensation pour la coarticulation.

Plus important encore dans la perspective de cette thèse, elles présentent le désavantage de ne pas décrire avec suffisamment de précision la première étape du traitement effectué au niveau de l'interface acoustico-phonétique. En effet, que les représentations des phonèmes soient articulatoires ou acoustiques, leur activation implique dans un premier temps l'extraction de certaines informations véhiculées par le medium acoustique. Quelle que soit la théorie adoptée, le questionnement initial demeure : où l'information linguistique se situe-t-elle dans le signal de parole ? Étant données deux productions de phonèmes, sur quelles différences acoustiques le système auditif s'appuie-t-il pour les distinguer ? Bien évidemment, un certain nombre de caractéristiques acoustiques essentielles pour la catégorisation phonémique ont été mises en évidence pour tenter d'invalider l'une ou l'autre théorie (comme la trajectoire du F3 dans l'expérience de Mann, paragraphe 3.1.3). Néanmoins, le débat qui oppose les partisans des théories motrices et ceux des théories auditives porte davantage sur la nature des représentations manipulées que sur les primitives mêmes qui activent ces représentations.

Pour répondre plus précisément à la question de l'extraction des primitives de la parole, il est nécessaire à ce stade d'expliciter de manière plus détaillée les mécanismes potentiellement à l'œuvre lors de cette étape. Celle-ci peut être envisagée comme une transmission de l'information depuis le domaine physique (ou domaine

[8] Suite à la constatation des limites des représentations motrices et auditives, une «Théorie de la Perception pour le Contrôle de l'Action» (*Perception for Action Control Theory*, PACT) a été proposée, qui se base sur des représentations perceptuo-motrices (gestes articulatoires modelés par la perception) (Schwartz et al., 2012)

acoustique) vers le domaine perceptuel, durant laquelle le système auditif convertit un signal acoustique et analogique en percept phonémique discret. Pour le psycholinguiste, l'interface acoustico-phonétique se présente comme une « boîte noire » réalisant un certain nombre de traitements à déterminer.

Dans la suite de ce chapitre, nous examinerons différents modèles qui ont été avancés pour décrire l'interface acoustico-phonétique et qui se focalisent sur la phase d'extraction des primitives. De ce fait, les modèles psycholinguistiques pour lesquels l'entrée ne correspond pas au signal acoustique brut mais directement à un niveau supérieur comme une succession de phonèmes [par exemple Shortlist A (Norris et al., 2000)], de probabilités d'identifications de phonèmes [Shortlist B (Norris & McQueen, 2008)] ou encore d'activation des traits phonétiques [par exemple TRACE (McClelland & Elman, 1986)], n'entrent pas dans le champ de la présente thèse. Seules les modélisations de la chaîne de traitement complète, à partir du niveau acoustique, seront donc examinées ici. Ceci nous permettra d'entrevoir, dans un second temps, les principes généraux de l'organisation de l'interface acoustico-phonétique.

3.2.1. Modèle LAFS

Le modèle LAFS (*Lexical Access From Spectra*), initialement conçu comme un système de reconnaissance automatique de la parole (système Harpy), suppose que le système auditif opère par une analyse au fil du temps du spectre du signal entrant (Klatt, 1979). L'analyse est effectuée au moyen d'un automate fini dont les états correspondent aux spectres successifs. Le graphe orienté à la base de ce modèle est obtenu en considérant l'automate fini des enchaînements d'éléments phonétiques – qui décrit les règles phonotactiques et phonologiques du langage – et en y remplaçant chaque transition entre deux états (i.e. entre deux réalisations de phonèmes) par la séquence de spectres nécessaire pour passer de l'un à l'autre. Le résultat est un nouveau graphe orienté encodant l'ensemble des enchaînements possibles de spectres acoustiques (Figure 13), qui forme alors un dictionnaire exhaustif de tous les enchaînements spectraux possibles. Lors de la perception d'un son de parole, ce graphe est alimenté en entrée par les spectres correspondant aux instants successifs (avec un pas de 10 ms). Le chemin obtenu en sortie du graphe correspond au décodage de la suite de phonèmes.

Le modèle LAFS fait partie des modèles dits « à exemplaires ». Aucune étape de détection d'indices acoustiques n'est nécessaire. En effet, la solution adoptée ici pour contourner le problème de l'absence d'invariants dans les productions de parole consiste à stocker un grand nombre d'exemplaires de chaque phonème et transition de phonèmes en mémoire, couvrant ainsi la variabilité observée dans le signal. Les effets de la coarticulation sont ainsi directement « incorporés » dans le graphe et ne constituent donc pas un obstacle au décodage mais, au contraire, une information supplémentaire pour l'identification des phonèmes.

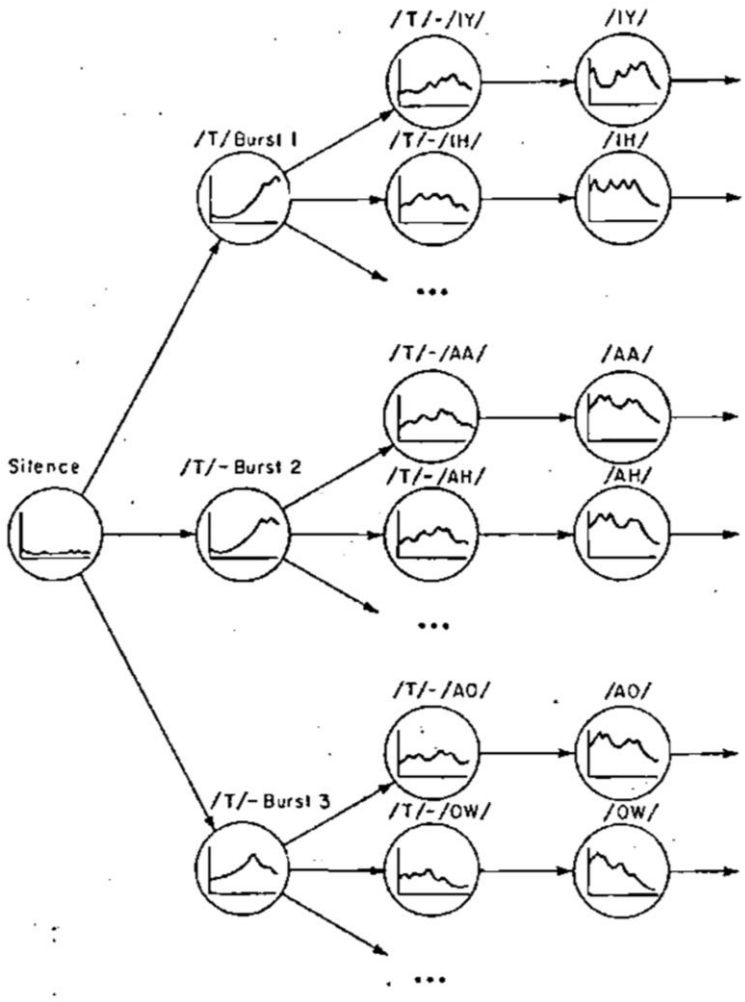


Figure 13 – Portion du graphe de l'automate fini décrit par le modèle SCRIBER (première version du modèle LAFS). À chaque état (cercle) est associé un spectre représenté à l'intérieur, et un contenu phonémique indiqué au-dessus. Les flèches indiquent les transitions possibles entre les différents spectres. Adapté de (Klatt, 1979).

La simplicité conceptuelle du modèle a malheureusement pour prix un système de règles extrêmement lourd et complexe et il est probable que le graphe encodant la totalité des enchaînements de spectres présents dans le français soit trop important pour la taille de la mémoire humaine.[9] Par ailleurs, le modèle LAFS s'avère incapable de rendre compte de la capacité de généralisation immédiate observée par exemple lors d'une discussion avec un locuteur inconnu. Il serait en effet contraint dans ce cas de construire un nouveau graphe décrivant tous les spectres utilisés par ce locuteur, ce qui nécessiterait une exposition préalable à un grand corpus de sons prononcés par ce locuteur.

[9] Cependant, il est à noter que d'autres modèles à exemplaires de la perception de la parole ont été proposés, basés sur un principe de stockage plus économique en termes de ressource mémorielle (Goldinger, 1998)

3.2.2. Modèle LAFF

Selon le modèle LAFF (*Lexical Access From Features*), les unités pré-phonémiques activées par les informations sensorielles ne seraient autres que les traits distinctifs décrits par la phonologie (Stevens, 2002). Il distingue deux types d'indices acoustiques jouant des rôles différents dans le processus : les repères acoustiques qui traduisent l'état du flux d'air dans le conduit vocal et les indices acoustiques secondaires qui indiquent la position des articulateurs. Pour conclure quant à la présence ou l'absence d'un trait donné, ce modèle traite le signal entrant en trois temps :

(1) L'étape de segmentation consiste à détecter les repères acoustiques qui marquent le mode d'écoulement de l'air. Ils indiquent la présence d'une consonne (constriction ou relâchement du conduit vocal) ou d'une voyelle (résonnances dans la cavité buccale). Ils constituent les indices acoustiques de premier type et se traduisent par des discontinuités ou des maxima de deux caractéristiques acoustiques particulières : la fréquence et l'amplitude du F1. Par une simple analyse spectrale du son, le système auditif est donc à même de repérer les positions temporelles des phonèmes et leur type (voyelle ou consonne). Il peut en outre déterminer certains traits distinctifs correspondant seulement à l'état du flux d'air et indépendants des articulateurs en action (par exemple, le trait de discontinuité identifiant les occlusives).

(2) Une fois ces repères acoustiques établis, et les segments de parole correspondant aux voyelles et consonnes délimités, des indices acoustiques secondaires sont extraits. Au contraire des premiers, ceux-ci renseignent sur les articulateurs mis en jeu à un moment donné (protrusion des lèvres, langue en position antérieure, etc…). Pour cela, le système mesure les valeurs de certains paramètres au voisinage des repères acoustiques et en fonction du type de phonème déterminé à la première étape (seules les informations utiles sont extraites). Les indices acoustiques secondaires incluent en particulier : les fréquences des formants, leur largeur de bande, l'énergie dans les basses fréquences, la fréquence fondamentale, la forme du spectre et l'intervalle de temps entre deux repères acoustiques.

(3) Enfin, les indices acoustiques primaires et secondaires sont combinés au moyen d'un ensemble de règles[10] permettant d'évaluer les traits distinctifs qui dépendent des articulateurs et permettront de déterminer l'identité du phonème. Cette étape est réalisée par des modules indépendants spécialisés dans la détection de chaque trait, prenant en entrée les indices acoustiques pertinents pour ce trait, ainsi que le

[10] Comme pour le modèle suivant, le processus de combinaison des indices n'est pas détaillé par les auteurs. On peut cependant penser à un mécanisme de décision du type de celui proposé par la Théorie de Combinaison des Indices (*Cue Combination Theory*) (Treisman, 1999)

contexte phonétique et prosodique, et restituant en sortie l'information concernant la présence ou l'absence du trait.[11]

L'implémentation de ce modèle nécessite la programmation de chacun des modules correspondant à la détection d'un trait (Slifka, 2005). Elle conduit à des résultats variables selon le type de phonèmes (avec plus de 90% de reconnaissances correctes pour les occlusives).

3.2.3. Modèle des canaux indépendants

Dès 1922, Fletcher remarqua que plusieurs indices étaient extraits de manière indépendante dans différentes bandes de fréquence (Allen, 1996; Fletcher, 1922). En effet, lorsqu'on applique un filtre passe-haut au signal de parole pour ne conserver que l'énergie dans les fréquences supérieures à 1550 Hz, le pourcentage de reconnaissances correctes sur des enregistrements de syllabes atteint 65% en moyenne. Le même pourcentage est obtenu lorsqu'on applique un filtre passe-bas à ce même signal pour ne conserver que l'énergie dans les fréquences inférieures à 1550 Hz. Par suite, la somme des pourcentages de reconnaissances correctes pour des stimuli paroliers traités par des filtres passe-haut et passe-bas complémentaires étant supérieure à 100%, on peut en déduire que, pour un certain nombre de syllabes, l'information sur l'identité du stimulus peut être obtenue indifféremment dans les parties supérieure ou inférieure du spectre.

Fletcher ne s'arrêta pas là et parvint à exprimer une relation entre les pourcentages d'erreur lors de la compréhension de la parole limitée à la bande haute seule (e_{hf}) limitée à la bande basse seule (e_{bf}) et pour le signal non filtré (e), quelle que soit la fréquence de coupure utilisée :

$$e = e_{hf} \cdot e_{bf} \qquad 3.1$$

L'auditeur détecte donc les indices acoustiques indépendamment dans chacune des deux bandes : les erreurs portant sur les indices haute-fréquence n'affectent pas la détection des indices basse-fréquence. Plus généralement, pour un découpage du signal en K bandes de fréquence, on observe la même indépendance des canaux et le taux d'erreur e est égal au produit des taux d'erreur des K signaux filtrés (Toscano & Allen, 2014). French et Steinberg proposèrent donc de diviser l'axe des fréquences en 20 bandes critiques contribuant de manière égale à la reconnaissance (c'est-à-dire que les taux d'erreurs moyens dans chaque bande sont égaux). Les intervalles de fréquence

[11] Le problème de la coarticulation se pose à ce niveau. Cependant, d'après Stevens, l'effet du contexte sur le processus d'identification est très limité. La coarticulation se traduirait uniquement par l'apparition d'indices acoustiques supplémentaires. Chaque trait distinctif possèderait donc un certain nombre de corrélats acoustiques qui le définissent auxquels s'ajouteraient dans le signal d'autres événements acoustiques non signifiants, conséquences des contraintes physiques sur les articulateurs.

ainsi déterminés possèdent des largeurs inégales en Hertz, mais représentent chacun une distance d'environ 1 mm le long de la membrane basilaire.

De ces observations découle logiquement une mesure d'intelligibilité pour un signal masqué par un bruit stationnaire, appelée indice d'articulation (AI, Articulation Index). Elle consiste à découper le signal dans les 20 bandes critiques, puis à estimer indépendamment le pourcentage d'erreur e_k dans chaque bande k en fonction du Rapport Signal sur Bruit (SNR, *Signal to Noise Ratio*) dans cette bande. Le pourcentage d'erreur pour le signal global est alors donné simplement par le produit de ces erreurs, comme nous l'avons vu ci-dessus. Malgré son apparente simplicité, l'indice d'articulation offre toutefois une très bonne prédiction de l'intelligibilité dans un grand nombre de conditions (Lobdell, 2009).

Cet ensemble de résultat fut synthétisé par Jont Allen sous la forme d'un modèle de fonctionnement de l'interface acoustico-phonétique (Allen, 1994). Il possède 4 niveaux, séparés par 3 traitements successifs (Figure 14). Deux couches ultérieures correspondant aux niveaux syllabique et lexical sont également mentionnées dans le modèle original mais, ne concernant pas directement l'objet de cette étude, elles ne seront pas détaillées ici.

(1) Avant toute analyse linguistique, le signal entrant est décomposé par la banque de filtres de la cochlée. Il y a approximativement 4000 cellules ciliées le long de la membrane basilaire dont les bandes passantes se superposent en grande partie (Allen, 2008). Des opérations non linéaires modélisant les traitements effectués au niveau du nerf auditif peuvent être également utilisées à ce stade. Pour obtenir une estimation de la fiabilité des informations provenant de chaque bande, le système effectue de plus une évaluation du niveau de bruit associé. Seuls les événements acoustiques d'amplitude trop importante pour appartenir au bruit seront retenus par la suite. La représentation temps-fréquence du résultat de cette première étape, appelée AI-gram, constitue une visualisation graphique des composantes du son perceptibles par le système auditif central (Lobdell, 2009; Lobdell et al., 2011; Régnier & Allen, 2008).

(2) Cependant, toutes les composantes audibles du signal de parole ne sont pas nécessairement utilisées pour la compréhension. Dans la deuxième étape, les signaux filtrés sont regroupés dans les K bandes critiques, puis différents groupes d'indices acoustiques localisés en fréquence sont extraits en parallèle pour chacune des K bandes, et convertis en événements (chaque événement correspondant simplement à une valeur booléenne indiquant la présence ou l'absence d'un indice particulier dans le signal). Cette information peut être exacte, dans le cas d'une communication claire, ou erronée si la communication est mauvaise.

(3) Les résultats de ces K processus indépendants sont finalement fusionnés en une seule estimation de l'identité du phonème. Ceci implique de combiner les indices et d'identifier puis de résoudre les conflits potentiels. Le processus par lequel une unité

phonémique est activée plutôt qu'une autre n'est pas détaillé par les tenants de ce modèle. Néanmoins, les travaux ultérieurs de l'équipe de Jont Allen sur le sujet (voir le paragraphe 4.5) suggèrent que les mécanismes mis en jeu durant cette étape sont relativement simples : un phonème est clairement défini par un petit nombre d'indices acoustiques. Au contraire, une mauvaise production d'un phonème peut contenir des indices conflictuels qui conduiront à une identification erronée du stimulus.

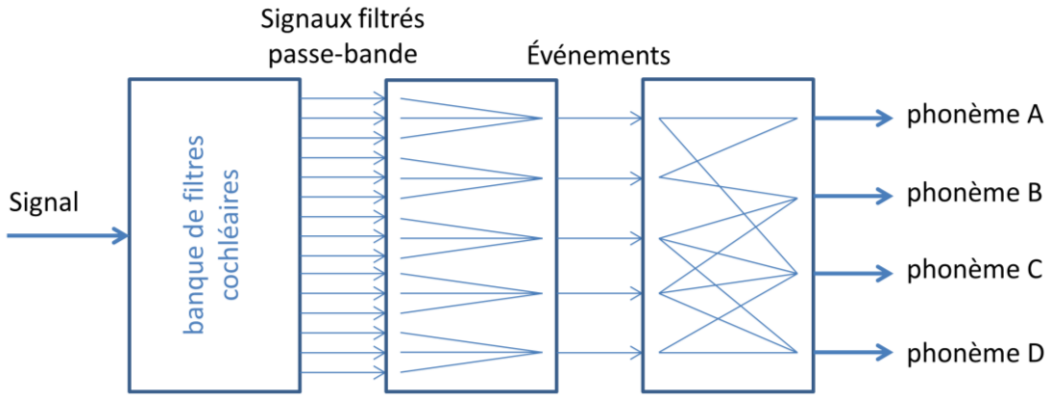


Figure 14 – Représentation schématique du modèle de l'interface acoustico-phonétique en canaux indépendants proposé par Jont Allen (Allen, 1994). Il comprend trois étapes : 1. Application des filtres cochléaires ; 2. Regroupement en bandes critiques et extraction des événements (ici, pour des questions de lisibilité, on pose K=5) ; 3. Combinaison des événements pour l'identification du phonème.

Le modèle lié à l'indice d'articulation est appelé multi-flux (*multi-stream*), car il suppose que les phonèmes sont traités dans plusieurs canaux fréquentiels indépendants pour obtenir autant d'informations partielles finalement unifiées en une seule interprétation du son. L'avantage de ce découpage tient à ce que, dans la mesure où certains canaux transmettent une information fiable (c.à.d. contiennent peu de bruit), l'information provenant des autres canaux potentiellement corrompus n'a pas besoin d'être considérée.

3.2.4. Modèle de la représentation multi-résolution

Chi et ses collaborateurs présentèrent en 1999 un modèle du traitement auditif biologiquement inspiré (Chi et al., 1999). Il est basé sur l'architecture de la voie auditive primaire (qui sera présentée au chapitre 6). Au lieu de se limiter à une représentation simulant les processus dans le système auditif périphérique, comme c'était le cas dans le modèle précédent, les chercheurs intégrèrent des traitements plus « haut-niveau » correspondant au système auditif central. Ainsi ce modèle comporte-t-il deux étapes de transformation du signal acoustique :

1) Une analyse temps-fréquence permettant d'obtenir un spectrogramme auditif (ou cochléogramme). Comme dans la première étape du modèle précédent, le signal est d'abord découpé en bandes de fréquences par la banque des filtres cochléaires puis plusieurs traitements non linéaires sont appliqués afin de simuler les caractéristiques des réponses des fibres du nerf auditif.

2) Dans un second temps, ce spectrogramme auditif est analysé par une seconde banque de filtres, plus complexe modélisant les champs récepteurs du cortex auditif primaire. Contrairement à la première étape, la seule réponse fréquentielle n'est pas suffisante pour caractériser ces neurones et il faut alors employer des réponses impulsionnelles temps-fréquence appelés champs récepteurs temps-fréquence (*Spectro-Temporal Receptive Fields*, STRFs). Comme nous l'expliquerons en détail dans la partie 6.1, un STRF est classiquement décrit par un filtre de Gabor possédant 3 paramètres : fréquence, échelle temporelle et échelle spectrale. La représentation neuronale obtenue possède donc 3 dimensions et elle est supposée reproduire les caractéristiques du signal encodées au niveau du cortex auditif. Elle est multi-résolution en temps et en fréquence, grâce aux échelles temporelle et spectrale qui caractérisent la rapidité des modulations du signal selon les deux dimensions.

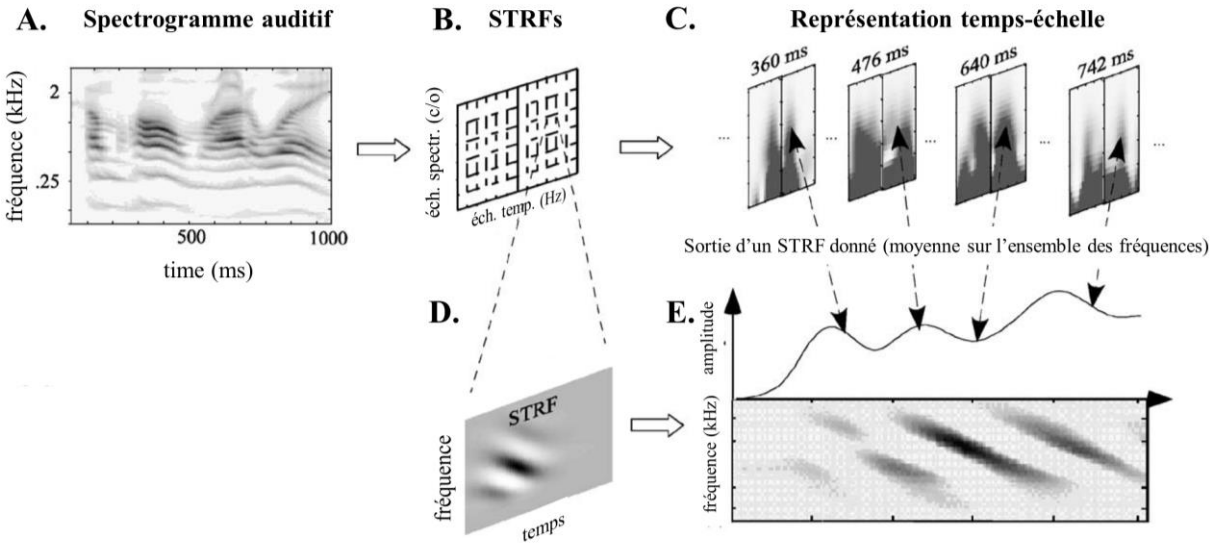


Figure 15 – Représentation schématique du calcul de la représentation multi-résolution pour la phrase « Come home right away ». A. Spectrogramme auditif du stimulus. B. Banque de filtres STRFs. C. Représentation corticale (temps-échelle). Chaque tranche correspond à une visualisation, à un instant donné, du signal dans l'espace échelle temporelle – échelle spectrale. D. Exemple d'un filtre STRF particulier (pour une valeur donnée de modulation temporelle et de modulation spectrale). E. Activation au cours du temps de ce filtre STRF, positionné à différentes fréquences, suite à la présentation du stimulus. Adapté de (Elhilali et al., 2003)

L'indice d'intelligibilité qui résulte de la représentation en scalogramme (représentation temps-échelle) est appelée indice de modulation spectro-temporel (STMI, *Spectro-Temporal Modulation Index*) (Elhilali et al., 2003). Dans le principe, il est assez proche de l'indice d'articulation ; cependant, tandis que l'AI est purement spectral (reflétant la sensibilité à des dégradations dues à l'addition d'un bruit stationnaire), le STMI rend également compte des problèmes dus à des dégradations temporelles (par exemple une réverbération). La corrélation entre cette mesure physique et le pourcentage de réponses correctes dans une tâche de compréhension de phrases dans le bruit est excellente (Chi et al., 1999). La fidélité du signal encodé par la banque de STRF a également été validée par un jugement d'intelligibilité : les auditeurs obtiennent près de 90 % de phonèmes reconnus à l'écoute du signal reconstruit (Chi et al., 2005).

Dans la perspective d'un modèle de compréhension de la parole, la représentation multi-résolution présente un avantage décisif : les différentes propriétés d'un son sont encodées de manière indépendante les unes des autres. Par exemple, un même groupe de neurones sera activé en réponse à la présentation de la voyelle /a/, quelle que soit la fréquence fondamentale du son. Cette capacité à extraire une caractéristique acoustique du stimulus en généralisant selon les autres dimensions est cruciale pour la compréhension de la parole comme nous l'avons vu dans la première partie. Au final, l'identification d'un phonème particulier correspond simplement à l'activation d'un groupe spécifique de STRFs.

Des versions plus avancées de cet algorithme de reconnaissance basées sur la représentation multi-résolution furent proposées par Mesgarani, Thomas et Hernansky (Mesgarani et al., 2010, 2011a, 2011b). Exploitant le fait que l'ajout de bruit ne dégrade pas cette représentation de manière uniforme, les chercheurs proposèrent de considérer des canaux indépendants composés d'un ensemble de STRF. Comme dans le modèle multi-flux précédent (partie 3.2.3), les canaux transmettant une information peu fiable sont identifiés et écartés par le système puis l'identité du phonème est établie uniquement sur la base des canaux restants.

3.2.5. Conclusions sur la structure de l'interface acoustico-phonétique

Les 4 modèles de l'interface acoustico-phonétique présentés ici furent mis au point pour tenter de décrire le processus par lequel les informations sensorielles qui parviennent au système auditif sont appariées avec les représentations phonémiques stockées en mémoire. Bien qu'ils soient en apparence relativement différents les uns des autres, il est intéressant à ce stade de noter certaines similarités :

- Tous les modèles sont purement ascendants (*bottom-up*). Aucune information provenant des niveaux supérieurs n'est utilisée par les niveaux inférieurs (pas de mécanismes de *feed-back*). Il s'agit d'une hypothèse forte qui ne se retrouve pas dans un certain nombre de modèles plus haut-niveau (notamment TRACE) selon

lesquels, par exemple, l'activation d'une unité lexicale peut renforcer l'activation d'une unité phonémique.

- Les modèles 2, 3 et 4 supposent l'existence d'une étape d'extraction d'informations acoustiques entraînant l'activation d'unités intermédiaires pré-phonémiques. Pour les modèles 2 et 3, cette phase est explicitement décrite. Pour le modèle 4 l'extraction correspond à l'excitation d'un filtre particulier (ou d'un ensemble de filtres) dans la banque de STRFs. En revanche, le modèle LAFS ne possède pas d'unité pré-phonémique à proprement parler, ce qui est un obstacle pour la généralisation à d'autres locuteurs, comme nous l'avons fait remarquer plus haut.

- De plus, les modèles 2, 3 et 4 ont en commun une phase de combinaison des indices acoustiques. À partir de l'ensemble des événements acoustiques détectés ou non et de leur organisation temporelle, une représentation de niveau supérieur (phonèmes, traits…) est activée, d'après un système de règles plus ou moins complexe. Comme nous l'avons noté plus haut l'absence de cette étape dans le modèle LAFS se solde par une taille démesurée du modèle résultant.

En remarquant les similarités entre les modèles 2, 3 et 4, une structure générale semble donc se dessiner pour l'interface acoustico-phonétique. Cette dernière prend la forme d'un canal de traitement entre le niveau physique (signal analogique en entrée) et le niveau perceptuel (état discret en sortie). Le canal est ici toujours composé d'au moins deux processus : l'extraction et la combinaison des primitives acoustiques. Ces primitives forment une description robuste et compacte du signal de parole, en préservant toute l'information pertinente pour la compréhension et en éliminant les aspects redondants et variables du signal. Il s'agit d'une étape cruciale puisqu'elle permet la conversion du signal continu en un ensemble de données discrètes (présence ou absence des différents indices). À ce stade, nous sommes amenés à nous demander quels sont précisément les indices acoustiques mis en jeu lors de la compréhension d'un signal de parole.

4. Méthodes psychoacoustiques pour l'identification des indices acoustiques de la parole

Le décodage d'un son de parole engage nécessairement deux processus distincts : la détection puis la combinaison des indices acoustiques. L'identification des informations contenues dans le signal de parole qui sont effectivement détectées et traitées par le système auditif est une question complexe. Comme nous l'avons vu au paragraphe 2.1, Alvin Liberman entama dès 1952 la première série d'expériences psychoacoustiques pour mettre à jour les indices acoustiques utilisés lors de la perception (Liberman et al., 1952). Aujourd'hui encore, cette question reste ouverte et plusieurs groupes de recherche abordent ce problème avec des angles d'approches très divers.

Dans le cadre de notre formulation « en canal » du problème (décrite à la fin de la partie précédente) composée d'un mécanisme inconnu interfaçant l'espace physique et l'espace perceptuel, deux procédés d'investigation peuvent être envisagés pour l'identification des primitives acoustiques de la parole :

- En faisant varier le signal en entrée et en observant la variation – ou l'absence de variation – du percept en sortie, les chercheurs peuvent espérer tirer des conclusions sur les mécanismes à l'œuvre au niveau de l'interface.

- Une autre solution consiste à présenter les signaux dans une condition dégradée (par exemple par l'ajout de bruit). Les confusions pouvant ou non en résulter sont alors une manifestation des traitements appliqués par l'auditeur, et nous renseignent donc indirectement sur les aspects du signal jouant un rôle dans sa catégorisation.

Ces deux voies seront empruntées parallèlement pour tenter de répondre à la question des primitives de la parole. Dans la suite de ce chapitre nous essaierons de dresser une généalogie des différentes méthodes successivement développées pour l'identification des indices acoustiques.

4.1. Le paradigme du continuum de parole synthétique

Nous avons déjà décrit en détail (cf. paragraphe 2.1) l'une des premières études employant un continuum de parole synthétique pour explorer la catégorisation des consonnes /b/, /d/ et /g/ (Liberman et al., 1957). La trajectoire du F2, seul élément variable entre les différents stimuli, suffisait à les identifier comme des exemplaires de /ba/, /da/, ou /ga/, ce qui semble indiquer que cet élément joue un rôle prépondérant dans la catégorisation des occlusives voisées.

Cette méthode fut appliquée de manière systématique par les chercheurs des Haskins Laboratories pour identifier les frontières phonétiques mises en jeu lors de la catégorisation de différentes classes de consonnes, parmi lesquelles les occlusives non voisées (Liberman et al., 1952) et voisées (Liberman et al., 1957), les fricatives voisée (Delattre et al., 1964), les nasales (Liberman et al., 1954) et certaines consonnes continues (O'Connor et al., 1957). Les expérimentateurs se plaçaient toujours dans le cadre d'un complexe CV avec différentes voyelles ce qui leur a permis d'envisager l'effet de la coarticulation sur la catégorisation de la consonne.

Bien évidemment, la mise en pratique de cette approche nécessitait de développer au préalable un outil propre à générer des sons de parole synthétique, de façon à contrôler rigoureusement les informations acoustiques contenues dans les stimuli. Aussi les chercheurs des Haskins Laboratories mirent-ils au point un dispositif appelé *Pattern Playback* permettant de transformer un spectrogramme schématique (composé uniquement du tracé des formants), peint à la main sur un support transparent, en un son de parole synthétique (Cooper et al., 1951; Liberman et al., 1952). De cette manière, n'importe quel ensemble de trajectoires formantiques pouvait être « joué » par l'appareil.

Dans chacune de ces études le paradigme expérimental employé comportait trois phases :

1) Formuler une hypothèse sur le ou les indices acoustiques impliqués dans la catégorisation entre les phonèmes A et B, sur la base d'une comparaison approfondie de spectrogrammes d'enregistrements naturels de ces phonèmes [voir par exemple (Liberman et al., 1952)].

2) Générer un continuum synthétique variant par pas réguliers le long de la ou des dimensions identifiées précédemment, avec le premier élément du continuum identifié sans ambiguïté comme A, et le dernier comme B.

3) Demander à un certain nombre d'auditeurs de catégoriser chacun des stimuli synthétiques obtenus, présentés dans un ordre aléatoire, comme une instance du phonème A ou du phonème B. La représentation du pourcentage de catégorisations le long du continuum permettait alors de vérifier la présence d'une rupture brutale entre les deux catégories, d'une part, et de localiser la position de la frontière phonémique au point du continuum où les deux réponses étaient données en proportions égales d'autre part.

Cette procédure en trois temps fut utilisée avec succès pour des espaces à une ou deux dimensions composés de complexes CV avec une voyelle constante. Lorsque le continuum varie selon deux caractéristiques (par exemple attaque de F2 et attaque de F3), le résultat est figuré sous la forme d'une représentation bidimensionnelle, et les frontières phonémiques apparaissent comme des contours délimitant les régions du

plan associées à une consonne ou à une autre (Liberman et al., 1954). Bien que la méthode employée soit différente, ces résultats peuvent être rapprochés du Triangle de Delattre (Figure 7), élaboré à la même période, qui localise les voyelles du français dans l'espace des valeurs de F1 et F2 (Delattre, 1948; Delattre et al., 1952).

Si la méthode du continuum de parole synthétique s'avère relativement efficace pour localiser une frontière phonémique sur une dimension donnée, une limite évidente concerne la première phase de mise en place de l'expérimentation : la recherche des caractéristiques cruciales du signal de parole. Cette étape est non seulement laborieuse et empirique (car elle demande d'examiner un grand nombre de productions d'un phonème de manière à identifier leurs points communs, puis de générer un son de parole « simplifié » en supprimant progressivement certaines composantes du son par la méthode des essais et des erreurs), mais elle est également biaisée. En effet il est envisageable qu'une dimension cruciale du son, peu apparente sur le spectrogramme, échappe à l'œil de l'expérimentateur durant cette phase, auquel cas la frontière phonémique ne sera pas recherchée sur le continuum approprié. En d'autres termes, l'expérience nécessite une connaissance a priori des indices acoustiques mis en jeu.

Une autre limite sérieuse de cette approche concerne le caractère peu naturel des stimuli synthétiques générés par le *Pattern Playback*, qui se traduit souvent par des taux d'identifications correctes inférieurs à 100 % même dans des conditions d'écoute optimales (Cooper et al., 1951; Li et al., 2010). La parole ainsi « simplifiée », réduite à un petit nombre de formants, s'avère donc artificielle et incomplète, ce qui indique que certaines informations acoustiques essentielles sont manquantes. Étant donné que le système auditif fait habituellement preuve d'une grande plasticité lui permettant de s'adapter à des conditions d'écoute très diverses (Shannon et al., 1995), il est permis de mettre en doute l'authenticité des indices acoustiques révélés par la méthode : ces derniers pourraient même se trouver relativement éloignés des indices acoustiques réellement utilisés lors de la compréhension de la parole naturelle. Cette seconde contrainte concernant la non-naturalité des stimuli fut partiellement contournée par la suite, grâce au développement d'outils de synthèse vocale plus perfectionnés permettant d'obtenir des sons de parole de meilleure qualité [voir notamment (Carré et al., 2001; Holt, 2005; Maddox et al., 2002)], ou par la transformation de signaux de parole naturelle.[12]

[12] Pour localiser la position précise de la frontière entre les consonnes occlusives voisées et non voisées, le paradigme du continuum synthétique fut à nouveau utilisé, cette fois sur un continuum de VOT (Lisker, 1957). La variation artificielle du VOT d'un enregistrement de parole de naturelle est réalisée simplement en insérant une plage de silence plus ou moins longue entre le relâchement de l'occlusion et le départ du voisement. Par cette manipulation, toutes les informations contenues dans le signal naturel sont conservées dans les stimuli du continuum. Cette méthode a néanmoins un champ d'application très limité puisque seuls des continuums de VOTs positifs peuvent être ainsi obtenus (i.e. uniquement des consonnes occlusives pour lesquelles le voisement démarre après le relâchement de l'occlusion). Là encore, les chercheurs passeront donc rapidement à l'utilisation de la synthèse vocale, avec les inconvénients mentionnés précédemment. La création de

4.2. La méthode du signal progressivement dégradé

Le raisonnement sous-jacent à la méthode du continuum de parole est de tester si une petite modification acoustique du signal de parole se traduit ou non par un changement perceptuel, pour en inférer si cet élément d'information est effectivement extrait par l'auditeur. La méthode du signal dégradé constitue un autre type d'analyse obéissant à la même logique. Il s'agit ici d'appliquer un traitement altérant progressivement le son de parole, selon une dimension particulière, et de repérer le niveau de dégradation à partir duquel le message devient inintelligible. Ce point devrait correspondre à la disparition d'une information acoustique cruciale pour la compréhension.

Le type de dégradation le plus couramment utilisé dans ce cadre est une diminution de la précision en fréquence. Cette opération est effectuée par un vocodage avec bruit (*noise-vocoding*), un traitement analogique du signal de parole qui retire une part importante des informations spectrales, tout en conservant la majeure partie des informations temporelles (Shannon et al., 1995). La procédure utilisée pour obtenir un son de parole vocodé est la suivante : l'enregistrement à traiter est filtré dans un petit nombre de bandes de fréquences disjointes. Les enveloppes temporelles des signaux obtenus sont ensuite extraites, puis appliquées à un bruit limité à la bande de fréquences correspondante. Enfin, les bruits modulés obtenus sont additionnés entre eux pour produire le son de parole vocodé. Une telle manipulation revient donc à contrôler la résolution en fréquence du stimulus. Lorsque le nombre de bandes de fréquences est réduit, le signal obtenu ne possède que très peu de détails spectraux.

Bien que ce type de dégradation ait été souvent combiné avec la neuro-imagerie pour identifier les mécanismes cérébraux mis en jeu dans la restauration d'un signal incomplet (Millman et al., 2011; Obleser & Kotz, 2011; Obleser & Weisz, 2012; Varnet et al., 2012a; Wild et al., 2012), elle fut aussi utilisée en tant que telle pour étudier les mécanismes d'apprentissage perceptuel (Davis & Johnsrude, 2007; Davis et al., 2005; Fu & Galvin, 2003; Hervais-Adelman et al., 2008, 2011; McGettigan et al., 2008) ou déterminer le nombre de bandes de fréquence minimum nécessaires pour préserver l'intelligibilité du message (Apoux & Bacon, 2004; Loizou et al., 1999; Lorenzi et al., 1999; Shannon et al., 1995). Il ressort du recoupement de ces différentes expériences que peu d'informations spectrales sont essentielles pour la compréhension. Pour un signal de parole réduit à seulement 5 bandes de fréquence, les auditeurs parviennent à maintenir un taux de 90% de reconnaissances correctes. Ils font par ailleurs preuve d'une capacité d'adaptation très rapide, passant de 0 % à 70 % de réponses correctes en

continuums par découpe / insertion d'un signal naturel ne fut plus utilisée que dans de très rares situations, comme la réduction de voyelle centrale (Seck, 2012; Serniclaes & Seck, 2013). L'hybridation progressive d'un enregistrement de phonème avec un autre fut également employée mais sans obtenir un rendu satisfaisant (Fowler, 2006).

l'espace de 30 phrases vocodées, sans même avoir besoin d'un retour sur la validité de leurs réponses, et ils sont ensuite capables de généraliser cette connaissances à d'autres formes de dégradation.

Des études similaires furent réalisées avec des stimuli progressivement dégradés temporellement (Drullman et al., 1994a, 1994b; Van Tasell et al., 1992) ou spectro-temporellement (Davis & Johnsrude, 2003; Obleser et al., 2008; Xu et al., 2005), conduisant à la conclusion que les informations fréquentielles et temporelles sont toutes deux importantes pour la compréhension mais que le système est capable d'une grande flexibilité et s'adapte aux différentes dégradations, sans doute en donnant davantage de poids aux indices considérés comme étant plus fiables (Holt & Lotto, 2010). À nouveau, cette adaptabilité de la perception aux dégradations du signal constitue un obstacle car cela signifie que les indices acoustiques identifiés pour des niveaux de dégradation élevés ne sont pas nécessairement ceux utilisés lors de la compréhension de parole naturelle.

La dégradation globale des stimuli par une diminution de la résolution temporelle ou spectrale vise uniquement à identifier la dimension (temps ou fréquence) porteuse de l'information parolière. Pour connaître la position précise des indices acoustiques dans le signal il est nécessaire d'altérer ce dernier de manière plus localisée, en en supprimant progressivement des portions restreintes jusqu'à ce qu'il devienne impossible à identifier. Furui mesura ainsi l'intelligibilité d'enregistrements de syllabes CV au début desquels un segment plus ou moins grand avait été tronqué (Furui, 1986). Ceci lui permit de déterminer, pour chaque enregistrement, la durée critique du segment supprimé entraînant une chute brutale des pourcentages d'identifications correctes (alors que la suppression d'un segment légèrement plus court n'altère pas l'intelligibilité de la consonne). Cette observation l'amena à conclure que les indices acoustiques utilisés pour identifier la consonne se situent nécessairement à cette position critique. Dans un deuxième temps, Furui répéta la même manipulation en tronquant progressivement la fin de l'enregistrement, localisant ainsi temporellement les indices acoustiques correspondant à chaque enregistrement de voyelle. Cette méthode par troncature progressive du signal fut également employée dans le cadre de la *3-Dimensional Deep Search* comme nous le verrons ci-dessous (voir également la Figure 18 pour un exemple d'application de cette approche à une production de la syllabe /ka/).

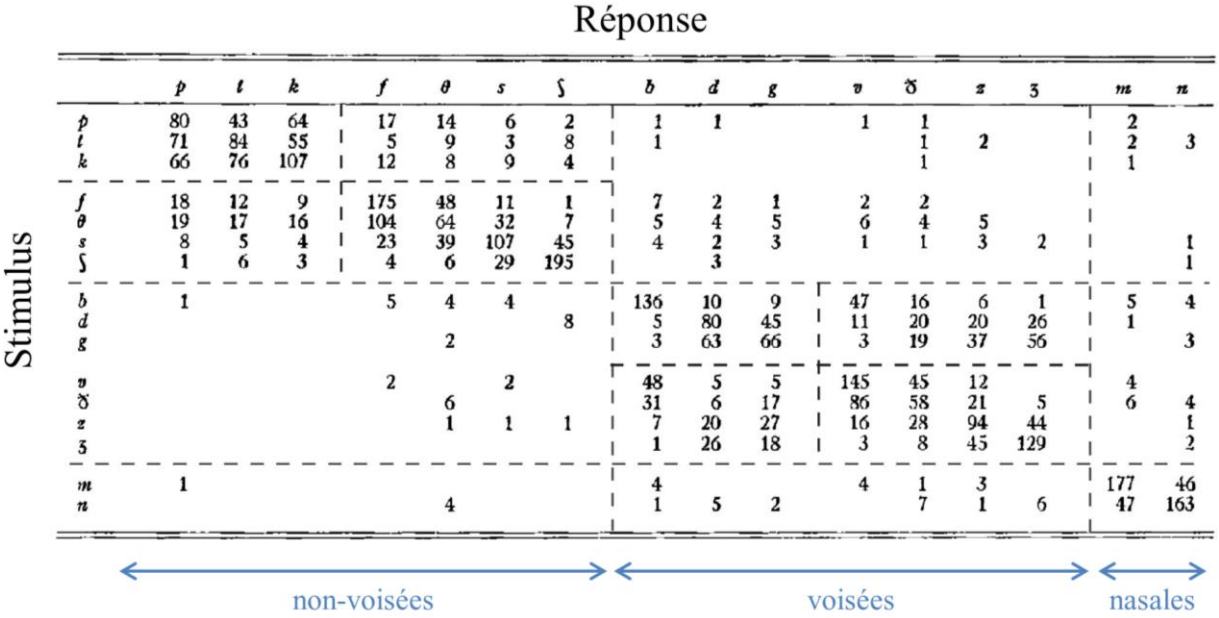
Le même type d'analyse peut également être appliqué selon l'axe des fréquences, en filtrant progressivement le signal de parole au moyen d'un filtre passe-haut ou passe-bas jusqu'à une fréquence de coupure critique à partir de laquelle le message devient brutalement inintelligible. Des exemples peuvent être trouvés dans les premiers travaux de Harvey Fletcher concernant l'indice d'articulation (Allen, 1994, 1996; Fletcher, 1922) reproduits plus tard comme un volet de la *3-Dimensional Deep Search* (voir le paragraphe 4.5 et la Figure 18) (Li & Allen, 2009).

4.3. L'analyse des matrices de confusion

Dans les expériences décrites précédemment, la compréhension du son de parole dégradé est mesurée uniquement en termes de pourcentages de reconnaissances correctes. Cependant, l'examen des différents types d'erreurs commises par les auditeurs permet une analyse plus fine des mécanismes en jeu. En particulier, le dénombrement des confusions qui peuvent survenir lors de la compréhension de parole bruitée amena à constater que celles-ci ne sont pas équiprobables pour tous les phonèmes (Miller & Nicely, 1955). Le caractère récurrent de certaines erreurs par rapport à d'autres nous renseigne sur les traitements effectués par le système pour différencier ces sons. Une confusion privilégiée entre deux phonèmes traduit le fait que ceux-ci sont « perceptivement proches » et donc partagent probablement des caractéristiques acoustiques primordiales pour l'identification, prêtant ainsi plus aisément à confusion (Allen, 2005).

L'étude des confusions suppose une tâche de compréhension de la parole durant laquelle des enregistrements de nombreux phonèmes sont présentés à des participants qui tentent de les identifier en choix ouvert (c'est-à-dire sans être astreints à choisir parmi un petit nombre de réponses prédéfinies). Par exemple, dans une expérience fondatrice, Miller et Nicely demandèrent à 5 participants de reconnaître différents enregistrements de syllabes CV composées d'une consonne variable (/p/, /t/, /k/, /f/, /θ/, /s/, /ʃ/, /b/, /d/, /g/, /v/, /ð/, /z/, /ʒ/, /m/ ou /n/) et d'une voyelle fixe (/a/) (Miller & Nicely, 1955). Plusieurs productions de chaque syllabe par différents locuteurs étaient présentées dans un bruit blanc (pour un total de 250 stimuli en moyenne pour chaque phonème). Le même genre de tâche relativement naturelle fut réalisée dans différents types de bruit : bruit blanc (Benkí, 2003; Miller & Nicely, 1955; Phatak et al., 2008), bruit naturel (Meyer et al., 2010, 2013), bruit de spectre équivalent à celui de la parole (Phatak & Allen, 2007; Trevino & Allen, 2013), ou parole concurrente (Varnet et al., 2012b). Une fois collectées les réponses de tous les participants, il est possible de les représenter sous la forme d'une matrice de confusion, chaque ligne (*i*) correspondant à un phonème présenté et chaque colonne (*j*) à un phonème répondu (voir un exemple Figure 16). Chaque coefficient C_{ij} de la matrice correspond alors au nombre de présentations du phonème *i* ayant donné lieu à la réponse *j*. Quand les conditions d'écoute sont idéales, la matrice est diagonale (la présentation du phonème *i* entraînant invariablement la réponse *i*). Au contraire quand le son est entièrement inaudible, et en l'absence de tout biais intrinsèque de l'auditeur, les réponses deviennent purement aléatoires et la matrice de confusions tend donc vers une matrice constante. Entre ces deux situations, on observe une répartition particulière des erreurs de reconnaissance. Ainsi dans chacune des études psycholinguistiques ci-dessus, plusieurs niveaux de bruits furent testés, correspondant à différents SNR. L'étude de Miller et Nicely, menée avec des SNRs de (-18 dB, -12 dB, -6 dB, 0 dB, +6 dB et +12 dB) donne donc lieu au tracé de 6 matrices, chacune correspondant aux confusions ayant lieu pour

un niveau de bruit spécifique. La Figure 16 présente la matrice correspondant au SNR de -6 dB.



Stimulus \ Réponse	p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n
p	80	43	64	17	14	6	2	1	1		1	1			2	
t	71	84	55	5	9	3	8	1				1	2		2	3
k	66	76	107	12	8	9	4					1			1	
f	18	12	9	175	48	11	1	7	2	1	2	2				
θ	19	17	16	104	64	32	7	5	4	5	6	4	5			
s	8	5	4	23	39	107	45	4	2	3	1	1	3	2		1
ʃ	1	6	3	4	6	29	195		3							1
b	1			5	4	4		136	10	9	47	16	6	1	5	4
d							8	5	80	45	11	20	20	26	1	
g					2			3	63	66	3	19	37	56		3
v				2		2		48	5	5	145	45	12		4	
ð					6			31	6	17	86	58	21	5	6	4
z					1	1	1	7	20	27	16	28	94	44		1
ʒ								1	26	18	3	8	45	129		2
m	1							4			4	1	3		177	46
n					4			1	5	2		7	1	6	47	163

Figure 16 - Exemple de matrice de confusion de consonnes dans un bruit blanc à -6 dB SNR. Les phonèmes sont organisés par traits phonologiques. Les pointillés marque la séparation entre consonnes non voisées, voisées et nasales, et entre occlusives et fricatives. Adapté de (Miller & Nicely, 1955).

Deux constatations s'imposent à la lecture de cette matrice. D'une part, certains phonèmes sont plus robustes au bruit que d'autres (notamment /f/, /ʃ/, /m/ et /n/) comme l'indique le nombre de réponses correctes sur la diagonale. D'autre part, certaines confusions sont plus fréquentes, en particulier entre les membres des trois groupes de confusion {/p/, /t/, /k/, /f/, /θ/, /s/, /ʃ/}, {/b/, /d/, /g/, /v/, /ð/, /z/, /ʒ/} et {/m/, /n/}, tandis que les confusions entre des phonèmes appartenant à deux groupes différents sont extrêmement rares. On remarque que ces groupes, définis par les confusions, coïncident avec certains traits phonologiques (marqués par des pointillés sur la Figure 16). Ainsi le premier groupe contient-il toutes les consonnes non voisées, le second les consonnes voisées non nasales, et le troisième les consonnes nasales. Qui plus est, on constate l'existence de sous-groupes correspondant à la distinction entre occlusives et fricatives. En résumé, une confusion a plus de chance de se produire entre deux consonnes partageant un grand nombre de traits distinctifs (comme /p/ et /k/ qui diffèrent uniquement par leur point d'articulation) qu'entre deux consonnes opposées sur de nombreux traits (comme /p/ et /n/ qui diffèrent à la fois du point de vue du voisement, de la place et de la nasalité). Cette observation confère donc une validité perceptive aux traits phonétiques, définis à l'origine uniquement sur la base de critères articulatoires. Ils semblent être transmis indépendamment puisque les confusions

portent rarement sur plusieurs traits simultanément. Cependant la correspondance entre les traits phonologiques et les groupes apparaissant sur la matrice n'est pas exacte : par exemple, les nasales semblent avoir ici le même statut que les non voisées et les voisées non nasales (Allen, 2006). Cette observation encouragea les chercheurs à délaisser la classification phonologique des phonèmes pour en définir une nouvelle, basée uniquement sur les données de perception rassemblées dans la matrice de catégorisation.

Les régularités observées au sein des erreurs commises par les auditeurs offrent une mesure indirecte de la « distance perceptuelle » entre chaque paire de phonèmes (Johnson, 2011; Mermelstein, 1976) : deux phonèmes proches donneront souvent lieu à des confusions, tandis que deux phonèmes éloignés ne seront confondus qu'à des niveaux de bruit très élevés. Cette interprétation est assez séduisante car elle conduit naturellement à positionner les phonèmes dans un espace perceptuel. Cependant, les méthodes de regroupement hiérarchique (*hierarchical clustering*) ou d'Analyse en Composante Principale, employées pour grouper les phonèmes d'après la matrice de confusion, sont complexes et ne garantissent pas une solution unique (Baayen, 2008; Hastie et al., 2001; Johnson, 2011; Phatak & Allen, 2007). Un autre problème plus fondamental se pose : à la différence d'une matrice de distances, la matrice de confusion n'est pas symétrique. Ainsi un /s/ est plus souvent pris pour /z/ qu'un /z/ pour un /s/ et, de manière générale, on observe un biais des réponses en faveur des consonnes voisées et des voyelles antérieures. Dès lors, il semble plus correct d'interpréter cette matrice dans le cadre de la Théorie de l'information de Shannon, comme proposé par (Allen, 1994). En divisant chaque ligne par le nombre total de présentations de chaque stimulus, nous obtenons une matrice de probabilités conditionnelles (probabilité de la réponse j sachant que le stimulus i a été présenté). Elle est appelée matrice de transition du canal discret, dont la capacité peut ensuite être décrite dans le cadre théorique proposé par Shannon (Palm, 2012).

Un problème subsiste néanmoins. La quantité importante de données collectées (N possibilités de réponse à chacun des N phonèmes présentés, pour chaque participant et chaque niveau de bruit) rend la lecture directe des matrices difficile. D'une part l'ordre arbitraire dans lequel sont représentés les phonèmes au sein de la matrice influe grandement sur la clarté des groupes de confusion : ces derniers, évidents sur la Figure 16, seraient bien moins lisibles avec une organisation différente des consonnes. D'autre part, la répartition des confusions dans la matrice présente une évolution avec le niveau de bruit, certains groupes de confusion étant plus marqués pour des niveaux de bruit faibles, d'autres pour des niveaux de bruit importants. Ce manque de transparence de la matrice de confusion en fait un outil limité pour l'exploration détaillée des erreurs commises par le système auditif lors de la perception.

4.4. L'analyse des profils de confusion

Un pas fut franchi par Jont Allen en 2005, grâce à un simple changement de représentation (Allen, 2005). Plutôt que de visualiser l'ensemble des confusions entre phonèmes présentés et répondus pour un SNR donné (c'est-à-dire les matrices de confusion pour chaque niveau de bruit), il choisit de représenter le profil de confusion, c'est-à-dire l'évolution des probabilités de chacune des réponses en fonction du SNR pour un stimulus donné. Allen ré-analysa ainsi les données collectées par (Miller & Nicely, 1955) ce qui lui permet de révéler sans ambiguïté l'évolution des groupes de confusion selon le niveau de bruit (Figure 17).

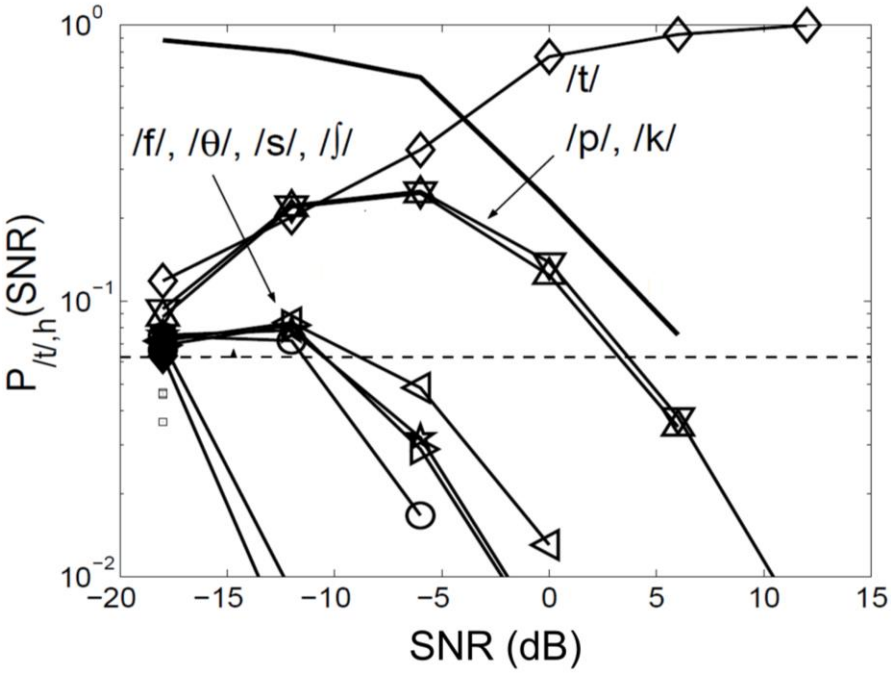


Figure 17 – Profil de confusion pour la syllabe /ta/ obtenus d'après les données de Miller et Nicely. Représentation de la probabilité de chaque réponse en fonction du SNR (traits fins). Le trait gras représente la probabilité totale d'erreur et le trait pointillé le niveau de chance (p=1/16). Adapté de (Allen, 2005).

L'exemple du profil de confusion pour la consonne /t/ présenté Figure 17 indique clairement que la probabilité de réponse correcte (i.e. réponse « t », symbole losange) augmente avec le SNR tandis que la probabilité d'erreur suit le mouvement inverse (trait épais). Cette dernière peut être subdivisée en 15 réponses erronées possibles différentes (seules /p/, /k/, /f/, /θ/, /s/, et /ʃ/ sont indiquées sur le profil de confusion, les autres consonnes n'étant presque jamais confondues avec /t/). Leurs probabilités ne suivent pas toutes une même trajectoire continûment décroissante. Deux groupes de confusion apparaissent clairement. Tandis qu'au-delà de -6 dB SNR le

stimulus /ta/ est identifié sans difficulté, à partir de ce niveau de bruit la probabilité de réponse /t/ devient à peu près égale aux probabilités de confusions avec les consonnes occlusives /p/ et /k/. De la même manière, un second groupe de confusions se détache autour de -18 dB SNR. Il s'agit des fricatives /f/, /θ/, /s/, et /ʃ/. Cette hétérogénéité des erreurs pour les différents groupes de phonèmes est interprétée par Jont Allen comme une preuve du partage de certaines caractéristiques acoustiques entre les membres d'un même groupe de confusion. Lorsque l'indice primaire est masqué par le bruit, ce son est confondu avec des phonèmes proches ne différant que par cette caractéristique. Ici, les productions de /ta/ utilisées comme stimuli dans l'expérience de Miller et Nicely présentent des similarités importantes avec les autres consonnes occlusives non voisées, favorisant les confusions au sein de ce groupe lorsque l'indice principal n'est plus perceptible, ainsi que des similarités moindres avec les fricatives. Pour les SNRs extrêmement faibles, auxquels la parole n'est plus audible, les réponses sont aléatoires et leurs probabilités rejoignent le niveau du hasard (p=1/16).

Ainsi Jont Allen a-t-il pu donner une description plus détaillée des résultats de Miller et Nicely : les confusions s'organisent par groupes de phonèmes proches dont l'importance relative évolue avec le SNR, ce qui établit l'existence d'indices acoustiques secondaires plus ou moins robustes au bruit. Il lui était malheureusement impossible à ce stade de donner une définition acoustique précise de ces indices, item par item, les seules données disponibles de l'article de Miller et Nicely étant uniquement constituées des taux de confusion moyens sur l'ensemble des productions d'une même syllabe. Phatak et ses collaborateurs décidèrent donc d'exploiter la même méthode d'analyse sur de nouvelles données (Phatak & Allen, 2007; Phatak et al., 2008). Ces expériences impliquaient toutes les deux 64 CV (16 consonnes associées à 4 voyelles), chacune produite par 14 locuteurs différents. À partir d'un même enregistrement, 6 stimuli bruités ont été générés pour 6 valeurs de SNR : −22 dB, −20 dB, −16 dB, −10 dB, −2 dB et +∞ dB (signal clair). Les chercheurs tracèrent le profil de confusion dans la tâche de masquage pour chaque production spécifique. Ce test leur permit, dans un second temps, de relier les groupes de confusion obtenus avec une analyse acoustique détaillée du stimulus. Pour la plupart des phonèmes étudiés, la courbe d'identifications correctes marque une rupture brutale à partir d'une certaine valeur de SNR. Comme nous l'avons vu, le niveau de dégradation à partir duquel le message devient inintelligible correspond à la disparition d'une information acoustique essentielle pour la compréhension. Par une représentation du stimulus aux différents niveaux de bruits, il devient alors possible de déterminer quelles régions temps-fréquence sont inaccessibles à l'auditeur à partir de ce seuil critique (Régnier & Allen, 2008). Ces éléments forment un petit groupe de candidats susceptibles d'être les indices acoustiques utilisés lors de la tâche. La Figure 18 (panneaux 1, 3 et 5) offre un exemple de ce raisonnement appliqué à une production de la syllabe /ka/. Régnier et Allen validèrent a posteriori leur approche en constatant que le niveau de bruit à partir duquel l'indice acoustique ainsi identifié n'est plus accessible constitue un excellent prédicteur du seuil critique de SNR pour la catégorisation correcte du son.

Le type d'analyse décrit ci-dessus exige une tâche relativement naturelle (écoute de productions de parole dans différents niveaux de bruit) et fait apparaître clairement les groupes de consonnes partageant des indices acoustiques semblables. De plus, le recoupement des scores de confusion avec une étude de la composition de chaque stimulus révèle un petit nombre d'objets sonores possiblement investis d'une fonction discriminative. Cette méthode ne nécessite pas de connaissances a priori concernant les indices acoustiques mis en jeu, contrairement à celle du continuum synthétique. Néanmoins, la quantité importante de données nécessaires au tracé des profils de confusion (64 CV x 14 locuteurs x 6 SNR) suppose un temps d'expérimentation extrêmement long (15 h par participants). Enfin, cette approche fondée uniquement sur un masquage progressif du stimulus ne permet pas de particulariser deux éléments du signal possédant la même énergie (et donc disparaissant au même seuil de bruit).

4.5. 3-Dimensional Deep Search

La seule connaissance de l'amplitude de l'indice acoustique recherché ne permettant pas d'identifier celui-ci de manière univoque, Jont Allen et ses collaborateurs entreprirent de le caractériser selon deux dimensions supplémentaires : en temps et en fréquence (Li et al., 2010). Le recoupement de ces trois informations pourrait ainsi permettre une localisation plus précise. À cette fin, ils complétèrent le protocole de masquage progressif décrit ci-dessus par une analyse des profils de confusion issus de deux autres expériences psychoacoustiques déjà mentionnées précédemment (cf. paragraphe 4.2). Ils réalisèrent, d'une part, une troncature progressive du signal pour isoler les indices acoustiques dans la dimension temporelle et, d'autre part, un filtrage en fréquence pour faire de même dans la dimension spectrale. Comme précédemment, les résultats de ces trois opérations (troncature, filtrage et masquage) étaient analysés indépendamment pour chaque stimulus, de façon à pouvoir ensuite relier les informations obtenues dans les trois dimensions aux représentations temps-fréquence des productions individuelles, d'où le nom de *3-Dimensional Deep Search* (3DDS).

Un exemple d'application de cette méthode pour une production de la syllabe /ka/ est donné Figure 18. La probabilité de la réponse /k/ est de 100 % lorsque le stimulus est intact. En revanche, lorsque la petite région spectro-temporelle encadrée en rouge est filtrée passe-haut, passe-bas, ou tronquée, la perception bascule brusquement vers une consonne proche (/p/ ou /t/). Les deux opérations pointent donc ici sans ambiguïté vers l'explosion brève autour de 1500 Hz comme objet sonore nécessaire pour l'identification correcte de la consonne. L'expérience de masquage progressif confirme ce résultat : le taux de réponses correctes à la présentation de cette production de /ka/ additionnée de bruit chute à partir d'un SNR de 0 dB. Or le modèle de l'AI-gram indique que l'explosion n'est plus perceptible au-delà de ce niveau de bruit, ce qui explique l'incapacité des participants à identifier correctement la consonne. Cet

ensemble d'observations convergentes indique que la région encadrée contient le substrat acoustique du percept /k/.

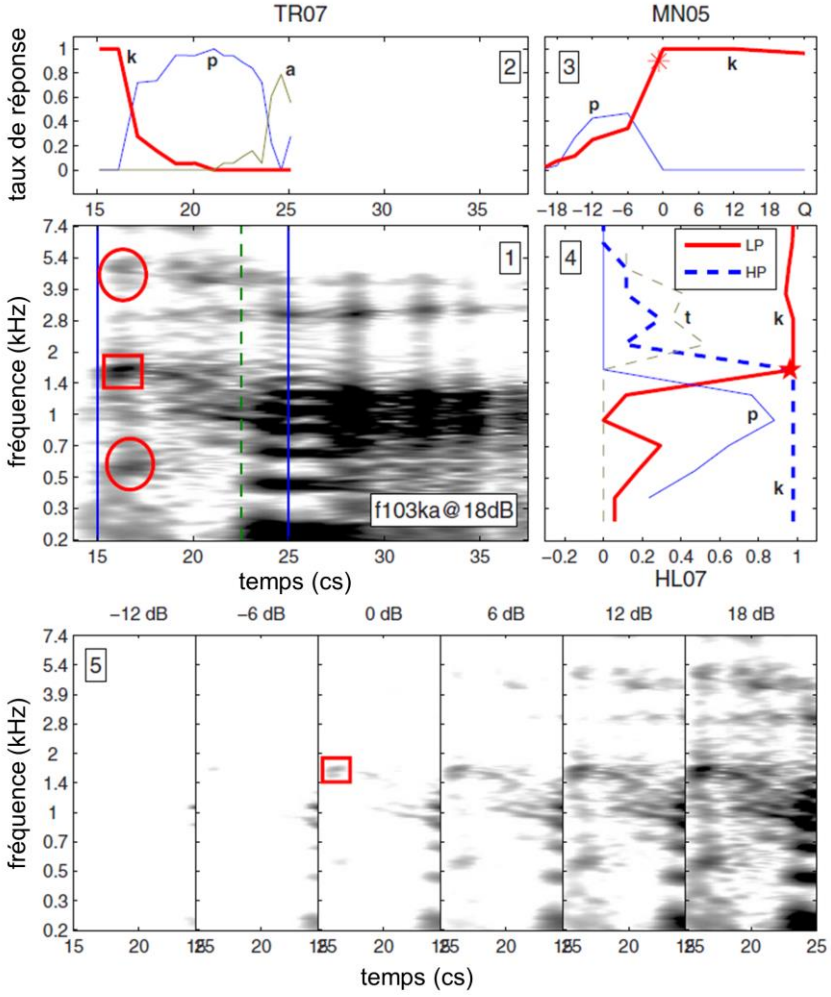


Figure 18 – Résultat de l'analyse 3DDS pour une production de la syllabe /ka/. Adapté de (Li et al., 2010). (1) représentation temps-fréquence du stimulus (AI-gram à 18 dB SNR). L'indice acoustique identifié est encadré en rouge et les indices conflictuels sont marqués par des ellipses. Le trait vert indique approximativement le début de la voyelle et les traits bleus la plage considérée pour les troncatures. (2) Profil de confusion obtenu lors de la troncature progressive du segment initial (taux de réponses correctes en rouge). (3) Profil de confusion obtenu lors du masquage progressif du stimulus (taux de réponses correctes en rouge). (4) Profils de confusion obtenus lors du filtrage progressif passe-bas (traits continus, taux de réponses correctes en rouge) ou passe-haut (traits discontinus, taux de réponses correctes en bleu) du stimulus. (5) Représentation du stimulus à différents SNRs (AI-grams).

Un avantage évident de cette méthode par rapport aux précédentes est qu'elle répond directement à notre question en délimitant une petite bande de fréquences et une position temporelle précise correspondant à une information acoustique cruciale

pour une compréhension correcte de la consonne. La troisième étape fait office de confirmation en démontrant que le masquage de cet indice par un bruit entraîne une chute subite du taux de reconnaissances correctes. Par ailleurs, cette approche est suffisamment flexible pour mettre à jour des indices secondaires conflictuels. Ce terme désigne des caractéristiques acoustiques du signal, partagées avec un autre phonème et qui conduisant à des confusions récurrentes lorsque l'indice principal est supprimé (Li & Allen, 2011; Li et al., 2010). Ainsi la production de /ka/ présentée Figure 18 est quasi systématiquement perçue comme un /pa/ lorsque l'indice principal est effacé par filtrage passe-bas. L'examen des profils de confusion amena Li et ses collaborateurs à conclure que le maximum secondaire de l'explosion, situé autour de 600 Hz, est dans ce cas interprété comme un indice en faveur de /pa/. Au contraire, lorsqu'un filtrage passe-haut efface l'indice principal, c'est le troisième maximum de l'explosion, vers 4000 Hz, qui biaise systématiquement les réponses des participants en faveur de /ta/.

De manière générale, l'utilisation de la 3DDS requiert le plus souvent un temps de passation extrêmement long (4h pour la tâche de troncature, 4h pour le filtrage et 15 h pour le masquage) et une part non négligeable d'interprétation pour conclure quant à la stratégie adoptée par le système pour catégoriser le stimulus. Il existe en effet des situations où les résultats des trois opérations ne concordent pas aussi parfaitement, notamment pour la consonne /b/ (Li et al., 2010). La 3DDS est mise en difficulté lorsque plusieurs indices acoustiques (dans ce cas un bruit large bande et la position du F2) sont détectés et combinés par le système auditif pour donner naissance au percept phonémique. En effet, dans ce cas, les scores de reconnaissances obtenus lors de la troncature, le filtrage passe-bas, passe-haut ou le masquage du signal marquent une chute rapide – dès qu'un seul de ces indices disparaît – et ne pointent donc pas vers une position spectro-temporelle unique. Une possibilité envisageable pour pallier ce problème consisterait à dégrader le signal de façon plus localisée. Idéalement la suppression sélective d'une portion du son de parole susceptible de porter l'information phonémique permettrait de confirmer son rôle dans la formation du percept. Ce fut la procédure adoptée par (Li & Allen, 2011). En atténuant ou en renforçant de manière ciblée de petites portions du spectrogramme, circonscrites au préalable par la 3DDS, les chercheurs purent valider a posteriori leur rôle essentiel dans le processus de reconnaissance de la parole. Là encore, la suppression d'un indice principal fait basculer la perception vers les phonèmes proches définis par les indices secondaires. Au contraire le renforcement de l'indice principal, et l'élimination des indices conflictuels, améliore l'intelligibilité du stimulus et sa robustesse face au bruit. Naturellement, cette approche ne peut être considérée comme une méthode d'identification des indices acoustiques en tant que telle car elle ne peut être utilisée qu'en complément d'une autre pour confirmer a posteriori le rôle d'indices circonscrits au préalable. Cependant elle pointe une limite de la 3DDS et de la plupart des procédures décrites précédemment: les traitements utilisés pour dégrader le signal sont relativement simples (filtre passe-bas, passe-haut, troncature, masquage uniforme par un bruit…) et ne permettent pas de

révéler des stratégies de catégorisation parfois complexes et impliquant différentes régions du signal.

4.6. La méthode des fonctions de pondération

Le problème des informations acoustiques utiles pour la compréhension de la parole fut abordé, durant la même période, par d'autres groupes de recherche, en s'appuyant sur le travail d'Ahumada et Lovell (Ahumada & Lovell, 1971). Leur objectif était d'étudier, de manière individuelle, l'importance de différentes « portions » du signal pour l'intelligibilité de ce dernier. La nécessité de mener une investigation indépendante pour chaque élément mis en jeu impliquait de réduire la résolution à un petit nombre de bandes de fréquence et/ou de segments temporels. Chacune des régions ainsi repérées se voyait alors attribuer un poids en fonction de son influence sur la reconnaissance ou non du stimulus par les participants, définissant ainsi une fonction de pondération qui rend compte de la stratégie d'écoute de chaque auditeur (Alexander & Lutfi, 2004). Cette fonction peut également être interprétée comme un filtre attentionnel décrivant la répartition des ressources cognitives sur différentes caractéristiques du son de parole.

Lors de ces expérimentations, un panel de stimuli variés fut utilisé : tons purs (Ahumada & Lovell, 1971; Alexander & Lutfi, 2004; Davidson et al., 2006), syllabes (Apoux & Bacon, 2004; Turner et al., 1998a), pseudomots (Doherty & Turner, 1996; Gilbert & Micheyl, 2005; Gilbert et al., 2002) ou phrases (Calandruccio & Doherty, 2007, 2008). La procédure consistait le plus souvent à diviser l'axe des fréquences en un nombre restreint de bandes (typiquement entre 3 et 6) puis à mesurer l'information portée par chacune d'elles, en évaluant l'intelligibilité du signal lorsque l'une des bandes est retirée par filtrage ou lorsque les bandes sont masquées par des niveaux de bruit variés. Dans les deux cas, les poids associés aux différentes bandes sont obtenus par une régression multiple entre le niveau de dégradation de chaque bande à chaque essai et la réponse (correcte vs. incorrecte) du participant à cet essai. Les premières études employèrent une corrélation ou une régression linéaire pour calculer ces poids (Doherty & Turner, 1996) mais il apparut rapidement que la régression logistique était plus indiquée dans ce contexte (Alexander & Lutfi, 2004). Des méthodes statistiques plus avancées, comme la régression logistique multiple avec régularisation L1 (*L1-regularized multiple logistic regression*), furent proposées et validées par la simulation (Schönfelder & Wichmann, 2012). Ces choix théoriques seront discutés plus en détail dans le chapitre 7 à propos des images de classification. Ainsi cette approche permet-elle d'étudier l'importance d'une bande de fréquence particulière dans la compréhension de la parole, contrairement à la méthode du signal progressivement filtré passe-haut ou passe-bas qui n'examinait que le rôle des hautes ou des basses fréquences (cf. paragraphe 4.2). Dans ce sens, elle offre donc davantage de finesse dans

la description de la stratégie d'écoute employée par un auditeur, même si la résolution spectrale demeure relativement faible.

Malgré de petites différences quantitatives, les différents participants démontrent une répartition des poids globalement similaire (Alexander & Lutfi, 2004; Apoux & Bacon, 2004; Doherty & Turner, 1996; Turner et al., 1998a), un résultat encourageant quand on considère que les tâches demandées (détection de ton, compréhension de la parole…) mettent en jeu des processus extrêmement automatisés et, donc, probablement identiques chez tous les auditeurs. À l'opposé, pour des tâches plus complexes, les participants semblent diversifier leurs stratégies d'écoute (Doherty & Lutfi, 1996; Gilbert & Micheyl, 2005). Les expériences menées sur des tâches de parole (identification de consonnes ou répétition de phrases), particulièrement intéressantes dans le cadre de notre étude, montrent qu'une bande de fréquences, approximativement située autour de 3500 Hz, joue un rôle primordial : la suppression des informations acoustiques qu'elle véhicule fait notablement chuter le taux de réponses correctes (Apoux & Bacon, 2004; Calandruccio & Doherty, 2007, 2008; Gilbert & Micheyl, 2005; Gilbert et al., 2002; Turner et al., 1998a). On observe également une certaine plasticité de cette stratégie privilégiant les hautes fréquences, puisqu'elle est modifiée sous certaines conditions d'écoute, comme la présence de parole concurrente (Gilbert & Micheyl, 2005) ou le type de dégradation utilisé (Apoux & Bacon, 2004; Turner et al., 1998a).

L'une des principales raisons du succès des fonctions de pondération dans l'exploration des stratégies d'écoute est liée à la possibilité d'étudier directement une autre forme de plasticité auditive. Ainsi, l'estimation des poids attribués aux différentes bandes pour la réalisation d'une même tâche auditive par un groupe de participants normo-entendants, d'une part, et par un groupe de participants souffrant de pertes auditives, d'autre part, offrit-elle une visualisation directe de la capacité d'adaptation de ces derniers aux contraintes liées à leur surdité (Alexander & Lutfi, 2004; Calandruccio & Doherty, 2008; Gilbert et al., 2002; Mehr et al., 2001).

Enfin, certains groupes de recherche proposèrent d'adapter cette méthode dans le cas d'un découpage temporel (Davidson et al., 2006; Pedersen & Ellermeier, 2008) ou spectro-temporel (Ahumada et al., 1975; Joosten & Neri, 2012) du signal (voir le paragraphe 7.2 pour la description de l'étude de Ahumada). Cette dernière possibilité donnera naissance à la méthode des Images de Classification dont il sera question au chapitre 7.

La multiplication des études basées sur les fonctions de pondération conduit à constater que celles-ci étaient généralement très similaires d'un auditeur à l'autre, quelle que soit la méthode utilisée pour les calculer (Apoux & Bacon, 2004) et le niveau de performance du participant (Calandruccio & Doherty, 2007). En revanche, cette méthode dépend fortement du choix des bandes de fréquences (Calandruccio & Doherty,

2007; Gilbert & Micheyl, 2005) ce qui pose problème quant à l'interprétation possible des poids au niveau cognitif.

4.7. La méthode des bulles auditives

La méthode des bulles (*Bubble Images*) provient du domaine visuel où elle a permis des avancées, notamment dans la compréhension des mécanismes de la reconnaissance des visages (Gosselin & Schyns, 2001; Murray, 2011). Son principe est voisin de celui des Images de Classification que nous aborderons plus loin (Murray, 2012). Dans son versant auditif, cette méthode a été élaborée par Michael Mandel en 2013, parallèlement à celle développée dans cette thèse, ce qui démontre l'engouement récent de la communauté scientifique pour cette problématique (Mandel, 2013). Toutefois, à l'heure actuelle, la méthode des bulles auditives n'ayant encore jamais été mise en œuvre dans le cadre d'un questionnement concret, ses applications possibles demeurent pour l'instant à l'état d'hypothèses.

Comme pour les fonctions de pondération, un signal (enregistrement de mot ou de non-mot) à catégoriser est présenté dans un bruit dont le niveau n'est pas constant. La régression de l'intelligibilité (réponse correcte ou réponse fausse) avec la disposition particulière du bruit permet de déterminer les régions cruciales pour la reconnaissance. Le bruit utilisé est ici un « masque à bulles » (*bubble noise*), c'est-à-dire un bruit blanc de niveau élevé présentant un ou plusieurs vides, de forme circulaire dans l'espace temps-fréquence, où le bruit est absent. L'auditeur peut ainsi accéder au signal uniquement à travers ces « trous » localisés. Une réponse correcte du participant indique donc que l'information transmise au niveau des trous correspondant à cet essai est suffisante pour catégoriser la cible. En rassemblant les données de centaines de catégorisations, il devient ainsi possible d'identifier les zones où réside l'information permettant de réaliser la tâche, c'est-à-dire les indices acoustiques.

L'application de cette approche à un ensemble de stimuli de la forme /aCa/ (/aʃa/, /ada/, /afa/, /aʒa/, /ata/, /ava/) indiqua que la partie cruciale du signal acoustique correspondait à l'attaque de la deuxième syllabe, entre 4 et 10 kHz (Mandel et al., 2014). Plus intéressant, il devenait alors possible d'utiliser cette information pour prédire l'intelligibilité d'une cible additionnée d'un masque à bulles particulier, conduisant à un taux de prédiction significativement supérieur au hasard. Les cartes d'intelligibilité obtenues démontrent en outre une grande généralité puisqu'elles permettent de prédire les réponses d'autres participants ou celles faites aux productions d'autres locuteurs. Cependant, la région identifiée par cette méthode est assez vaste (et la caractérisation des indices acoustiques demeure donc relativement imprécise), ce qui se traduit par des taux de prédiction correctes relativement faibles, bien que significatifs (environ 60%). De plus, bien que cette méthode soit plus objective que celle des

fonctions de pondération puisqu'il n'est pas nécessaire de découper arbitrairement l'axe des fréquences, le résultat demeure dépendant du choix de la base de bulles utilisées.

En résumé, à l'heure actuelle aucune des méthodologies développées pour l'identification des primitives acoustiques de la parole n'est entièrement satisfaisante. L'utilisation d'un continuum synthétique permet de contrôler précisément les caractéristiques acoustiques des stimuli. Cependant sa mise en place requiert une connaissance a priori des composantes du son de parole importantes pour la compréhension. Il en résulte que les stimuli présentés semblent peu naturels et que la méthode ne peut être utilisée qu'en tant que test d'hypothèse. Les approches par dégradation (signal progressivement dégradé et 3DDS) sont quant à elles confrontées à la grande flexibilité du système acoustique. Les profils de confusion et la 3DDS nécessitent un nombre élevé d'essais et une interprétation du résultat qui peut s'avérer très délicate, voire subjective, lorsque plusieurs indices sont mis en jeu. Enfin, la plupart des méthodes mentionnées ne permettent qu'un encadrement relativement peu précis de la région temps-fréquence critique. Un tableau récapitulatif des avantages et inconvénients des différentes méthodes sera proposé dans la partie Discussion de cette thèse (Table 1).

5. Plasticités du système auditif

La présentation des différents modèles de l'interface acoustico-phonétique au chapitre 3.2 pourrait laisser croire que l'association entre indices acoustiques et phonèmes est figée et qu'une même production de parole sera toujours traitée de la manière identique par le système auditif. Au contraire, celle-ci démontre une grande plasticité, observable sur différentes échelles de temps. De plus, elle peut dans certains cas se révéler déficiente.

Nous avons déjà noté plus haut une capacité d'adaptation à court terme, notamment en réponse à la présence de bruit concurrent. La redondance des informations contenues dans le signal explique que la parole puisse être perturbée dans de larges proportions sans que l'intelligibilité n'en pâtisse. Ceci suggère que l'auditeur peut porter sélectivement attention à différents éléments selon la tâche à réaliser et les conditions d'écoute. Le paragraphe 5.1 ci-dessous rassemble un certain nombre d'observations concernant l'interaction entre le processus d'extraction des indices acoustiques et la présence de bruit. Sur le plus long terme, un entraînement spécifique des compétences auditives peut améliorer le traitement des phonèmes. C'est notamment le cas pour les musiciens experts qui manifestent un meilleur encodage des sons de parole au niveau du tronc cérébral et des taux d'erreur réduits. Le paragraphe 5.2 détaillera l'état actuel des recherches concernant l'impact de l'apprentissage musical sur la compréhension de la parole. Enfin, la partie 5.3 s'intéressera à une partie de la population présentant des difficultés lors de la perception et de l'utilisation du langage écrit et oral : les dyslexiques développementaux. Une origine de ces troubles pourrait être un déficit au niveau de l'appariement entre éléments acoustiques et unités linguistiques.

5.1. Compréhension dans le bruit et indices acoustiques

La multiplicité et la redondance des informations acoustiques supportant la compréhension du langage oral ont déjà été soulignées au paragraphe 2.2.5. De manière générale, aucun de ces indices acoustiques n'est absolument nécessaire pour la perception correcte d'un phonème ou d'un mot et plusieurs indices acoustiques sont suffisants. Par ailleurs, l'analyse des matrices de confusion suggère que les différents indices acoustiques ne sont pas tous également robustes au bruit et aux dégradations (Benkí, 2003; Régnier & Allen, 2008) : par exemple, l'indice haute fréquence du /t/ résiste bien à l'ajout de bruit avec un spectre de parole mais pas à un bruit blanc, plus énergétique dans les hautes fréquences. Ces observations soulèvent la question de savoir dans quelle mesure l'extraction de telle ou telle information est conditionnée par la tâche demandée et la situation d'écoute.

La capacité du système auditif à adapter le traitement des sons de parole en fonction des perturbations a été mise en évidence à plusieurs reprises. Elle prend le plus souvent la forme d'une pondération des indices acoustiques : les indices principaux pour une tâche ou une situation d'écoute données pouvant devenir secondaires pour d'autres. Lors de la catégorisation de deux types de sons simples (composés de sinusoïdes modulées en amplitude), les auditeurs sont capables d'adapter l'utilisation des différents indices disponibles en fonction de leur fiabilité et ce, même en cours d'expérience, si cela améliore leurs performances (Lotto and Holt, 2011; Scharinger et al., 2014). La même plasticité du système auditif a été observée dans la compréhension de la parole. Pour reprendre l'exemple du trait de voisement – qui se traduit non seulement par un VOT plus long mais également une f_0 plus grave et des transitions formantiques plus longues (paragraphe 2.2.5) – tandis que le VOT joue un rôle prépondérant en situation d'écoute claire, l'ajout de bruit rend cet indice ambigu et l'auditeur se reporte alors sur les indices secondaires que sont la f_0 et les transitions de formants (Serniclaes and Arrouas, 1995).

Cette plasticité du traitement dépend non seulement de la simple présence de bruit mais également de sa composition spectrale. Comme le suggère le modèle des canaux indépendants évoqué au paragraphe 3.2.3, lorsque le spectre de parole est partiellement masqué par le bruit, le système auditif est en mesure de se reporter sur des bandes de fréquence pour lesquelles le SNR est localement plus favorable. Nous avions illustré ce point avec un cas extrême : certains sons de parole restent intelligibles après un filtrage aussi bien passe-haut que passe-bas relativement à la même fréquence de coupure. C'est donc que des indices hautes- ou basses-fréquences sont utilisés selon le type de dégradation. Ce mécanisme intervient également en situation de parole dans le bruit, comme le démontre la comparaison des fonctions de pondération obtenues pour des tâches de compréhension en présence ou non d'un bruit conversationnel (Gilbert & Micheyl, 2005). La bande de fréquences médiane (1750-3750 Hz), très utilisée, est délaissée lorsqu'un bruit parolier est ajouté. Enfin le rôle des indices secondaires conflictuels décrits au paragraphe 4.5 (Li & Allen, 2011; Li et al., 2010) représente une preuve supplémentaire de la réallocation de l'attention auditive en situation bruitée : lorsque l'indice principal se trouve masqué par le bruit, l'indice secondaire erroné entre en jeu, faisant basculer brutalement la perception vers un autre phonème.

5.2. Effets de l'expertise musicale sur la compréhension de la parole

Si l' « Effet Mozart », en vogue dans les années 1990, et selon lequel la simple exposition à la musique classique permettrait un développement des facultés cognitives, a depuis été largement battu en brèche (Pietschnig et al., 2010; Schellenberg & Peretz, 2008), en revanche l'influence d'une pratique musicale soutenue sur les autres

domaines de la cognition demeure une question ouverte. Il a été démontré, notamment, que la compréhension de la parole peut tirer bénéfice d'un entraînement intensif de capacités non langagières tel que celui fourni lors de l'apprentissage d'un instrument. Six mois de pratique musicale suffisent pour améliorer les performances dans des tâches linguistiques et modifier la configuration des activités cérébrales observées (Moreno et al., 2009), ce qui en fait un parfait exemple de plasticité du système auditif à moyen terme. Néanmoins les mécanismes sous-tendant ce transfert inter-domaine demeurent encore largement méconnus.

D'un point de vue acoustique, les signaux de musique et de parole présentent de nombreuses similarités (Wolfe, 2002). Dans leur structure tout d'abord, puisqu'ils sont tous deux composés de segments relativement stables séparés par des transitions brutales et/ou de brefs silences. Par ailleurs, ils partagent des éléments communs : tous deux sont caractérisés par leur rythme et leur fréquence fondamentale (traduisant la mélodie ou la prosodie), ainsi que par leur structure harmonique (timbre ou complexe formantique). Un deuxième parallèle entre parole et musique peut être envisagé au niveau de leur système de codage. Dans la parole, la structure harmonique, élément fondamental pour le décodage des phonèmes, est perçu de façon catégorielle, au contraire de la prosodie. Dans la musique, en revanche, les notes sont codées en termes de hauteur et de durée et perçues catégoriellement, tandis que le timbre peut varier continûment. Du fait de ces homologies, des fonctions similaires sont sollicitées pour traiter ces deux types de signaux : séparation des flux (*streaming*), mémoire de travail auditive, discrimination des fréquences et des durées, attention auditive, détection dans le bruit, etc. (Patel, 2011).

Toutes ces capacités auditives sont effectivement renforcées lors d'une pratique musicale soutenue. Conformément à nos attentes, ceci se traduit par de meilleures performances des musiciens experts par rapport aux non-musiciens dans des tâches auditives non linguistiques (Gaab et al., 2005; Kishon-Rabin et al., 2001; Micheyl et al., 2006; Rammsayer & Altenmüller, 2006; Strait et al., 2010). Ces changements sont sous-tendus par une véritable réorganisation des aires corticales impliquées dans perception de la musique (expansion des aires auditive, motrice et visuo-spatiale) (Gaser & Schlaug, 2003a, 2003b; Schneider et al., 2002) de même que par une sensibilité accrue de l'activité cérébrale aux caractéristiques spectrales des sons (Brattico et al., 2009; Pantev et al., 1998; Shahin et al., 2005).

Par extension, l'apprentissage d'un instrument s'accompagne de bénéfices notables pour le traitement des signaux de parole. On observe une amélioration globale des performances des musiciens, comparativement aux non-musiciens, notamment pour les tâches de segmentation ou d'extraction de la prosodie (François et al., 2014; Magne et al., 2006). L'intérêt des chercheurs s'est essentiellement porté sur l'encodage plus précis et plus robuste des sons de parole chez les musiciens experts, quantifiable par la mesure des potentiels évoqués auditifs au niveau du tronc cérébral (Bidelman &

Krishnan, 2010; Parbery-Clark et al., 2009a, 2009b, 2012; Strait et al., 2012; Weiss & Bidelman, 2015). Finalement, il a été suggéré par certaines équipes de recherche que ces meilleures représentations des sons permettaient in fine un renforcement de la perception de parole dans le bruit (Parbery-Clark et al., 2009b). Cependant, les mécanismes à la base de ce transfert demeurent encore inexpliqués (Schellenberg & Peretz, 2008). De plus, la reproductibilité de ces résultats dans des conditions proches ou équivalentes a été récemment remise en cause (Boebinger et al., 2015).

5.3. La dyslexie développementale

L'Organisation Mondiale de la Santé estime à 10 % la proportion d'enfants atteints de dyslexie développementale. En termes diagnostiques, ce trouble spécifique de l'acquisition de la lecture est caractérisé par des performances individuelles nettement au-dessous du niveau attendu, pour la classe d'âge et le QI du patient, qui ne peuvent être expliquées uniquement par des déficits sensoriels, ni par une scolarisation inadéquate (World Health Organization, 2010). En pratique, la dyslexie développementale se manifeste, dans les premières années de scolarité, par une difficulté d'apprentissage de la lecture qui se traduit ensuite par des difficultés en orthographe. À l'âge adulte, on observe une persistance de ces difficultés, même s'il est largement admis que les dyslexiques parviennent le plus souvent à mettre en place des stratégies compensatoires.

Cette définition concise contraste avec l'hétérogénéité des déficits cognitifs associés. Ainsi les difficultés de manipulation de l'information phonologique chez les dyslexiques développementaux se traduisent notamment par de faibles performances dans des tâches mettant en jeu la conscience phonologique (tâches d'épellation, de contrepèterie ou de suppression de phonème), l'accès lexical (tâche de dénomination d'objet), la mémoire de travail verbale (tâche de répétition d'une liste de mots) ou, encore, la compréhension de la parole dans le bruit (Boets et al., 2006; Brady et al., 1983; Dole et al., 2012; Law et al., 2014; Ramus et al., 2003; Ziegler et al., 2009, 2011). Certaines études ont également rapproché la dyslexie développementale d'une anomalie de la perception catégorielle des phonèmes, révélée au moyen du paradigme du continuum synthétique : comparativement aux normo-typiques, les participants atteints de dyslexie obtiennent de meilleures performances dans la discrimination de certains allophones (Bogliotti et al., 2008; Noordenbos et al., 2012a; Noordenbos & Serniclaes, 2015), un résultat confirmé par la neuro-imagerie (Dufor et al., 2009; Noordenbos et al., 2012b, 2013), et que certains modèles considèrent comme fondamental (Serniclaes et al., 2004).

Ces nombreuses comorbidités ont été souvent regroupées dans le cadre de l'hypothèse du déficit de traitement phonologique (Ramus, 2003; Sprenger-Charolles et al., 2000; Ziegler et al., 2008). Selon cette théorie, les problèmes des dyslexiques se

fonderaient sur une anomalie lors du stockage ou de la récupération des sons du langage, qui serait source de difficultés durant l'apprentissage de la lecture. Une autre explication, également soutenue dans la communauté scientifique, attribue le panel de déficits observés à des troubles sensori-moteurs plus généraux dus à des dysfonctionnements de circuits neuronaux magnocellulaires (Giraud & Ramus, 2013; Lehongre et al., 2013, 2011).

Quelle que soit son origine, la dyslexie développementale offre un exemple éclairant de la plasticité de l'interface acoustico-phonétique. Loin d'être identiques chez tous les individus, les mécanismes à l'œuvre se révèlent parfois déficients, impactant leurs performances dans les tâches phonologiques. Néanmoins, des stratégies compensatoires sont parfois mises en place, laissant croire que la dyslexie pourrait disparaître à l'âge adulte. Ces observations, de même que les adaptations à court ou long terme décrites dans les paragraphes 5.1 et 5.2, indiquent clairement que les processus mis en jeu au niveau de l'interface acoustico-phonétique ne sont pas figés mais présentent au contraire une assez large variabilité selon la situation d'écoute, l'apprentissage, ou le groupe d'auditeurs étudié.

6. Encodage et décodage de la parole dans le cerveau

Comme nous l'avons déjà souligné, un problème fondamental dans l'étude de la parole est de caractériser les indices acoustiques utilisés par le cerveau et la manière dont ils sont combinés pour construire une représentation compacte et robuste du signal. Il s'agit donc de mettre à jour le codage qui gouverne la manipulation et la reconnaissance des sons du langage dans le cerveau. Au chapitre 4, nous avons proposé différentes méthodes purement comportementales permettant l'identification des primitives de parole. Un second angle d'approche, complémentaire du premier, est offert par la neuro-imagerie. Les récents progrès techniques dans ce domaine nous permettent à présent d'observer la spécialisation de régions plus ou moins vastes du cerveau pour un type de stimuli ou un type de traitements particuliers. Dans une perspective d'exploration des substrats cérébraux du langage, notre question initiale prend une formulation plus neurobiologique et devient : comment les neurones auditifs recueillent-ils et structurent-ils l'information acoustique reçue pour la traduire en information linguistique ?

Dans ce chapitre, nous envisagerons quelques pistes de réponses en nous concentrant sur une technique d'investigation particulière, le Champ Récepteur Spectro-Temporel (*Spectro-Temporal Receptive Field*, STRF) décrit dans la partie 6.1. Dans un second temps, cette technique nous aidera à décoder les représentations utilisées par le système auditif et à suivre leur évolution le long de la voie auditive primaire (partie 6.2). Enfin, la pertinence de ce type d'encodage pour la reconnaissance de la parole sera examinée dans la partie 6.3.

6.1. Le Champ Récepteur Spectro-Temporel (STRF)

Le STRF est probablement l'un des outils les plus couramment utilisés pour étudier la façon dont les neurones auditifs décomposent et analysent les sons. Schématiquement, il consiste à enregistrer, au moyen d'une électrode implantée, les réponses électriques d'un neurone (ou d'un groupe de neurones) lors de la présentation de stimuli acoustiques puis à déterminer le lien entre cette activité et certaines caractéristiques du stimulus. Cette opération a été majoritairement pratiquée sur des pinsons écoutant les vocalisations de leurs congénères (Woolley et al., 2009) mais elle a pu, en de rares cas, être aussi réalisée sur des enregistrements intracrâniens effectués chez des patients épileptiques (Pasley et al., 2012). Ces études ont ainsi permis de mettre en évidence la spécialisation de certains neurones auditifs pour des caractéristiques particulières du signal.

À l'origine, les STRFs étaient estimés par une simple sommation des stimuli entraînant la décharge électrique de la cellule (selon le principe de la corrélation inversée décrite dans la partie 7.3.1). Si cette méthode est relativement efficace dans le cas de stimuli composés uniquement de bruit blanc, elle se trouve en revanche largement mise en défaut lorsque le son possède une structure interne hautement corrélée, comme c'est le cas pour les signaux naturels (Laudanski et al., 2012; Theunissen & Elie, 2014). D'autres techniques statistiques, plus efficaces, furent donc mises au point, telle la régularisation (Calabrese et al., 2011; David et al., 2007; Wu et al., 2006), décrite dans la partie 7.7 dans le cas des Images de Classification. Dans ce contexte, le STRF correspond à la réponse impulsionnelle temps-fréquence du neurone, $h(\tau, f)$, convoluée avec le spectrogramme du signal d'entrée $S(t, f)$. La probabilité de décharge $\hat{r}(t)$ est ensuite obtenue par une transformation non linéaire g du résultat. Ceci nous fournit une modélisation simple pour les neurones auditifs, également utilisée pour les neurones visuels (Pillow, 2007), et schématisée Figure 19. Elle est parfois complétée par un filtre post-décharge (*post-spike filter*) sur une boucle de rétroaction, modélisant la dynamique temporelle des réponses.

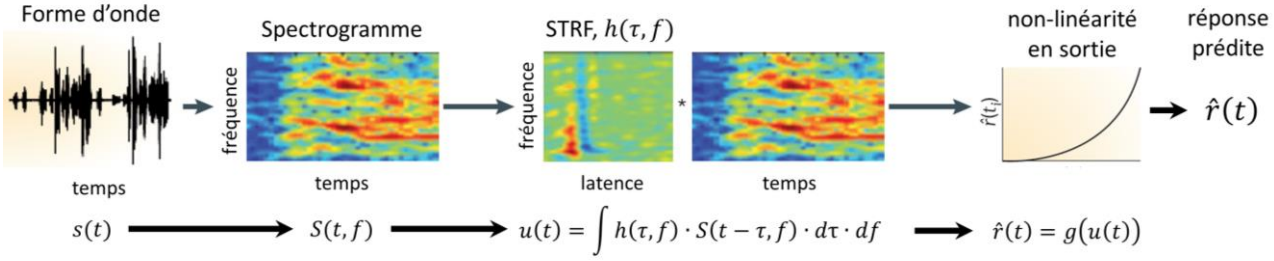


Figure 19 – Diagramme du modèle de neurone auditif basé sur le STRF, sans filtre post-décharge. Les étapes du traitement correspondent au calcul du spectrogramme $S(t, f)$ à partir de la forme d'onde, la convolution avec le STRF $h(\tau, f)$, puis la prédiction de la probabilité de décharge $\hat{r}(t)$ par une transformation non linéaire du résultat. Adapté de (Theunissen & Elie, 2014).

Le STRF correspond au profil de poids du Modèle Linéaire Généralisé (cf. paragraphe 7.5) prédisant le mieux les réponses du neurone à un ensemble de stimuli acoustiques. Il permet donc de mettre en évidence les structures spectro-temporelles du son auxquelles le neurone est le plus sensible. Il présente généralement des régions excitatrices (en rouge sur l'exemple présenté Figure 19), où la présence d'énergie accroît l'activité du neurone, ainsi que des régions inhibitrices (en bleu) qui diminuent la probabilité de décharge (Zhao & Zhaoping, 2011). Le STRF présenté Figure 19 correspond donc à un neurone détectant la présence d'une explosion d'énergie large-bande. Il est très précis temporellement mais peu sélectif en fréquences, car les régions excitatrices et inhibitrices sont étroites sur l'axe temporel mais étendues sur l'axe des fréquences. Ainsi, un STRF est souvent caractérisé par sa largeur de bande en temps

(*Best Rate*, BR) et en fréquence (*Best Scale*, BS), en plus de la fréquence à laquelle il répond préférentiellement (*Best Frequency*, BF) (Mesgarani et al., 2008).

Pour un neurone répondant de manière linéaire, le STRF représente le spectrogramme du son conduisant au taux de décharge le plus important. Cependant, dans la réalité, les réponses des neurones sont le plus souvent non linéaires et le STRF peut refléter les caractéristiques des stimuli utilisés, en plus de celles du neurone (Christianson et al., 2008). Cette technique demeure néanmoins un outil incontournable pour l'étude du système auditif.

6.2. L'évolution des STRFs le long de la voie auditive primaire

L'information acoustique de la parole est transmise le long de la voie auditive primaire depuis la cochlée jusqu'au colliculus inférieur, relais du tronc cérébral, puis au thalamus pour atteindre la partie du cortex temporal appelée Aire auditive primaire (A1). L'estimation de STRFs peut être réalisée à différents niveaux pour étudier l'encodage de l'information acoustique au fil de son traitement par le système auditif.

Dès le colliculus inférieur, les STRFs mesurés présentent une sélectivité pour certaines caractéristiques des sons naturels (Theunissen & Elie, 2014). De plus, Woolley et ses collaborateurs observèrent que les champs récepteurs des neurones auditifs du pinson présentent des structures de plus en plus complexes le long de la voie auditive primaire (Woolley et al., 2009). Tandis que le colliculus inférieur contient principalement des neurones rapides, plus ou moins spécifiques en fréquence, qui encodent l'attaque ou la présence d'une fréquence particulière, les neurones auditifs du thalamus présentent des latences plus grandes et encodent des caractéristiques sonores plus complexes comme la modulation de fréquence (Amin et al., 2010). Enfin le cortex auditif primaire voit l'apparition de STRFs plus lents et sensibles à des combinaisons d'indices (Theunissen & Elie, 2014). Des résultats cohérents ont été obtenus chez le chat (Atencio et al., 2012). Par ailleurs, il a été observé chez plusieurs mammifères que, au-delà du colliculus inférieur, les neurones auditifs deviennent hautement non linéaires et qu'ils encodent majoritairement des informations abstraites sur le signal reçu (comme la présence d'un objet auditif particulier ou une caractéristique pertinente du point de vue comportemental) (Atiani et al., 2014; Chechik & Nelken, 2012; Machens et al., 2004).

Les champs récepteurs présentent également une certaine plasticité qui leur permet de s'adapter à la tâche (Fritz et al., 2003), aux stimuli utilisés (Woolley et al., 2006), ou au bruit ambiant (Lesica & Grothe, 2008), réalisant ainsi un encodage dynamique qui maximise l'information transmise en minimisant l'énergie dépensée (Zhao & Zhaoping, 2011). Cette adaptabilité des STRFs, qui peut être reliée avec les résultats comportementaux décrits au paragraphe 5.1, explique en partie la grande robustesse du système de compréhension de la parole face aux dégradations du signal.

Chez le furet, l'insensibilité des neurones auditifs au bruit augmente progressivement le long de la voie auditive primaire : leurs fréquences de décharge sont de plus en plus indépendantes du SNR au fil des niveaux de traitement (Rabinowitz et al., 2013). Au niveau du cortex auditif primaire, la distribution des activités s'affranchit presque entièrement du bruit grâce à un mécanisme d'ajustement simultané des paramètres des neurones à la moyenne et à la variance du signal entrant (Mesgarani et al., 2014b).

6.3. Encodage de la parole par les STRFs

En complément des études chez l'animal concernant l'encodage d'un son (souvent une somme de fréquences modulées ou la vocalisation d'un congénère) par les STRFs le long de la voie auditive primaire, récemment, certains groupes de recherches se sont plus particulièrement intéressés à la manière dont ces champs récepteurs pourraient expliquer les comportements observés chez l'homme lors de la compréhension de la parole. Nous avons déjà évoqué précédemment le modèle de la représentation multi-résolution (cf. paragraphe 3.2.4) qui émet l'hypothèse selon laquelle l'excitation par un stimulus d'ensembles de neurones spécifiques, au sein d'une banque de STRFs, pourrait permettre la reconnaissance des phonèmes prononcés (Chi et al., 2005). Pour des raisons de modélisation, il s'agissait ici de champs récepteurs idéaux représentés par des filtres de Gabor couvrant tout l'espace BR x BS x BF. Comme il le fut démontré ensuite, les différents phonèmes sont relativement localisés dans cet espace (Mesgarani et al., 2014a, 2008; Nelken et al., 2014). En théorie, ce type d'encodage pourrait donc convenir pour la reconnaissance de la parole. En effet, la présentation de productions de différents phonèmes à des furets déclenche l'activation de groupes spécifiques de neurones, formes de représentations corticales des phonèmes. L'excellente corrélation de l'indice d'intelligibilité dérivé de ce modèle, le STMI (Elhilali et al., 2003), avec les taux de reconnaissance mesurés expérimentalement constitue une première preuve en faveur de la représentation multi-résolution pour la parole. La validité neurobiologique du modèle a été récemment corroborée par la reconstruction du signal de parole écouté par un patient, à partir de l'activité neuronale enregistrée au niveau de son Gyrus Temporal Supérieur (Pasley et al., 2012). L'équipe du Kight's lab montra ainsi que l'utilisation de la représentation multi-résolution dans ce cadre permettait une reconstruction fidèle du signal, suffisante pour identifier les mots au moyen d'un algorithme basique de reconnaissance de la parole.

L'utilisation par le système auditif d'un type de codage adapté au langage assure une sélectivité des neurones pour les caractéristiques acoustiques spécifiques aux signaux langagiers et absentes de la plupart des signaux de bruits (comme les transitions de formants). Autrement dit, les différents types de bruit possèdent des distributions caractéristiques dans l'espace des STRFs (Chi et al., 2005), ce qui permettrait au système d'isoler et d'éliminer spécifiquement les canaux corrompus (Mesgarani et al., 2011a, 2011b).

Ces quelques considérations sur les STRFs, chez l'animal et l'humain, nous renseignent sur les opérations potentiellement effectuées par le système auditif pour transformer le signal acoustique en information linguistique. Néanmoins, il est difficile d'établir un lien direct entre le profil des STRFs et les primitives parolières. Comme nous le verrons dans la partie Discussion (paragraphe 14.2.5), une solution pourrait être apportée par le calcul de « champs récepteurs comportementaux ». En effet, dans le domaine visuel, il est d'ores et déjà possible de calculer de telles représentations par la méthode des Images de Classification, exposée au chapitre suivant.

7. La méthode des Images de Classification

À ce stade de la partie théorique, nous avons présenté en détail l'état de la recherche sur les primitives acoustiques de la parole et constaté que leur identification demeurait encore une question ouverte faute d'une technique efficace permettant de les appréhender. Dans ce chapitre, nous nous écarterons un instant du cadre de la psychoacoustique pour nous intéresser à une problématique proche dans le domaine de la vision en décrivant comment certaines primitives visuelles peuvent être caractérisées au moyen d'une méthodologie récente nommée Image de Classification (*Classification Image*, CI).

L'exemple d'une étude menée par l'équipe de Gold en 2000 fournira ici un premier aperçu du principe de la CI (Gold et al., 2000). Dans cette expérience de psychophysique visuelle, les auteurs s'intéressaient au problème du traitement des contours illusoires, une illusion d'optique représentant un nombre restreint de formes espacées dans l'image, mais interprétées par le système visuel comme une figure géométrique simple au premier plan masquant d'autres figures simples au second plan. Dans le cas des carrés de Kanizsa présentés Figure 20B, par exemple, deux carrés bombés (respectivement « fin » et « épais ») apparaissent clairement en superposition sur quatre petits cercles blancs, alors que l'image n'est en fait composée que de quatre portions de cercle légèrement orientées vers l'intérieur ou vers l'extérieur.

Dans cette étude, Gold et ses collaborateurs mirent en place un paradigme ingénieux pour éclairer les traitements employés par le système visuel dans la reconstitution des contours illusoires. Pour chacune des conditions présentées Figure 20 (Contours complets, Contours de Kanizsa, Contours de Kanizsa masqué, ou Stimulus fragmenté), ils présentèrent à des participants une série de cibles T_0 ou T_1 dans un ordre aléatoire. Ces images étaient superposées à un bruit blanc gaussien, et les participants devaient à chaque essai identifier la cible contenue dans le stimulus. La question que se posaient les auteurs était la suivante : quelle information est utilisée par l'observateur pour effectuer cette catégorisation ? Se concentre-t-il exclusivement sur les parties du stimulus qui diffèrent entre les deux cibles (i.e. les portions de cercle) ou prend-il en compte d'autres régions de l'image ?

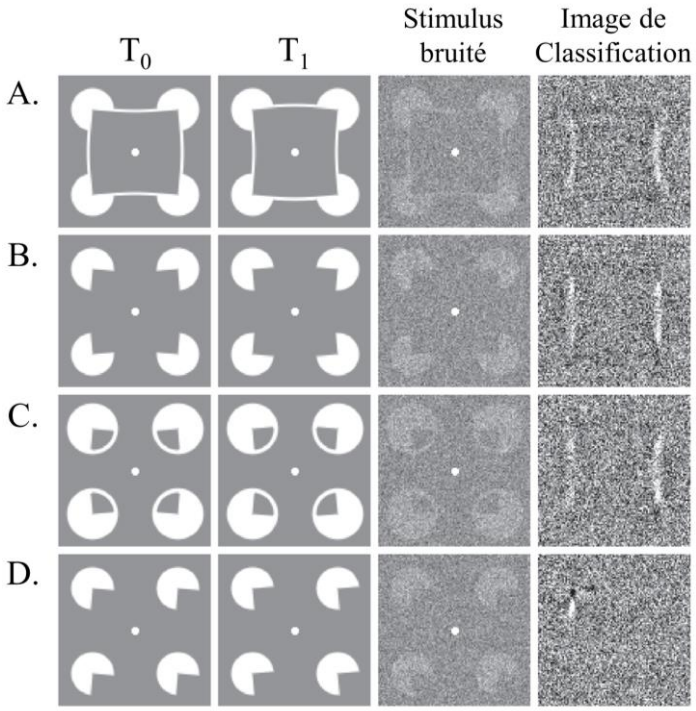


Figure 20 – Couples de cibles T0 et T1 utilisés dans l'expérience de perception des contours illusoires (Gold et al., 2000), pour les différentes conditions. Dans chaque cas, un exemple de stimuli et l'Image de Classification obtenue. A. Contours complets. B-C. Contours illusoires (contour de Kanizsa et contour de Kanizsa masqué). D. Stimulus fragmenté, sans contour illusoire. Adapté de (Murray et al., 2005).

La solution adoptée dans cet article fut de calculer, pour chaque participant et dans chaque condition, la corrélation entre la disposition exacte du bruit à chaque essai et la réponse correspondante du participant. La matrice de corrélation obtenue, appelée CI, permet de visualiser comment la présence de bruit en chaque point de l'image interfère avec la décision de l'observateur (c'est-à-dire quelles configurations du bruit influencent l'observateur en faveur d'une réponse particulière, T_0 ou T_1). Il apparaît clairement sur les CIs correspondant aux deux conditions Contour illusoire (Figure 20, B et C) que la décision du participant repose seulement sur certaines portions délimitées du stimulus (correspondant aux zones plus foncées et plus claires). La présence de bruit le long des contours illusoires verticaux du carré influe sur la catégorisation du stimulus, tandis qu'aucune corrélation n'est observée dans le reste de l'image.

Ainsi le bruit perturbe la décision de l'observateur, non seulement lorsqu'il est localisé à proximité de l'angle du carré de Kanizsa mais aussi tout le long de ses cotés subjectifs verticaux. Ce résultat offre une démonstration directe de la réalité perceptuelle des contours illusoires. On constate en effet que le traitement effectué par les participants pour juger de la forme du contour illusoire est très similaire à celui de la condition Contour complet (Figure 20A) pour laquelle les côtés du carré sont apparents.

Au contraire, la condition Stimulus fragmenté (Figure 20E), qui contient pourtant la même quantité d'information objective (quatre portions de cercle), est traitée différemment, les participants se concentrant exclusivement sur l'angle supérieur gauche. Le contour illusoire semble donc être véritablement reconstruit par le système visuel lors de la perception, sur la base des informations situées tout le long de son parcours, et traité de la même manière qu'un contour réel.

La méthode des CIs offre ici une visualisation directe des portions du stimulus visuel utilisées pour sa catégorisation. En faisant apparaître les zones de l'image dans lesquelles la présence de bruit impacte la décision de l'observateur, elle permet d'identifier les informations visuelles déterminantes pour le traitement de ce stimulus.

Le succès de l'application de cette méthode dans le contexte d'une tâche de catégorisation visuelle conduit à se demander si une transposition au problème des primitives auditives de la parole, évoqué dans les chapitres précédents, serait envisageable. En effet, il s'agit là aussi d'identifier les informations cruciales impliquées dans la catégorisation de stimuli, acoustiques dans ce cas. Pour déterminer si un tel projet est réalisable et quelles sont les contraintes à prendre en compte pour le calcul d'une variante auditive des CIs, il est nécessaire de s'intéresser dans un premier temps aux développements théoriques qui ont été réalisés dans ce champ de recherche durant les cinquante dernières années.

7.1. Formalisation du problème

Avant tout, la définition d'un cadre mathématique pour la description des CIs est indispensable, afin de rassembler toutes les études employant cette technique sous une seule et même formalisation, ce qui facilitera leur comparaison.

L'approche par la méthode des CIs est le plus souvent basée sur une tâche de catégorisation entre deux signaux-cibles, notés par les vecteurs $\underline{t}_0$ et $\underline{t}_1$ (dans le cas où la cible correspond à une image, on se ramène au cas unidimensionnel par une simple vectorisation). À chaque essai i ($\epsilon \, [\![1, N]\!]$), le stimulus $\underline{s}_i$ présenté au participant est composé de l'un de ces deux signaux, masqué par un bruit de fond (ou bruit externe) à un SNR donné, selon l'équation :

$$\underline{s}_i = \alpha_i \cdot \underline{t}_{k_i} + \underline{n}_i \qquad 7.1$$

avec k_i le numéro du signal associé à cet essai, $\underline{n}_i$ le bruit, et α_i un facteur permettant de fixer le SNR ($\alpha_i = 10^{\frac{SNR_i}{20}}$, pour $\underline{n}_i$ et $\underline{t}_{k_i}$ tous deux normalisés en puissance et SNR_i exprimé en dB). Après chaque présentation d'un stimulus, il est demandé au participant

d'indiquer la cible qu'il pense avoir reconnue. La réponse est notée r_i. Elle est le plus souvent binaire (= 0 pour $\underline{t}_0$ ou 1 pour $\underline{t}_1$).

Le calcul des CIs est parfois basé directement sur les composantes du stimulus $\underline{s}_i$ ou du bruit $\underline{n}_i$, mais pour la généralité du raisonnement nous introduisons ici la possibilité d'effectuer un changement de représentation noté φ, et transformant les signaux en $\underline{S}_i = \varphi(\underline{s}_i)$, $\underline{T}_i = \varphi(\underline{t}_i)$ et $\underline{N}_i = \varphi(\underline{n}_i)$. Dans le cas où cette transformation aboutit dans un espace bidimensionnel (par exemple un passage dans le domaine temps-fréquence), pour une meilleure lisibilité, on considère que les matrices résultantes, $\underline{\underline{S}}_i$, $\underline{\underline{T}}_i$ ou $\underline{\underline{N}}_i$, sont réorganisées sous leur forme vectorielle, $\underline{S}_i$, $\underline{T}_i$ ou $\underline{N}_i$,.

7.2. Premiers développements dans le domaine auditif

7.2.1. Psychophysique molaire et psychophysique moléculaire

Bien que la méthode des CIs ait été essentiellement développée pour des applications dans le domaine visuel, comme celle évoquée plus haut, il faut chercher les origines historiques de cette branche de la psychophysique dans le domaine acoustique. La première conceptualisation, encore relativement abstraite, de cette approche remonte à un article publié en 1964 formalisant l'opposition entre la psychophysique molaire (*molar psychophysics*) et la psychophysique moléculaire (*molecular psychophysics*) (Green, 1964).

Dans une expérience de psychophysique « standard », un même type de stimulation est présenté un grand nombre de fois à un (ou plusieurs) participants. Ces stimuli sont la plupart du temps différents mais ils partagent certaines propriétés communes (e.g. présence d'une même caractéristique acoustique, d'une même signification, d'une même structure…) qui permettent de les regrouper sous un petit nombre de conditions (facteurs contrôlés). Du fait de l'existence de facteurs non contrôlés (bruit ambiant, état mental du participant…), les réactions d'un même participant ne seront pas toujours identiques d'un essai à l'autre même sous des conditions expérimentales similaires. Pour contourner ce problème, la psychophysique molaire emploie une approche consistant à regrouper l'ensemble des réponses du ou des participants à un même type de stimuli et à les caractériser au moyen de mesures statistiques telles que la moyenne ou la variance (Gilkey & Robinson, 1986). Dans ce cas, la variable étudiée porte sur l'ensemble d'un grand nombre de jugements psychophysiques, le passage à la moyenne permettant de diminuer le poids des facteurs non contrôlés. Malheureusement, dans la plupart des cas, le regroupement des réponses correspondant à des essais différents implique la perte d'informations importantes, notamment celles concernant la nature exacte du stimulus présenté à chaque essai.

En opposition avec cette approche, Green proposa le concept de psychophysique moléculaire. Le principe est ici de déterminer les facteurs qui gouvernent le comportement de l'observateur en tentant de prédire sa réponse à chaque essai, en fonction des caractéristiques exactes du stimulus présenté. Naturellement, l'existence de facteurs non contrôlés interdit toute prédiction parfaite de la réaction à un stimulus donné. Néanmoins, selon l'auteur, une approche expérimentale utilisant la prédiction en essai-par-essai devrait permettre d'identifier les propriétés du stimulus qui influent sur la décision de l'auditeur[13].

Elle sera effectivement mise en application quelques années plus tard, dans une série de trois études, pour appréhender le problème de la détection de ton dans le bruit (*Tone-in-Noise detection*). En raison de son caractère fondamental dans le domaine de l'audition, cette tâche a été largement étudiée par la communauté psychoacoustique. Elle consiste à détecter un ton à une fréquence déterminée, présent sur la moitié des essais, et masqué par un bruit de fond large-bande présent sur la totalité des essais. À l'heure actuelle, après des décennies de recherche, il n'existe toujours pas de réponse complète et définitive à la question de la stratégie employée par un auditeur pour réaliser cette tâche. Plusieurs modèles ont été avancés, notamment le Détecteur Énergétique (qui exploite le fait que l'énergie totale du signal cible est supérieure à celle du signal non-cible), le Détecteur Spectral (qui opère une analyse différentielle du spectre du signal autour de la fréquence de la cible), ou le Détecteur d'Enveloppe (basé sur la différence entre les deux signaux dans le domaine des modulations de fréquence) mais aucun d'eux n'a permis jusque-là de prédire de manière fiable les réponses d'un participant à un stimulus donné (Schönfelder & Wichmann, 2012).

7.2.2. CIs par régression multiple

En 1971, Ahumada et Lovell abordèrent la problématique de la détection de ton dans le bruit par une méthode nouvelle (Ahumada & Lovell, 1971), précurseur du calcul des CIs mais également des fonctions de pondération évoquées au paragraphe 4.6. Dans cette expérience, les participants écoutaient une série de 50 stimuli bruités de 100 ms chacun, dans un ordre aléatoire, 25 d'entre eux contenant un ton pur à la fréquence 500 Hz ($\underline{t}_1$) et 25 contenant uniquement le bruit ($\underline{t}_0$). Il leur était demandé à chaque essai i de juger de la présence de la cible sur une échelle de 1 (absence certaine) à 4 (présence certaine) ($r_i \in \{1, 2, 3, 4\}$). Suivant le principe de la psychophysique moléculaire, les auteurs décidèrent ici de conserver un enregistrement de tous les stimuli présentés. Ceci leur permit dans un second temps de calculer l'énergie $\underline{N_i}$ de chaque bruit dans 5 bandes de fréquence situées de part et d'autre de la fréquence cible (ou 9 bandes de fréquence dans une seconde expérience), puis d'utiliser ces données pour tenter de prédire les

[13] Green note cependant déjà que ceci présuppose que l'observateur ne répond pas de manière aléatoire (c'est-à-dire que le poids des facteurs non contrôlés sur la décision reste limité), ce qui n'est pas assuré en pratique. Nous reviendrons sur ce point dans la partie Discussion de cette thèse (cf. paragraphe 14.2.2).

réponses du participant. L'espérance de la réponse, $E[r_i]$, est exprimée en fonction de la répartition du bruit au moyen d'une régression linéaire multiple, d'équation :

$$E[r_i] = \underline{N_i} * \underline{\beta} + \beta_0 \qquad 7.2$$

Les coefficients $\underline{\beta}$ sont interprétés comme des poids attentionnels associés par l'auditeur à chaque bande de fréquence pour la détection du ton cible, et β_0 comme un biais subjectif général en faveur de la perception d'un ton-cible. Ils sont obtenus en minimisant le critère des moindres carrés

$$\underline{\hat{\beta}}, \hat{\beta}_0 = \underset{\underline{\beta}, \beta_0}{\operatorname{argmin}} \sum_{i=1}^{N} (r_i - (\underline{N_i} * \underline{\beta} + \beta_0))^2 \qquad 7.3$$

Comme attendu, chez une majorité des participants, la représentation des coefficients de régression $\underline{\hat{\beta}}$ en fonction de la fréquence, appelée courbe de réponse fréquentielle, présente un pic positif à la fréquence de la cible, indiquant que la présence de bruit dans cette bande particulière augmente la probabilité que l'auditeur détecte un ton cible (que celui-ci soit effectivement présent ou non). Au contraire, les bandes de fréquences de part et d'autres ce pic sont souvent associées à des poids négatifs, la présence d'énergie autour de la cible empêche la détection de celle-ci. Ces résultats semblent donc plaider en faveur de l'hypothèse du Détecteur Spectral : l'attention des auditeurs se focalise sur une gamme de fréquence restreinte, hors de laquelle la présence de bruit n'influe pas sur le traitement. De plus, les pondérations négatives au-dessus et au-dessous de la fréquence cible correspondent à l'idée d'une analyse différentielle du spectre par le calcul de la différence entre l'énergie d'une bande de fréquence et celle des bandes adjacentes.

Néanmoins un certain nombre d'observations brouillaient le tableau. Tout d'abord, l'importante variabilité entre les courbes de réponse fréquentielle obtenues suggérait l'utilisation de plusieurs indices pondérés de manière différente par les participants : l'énergie dans la bande de fréquence correspondant à la cible, déjà évoquée, mais également l'énergie totale du stimulus et la répartition globale de l'énergie entre les hautes et les basses fréquences. Par ailleurs, Ahumada et Lovell remarquèrent que la restriction du calcul de la régression linéaire aux essais « cible » uniquement ou aux essais « non-cible » uniquement conduisait à des courbes de réponse fréquentielle légèrement différentes. Cette observation n'est pas compatible avec le modèle linéaire du Détecteur Spectral, censé appliquer le même traitement à tous les stimuli (voir paragraphe 7.4).

Ce type d'approche fut repris et amélioré par Ahumada et ses collaborateurs quatre ans plus tard. Plutôt que de caractériser la répartition spectrale du bruit seulement en termes de spectre, ils calculèrent une représentation temps-fréquence de son énergie, assez grossière, contenant 25 pixels (5 intervalles temporels x 5 bandes de fréquence). En utilisant la même équation que précédemment, $\underline{N_i}$ correspondant cette

fois à la représentation temps-fréquence vectorisée du bruit, ils montrèrent que la région influant le plus sur la réponse du participant était le pixel central, qui correspondait à la position spectro-temporelle de la cible. De plus, les participants associaient des poids négatifs aux régions précédant temporellement le signal et le bordant en fréquence. À nouveau, cette pondération négative indique que le critère de décision de l'auditeur correspond à une comparaison de l'énergie dans la région de la cible avec l'énergie dans les régions temporelles et fréquentielles adjacentes plutôt qu'à une mesure absolue de l'énergie. Enfin, comme dans l'expérience précédente, une différence de pondération est observée entre les essais « cible » et « non-cible », remettant en cause la linéarité du modèle.

Ces deux premières tentatives de prédiction essai-par-essai des réponses d'un participant (ici la présence ou l'absence d'un ton) sur la base de la répartition exacte du bruit masquant le signal, sont considérées comme les prototypes de la technique des CIs (Murray, 2011).

7.3. CIs visuelles

En principe, une telle méthode permettant d'analyser les portions d'un spectrogramme influant sur une catégorisation auditive dans le bruit pourrait s'avérer extrêmement utile pour la recherche des primitives auditives de la parole. Cependant, sa très faible résolution spectro-temporelle (5 x 5 pixels) proscrit toute application directe à des tâches plus complexes qu'une détection de ton pur. Les développements ultérieurs apportés à la technique des CIs dans le domaine visuel permirent d'augmenter le nombre de composantes prises en compte dans le modèle.

7.3.1. CIs par corrélation inversée

Après avoir laissé de côté son idée pendant quelques années, Ahumada trouva dans la question de l'Acuité de Vernier (*Vernier Acuity*) une nouvelle application possible de la psychophysique moléculaire (Ahumada, 1996; Beard & Ahumada, 1998). Dans cette expérience, deux configurations différentes étaient présentées aux participants. Tous les signaux contenaient un segment identique à gauche de l'image. La partie droite était quant à elle occupée par un autre segment, soit aligné avec le premier, soit légèrement décalé d'un pixel vers le haut (Figure 21). Comme précédemment ces images étaient présentées dans un bruit blanc visuel, et il était demandé à l'observateur de catégoriser les images alignée ($\underline{t}_1$) et non alignée ($\underline{t}_0$). En l'absence de bruit, le système visuel humain est remarquablement performant pour cette tâche, puisqu'il est capable de détecter des décalages inférieurs à sa résolution minimale imposée par le diamètre des photorécepteurs de la rétine, ce qui amena les chercheurs à utiliser le terme

d' hyperacuité (Li et al., 2006). Cette expérience visait à identifier les indices visuels utilisés dans la perception de l'alignement.

La partie gauche étant identique pour les deux signaux, elle ne permet pas de les discriminer, et un processus effectuant cette catégorisation de manière optimale doit donc se concentrer uniquement sur la partie droite de l'image, ainsi qu'on le constate en calculant la différence entre les deux signaux (Gabarit Idéal Figure 21).

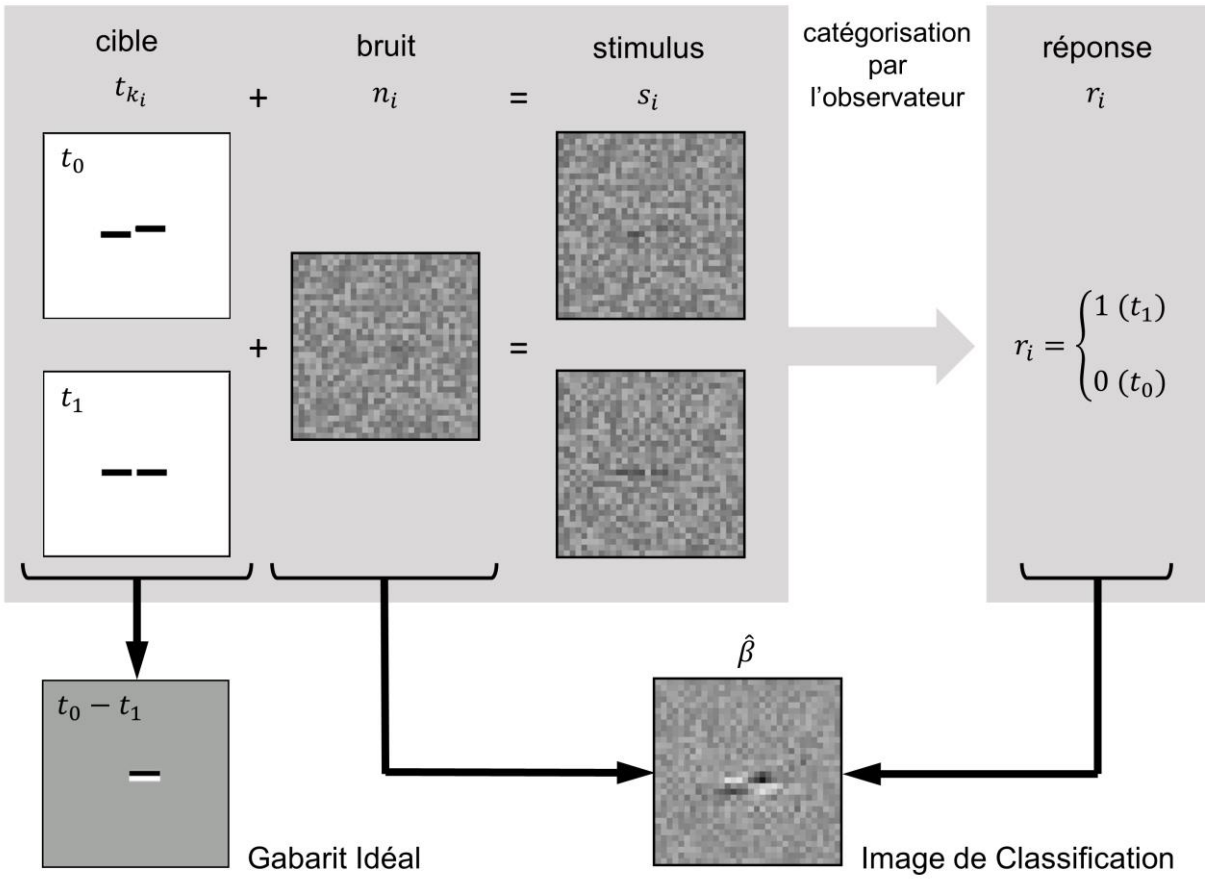


Figure 21 – Diagramme schématique du paradigme expérimental employé pour le calcul des CIs. Les exemples de cibles, stimuli, gabarit idéal et la CI correspondante sont tirés de la tâche d'acuité de Vernier (Ahumada, 1996).

Pour examiner les traitements effectivement mis en jeu par un observateur réel durant cette tâche, Ahumada proposa de réutiliser l'approche issue de la psychophysique moléculaire, à laquelle il donna pour la première fois le nom d' « Image de Classification ». Abandonnant la régression linéaire multiple, il choisit ici, plus simplement, de dériver une matrice (de dimensions 64 x 64) présentant la corrélation entre la luminance du bruit à chaque pixel et la réponse de l'observateur. Comme pour les études précédentes, les pixels possédant des valeurs positives ou négatives

importantes marquent les régions influant sur la catégorisation, tandis que les régions non impliquées dans le traitement du stimulus sont associées à des valeurs proches de zéro. Pour chaque pixel j, la corrélation entre la luminance du bruit, N_j, et la réponse binaire de l'observateur correspondante, r, est donnée par :

$$corr(N_j, r) = \frac{E[(N_j - E[N_j])(r - E[r])]}{\sigma_{N_j}\sigma_r} \quad 7.4$$

Cette équation peut être simplifiée, comme indiqué dans (Murray, 2011), en faisant l'hypothèse que l'observateur est non biaisé (c'est-à-dire $P(r = 0) = P(r = 1) = 0.5$, soit $E[r] = 0.5$), et en notant que le bruit est blanc et donc de moyenne nulle ($E[N_j] = 0$)

$$\begin{aligned} corr(N_j, r) &= \frac{E[N_j(r - 0.5)]}{\sigma_{N_j}\sigma_r} \\ &= \frac{E[N_j(r - 0.5)|r = 0] \cdot P(r = 0) + E[N_j(r - 0.5)|r = 1] \cdot P(r = 1)}{\sigma_{N_j}\sigma_r} \\ &= \frac{E[N_j(r - 0.5)|r = 0] + E[N_j(r - 0.5)|r = 1]}{2\sigma_{N_j}\sigma_r} \\ &= \frac{-0.5 \cdot E[N_j|r = 0] + 0.5 \cdot E[N_j|r = 1]}{2\sigma_{N_j}\sigma_r} \\ &= \frac{E[N_j|r = 1] - E[N_j|r = 0]}{4\sigma_{N_j}\sigma_r} \end{aligned} \quad 7.5$$

Le facteur $4\sigma_{N_j}\sigma_r$ étant constant, Ahumada choisit de définir la CI $\underline{\beta}$ par l'équation

$$\underline{\beta} = E[\underline{N}|r = 1] - E[\underline{N}|r = 0] \quad 7.6$$

$\underline{\beta}$ est donc proportionnel à la matrice de corrélation $corr(\underline{N}, r)$. En pratique, ce calcul revient à prendre simplement la moyenne des bruits correspondant aux essais pour lesquels l'observateur a donné la réponse 1 (« les deux segments sont alignés ») et la moyenne des bruits correspondant aux essais pour lesquels il a donné la réponse 0 (« les deux segments ne sont pas alignés »), puis à calculer la différence entre ces deux images. Cette approche, qui est aussi celle utilisée dans l'étude de (Gold et al., 2000) mentionnée plus haut, est appelée Image de Classification par corrélation inversée.

En appliquant cette méthode de calcul à la tâche d'acuité de Vernier (voir Figure 21), Ahumada constata tout d'abord que la moitié droite de la CI reproduisait bien la configuration attendue de la pondération, donnée par la différence entre les deux cibles (gabarit idéal), à savoir des poids négatifs sur le segment non aligné à droite et des poids positifs sur le segment aligné à droite. Cependant, le traitement effectué par l'observateur réel se révélait clairement différent de celui attendu d'un observateur

idéal. En effet, la partie gauche du stimulus, identique dans les signaux $\underline{t}_1$ et $\underline{t}_0$, apparaissait clairement jouer un rôle dans la décision puisqu'elle était affectée de poids non nuls. La stratégie adoptée par le participant pour réaliser cette tâche de détection de l'alignement est donc suboptimale puisqu'elle s'appuie en partie sur des régions de l'image ne contenant aucune information permettant de différencier les deux cibles. Les auteurs expliquèrent ce résultat par l'incertitude spatiale de l'observateur : celui-ci ne considère pas la position absolue du segment situé à droite pour prendre sa décision mais plutôt de sa position relative par rapport au segment de gauche. Dès lors, la présence de bruit masquant n'importe lequel des deux segments perturbe la catégorisation, expliquant la présence de poids non nuls sur les deux moitiés de la CI.

7.3.2. Exemples d'applications des CIs dans le domaine visuel

Le succès de l'application des CIs à la tâche d'acuité de Vernier provoqua un vif engouement de la communauté scientifique pour cette nouvelle technique (Eckstein & Ahumada, 2002). Dans le domaine de la psychophysique visuelle de nombreuses questions furent abordées par la suite au moyen de la méthode de corrélation inversée ou de ses dérivées que nous présenterons plus loin :

- l'acuité de Vernier (Ahumada, 1996; Ahumada & Beard, 1998; Barth et al., 1999; Beard & Ahumada, 1998; Li et al., 2006)

- la perception de la forme, avec les contours illusoire (Gold et al., 2000; Nagai et al., 2008) et l'intégration de contour (Kurki et al., 2014)

- la perception de motifs simples : profils gaussiens (Abbey & Eckstein, 2002, 2006; Mineault et al., 2009; Solomon, 2002), ondelettes de Gabor (Ahumada, 2002; Beard & Ahumada, 1999; Solomon, 2002) et damiers (Beard & Ahumada, 1999)

- le traitement de stimuli possédant un décours temporel, comme les modulations temporelles chromatiques (Bouet & Knoblauch, 2004) ou de luminance (Thomas & Knoblauch, 2005), la détection d'une cible immobile dans un bruit spatiotemporel (Neri & Heeger, 2002) et la détection d'une cible en mouvement dans un espace tridimensionnel (Neri & Levi, 2008), ou encore la reconnaissance des mouvements biologiques (van Boxtel & Lu, 2015)

- l'identification de lettres (Liu et al., 2014; Morin Duchesne et al., 2014; Nandy & Tjan, 2007; Rieth et al., 2011; Watson & Rosenholtz, 1997)

- la perception de la luminosité (Kurki et al., 2009; Thomas & Knoblauch, 2005) ou de la couleur (Bouet & Knoblauch, 2004; Thorsten Hansen, 2005)

- le traitement des visages et la reconnaissance des émotions faciales (Gold et al., 2004; Kontsevich & Tyler, 2004; Liu et al., 2014; Mangini & Biederman, 2003;

Martin-Malivel et al., 2006; Nagai et al., 2013; Rieth et al., 2011; Sekuler et al., 2004; Wichmann et al., 2005)

- l'apprentissage perceptuel, sur des motifs abstraits ou des visages (Gold et al., 2004), sur une tâche de discrimination d'orientation (Kurki & Eckstein, 2014), ou d'acuité de Vernier (Li et al., 2004)

- l'identification d'objets (Olman & Kersten, 2004; Vondrick et al., 2014)

- la perception multimodale (Pascucci et al., 2011)

- la vision stéréoscopique (Neri & Levi, 2008; Neri et al., 1999)

7.4. Le modèle de l'observateur linéaire

Le calcul de la CI par corrélation inversée est une approche à la fois simple et intuitive permettant d'identifier les indices utilisés dans une tâche de catégorisation visuelle. En effet, il nécessite simplement de présenter de manière répétée deux signaux dans un bruit blanc, en demandant au participant de les catégoriser, puis de regrouper les bruits en fonction des réponses correspondantes de l'observateur. Une moyenne de tous les essais pour lesquels le participant a répondu 1 (respectivement 0) permet de faire ressortir les particularités du bruit qui biaisent la catégorisation en faveur de $\underline{t}_1$ (respectivement $\underline{t}_0$).

Cette description du principe de la corrélation inversée masque toutefois un certain nombre d'hypothèses implicites. On sait notamment que la corrélation, sur laquelle est basé le calcul, ne peut mettre en évidence que les relations linéaires existant entre les deux variables. Les effets non linéaires du bruit sur la catégorisation ne sont donc pas représentés sur la CI, ou seulement de manière indirecte. Pour faire ressortir les limitations de cette méthode il est nécessaire d'étudier le modèle de détection sous-jacent nommé observateur linéaire.

Le modèle de l'observateur linéaire est issu de la Théorie de la Détection du Signal (*Signal Detection Theory*, SDT), un cadre théorique qui décrit l'extraction d'un signal noyé dans le bruit et développe un certain nombre de méthodes et d'outils fondamentaux pour la psychophysique moderne (Abdi, 2007; Knoblauch & Maloney, 2012). Il permet de caractériser les différentes stratégies linéaires pouvant être adoptées lors d'une tâche de catégorisation. D'après ce modèle, la détection se décompose en deux étapes (Harvey, 2004) :

1) Le traitement sensoriel transforme la stimulation physique multidimensionnelle $\underline{s}$ en une variable interne d unidimensionnelle. Le modèle suppose que cette opération est effectuée par comparaison à un gabarit (*template-matching*)

(Ahumada, 2002; Knoblauch & Maloney, 2012; Murray, 2011). Ce gabarit, $\underline{\beta}$, stocké en mémoire et de mêmes dimensions que la représentation du stimulus $\underline{S}$, indique le poids perceptuel β_j de chaque composante S_j du stimulus pour la décision finale. Plus précisément, l'observateur linéaire calcule, au moyen d'un produit scalaire, la ressemblance de $\underline{S}$ avec un gabarit idéal $\underline{\beta}$ stocké en mémoire:

$$d = \underline{S} * \underline{\beta} + \varepsilon \qquad 7.7$$

Le terme ε, nommé bruit interne, correspond aux facteurs aléatoires dépendants du participant et pouvant venir parasiter le processus (problème attentionnel, état mental au moment de la tâche…). Par la suite, en accord avec la littérature, ε sera supposé gaussien de variance σ_ε^2 et indépendant de la valeur de $\underline{S} * \underline{\beta}$.

2) Dans un second temps a lieu la décision à proprement parler. Elle repose simplement sur une comparaison à un seuil c (aussi appelé critère de décision) transformant la variable interne continue en une réponse booléenne :

$$r = \begin{cases} 0 \; si \; d > c \\ 1 \; sinon \end{cases} \qquad 7.8$$

En moyenne, le modèle a une plus grande probabilités de répondre « 0 » pour les stimuli ressemblant le plus à $\underline{\beta}$. À l'inverse, il a une plus grande probabilité de répondre « 1 » pour les stimuli ressemblant le plus à $-\underline{\beta}$.

L'observateur linéaire est donc entièrement défini par la donnée de $\left\{\underline{\beta}, c\right\}$. Comme tout modèle, il ne permet qu'une estimation des données réelles. Néanmoins, sa simplicité en fait un outil intéressant pour essayer de comprendre la stratégie d'un observateur réel dans une tâche de catégorisation : la détermination du couple de paramètres $\left\{\underline{\hat{\beta}}, \hat{c}\right\}$ qui prédit le mieux ses réponses nous renseigne sur l'importance accordée par le participant aux différentes composantes du stimulus.

7.4.1. Observateur idéal

Parmi l'ensemble des observateurs linéaires, l'observateur idéal est celui qui possède un taux maximal de catégorisations correctes (Ahumada, 2002; Eckstein et al., 1997; Knoblauch & Maloney, 2012). À ce titre, il constitue un point de comparaison pour l'évaluation et l'interprétation des autres modèles. Sa stratégie consiste à calculer la ressemblance du le stimulus perçu $\underline{S}$ avec les deux signaux cibles ($\underline{t}_0$ et $\underline{t}_1$) à discriminer :

$$\begin{aligned} d_0 &= \underline{S} * \underline{t}_0 + \varepsilon_0 \\ d_1 &= \underline{S} * \underline{t}_1 + \varepsilon_1 \end{aligned} \qquad 7.9$$

puis à donner pour réponse le signal qui ressemble le plus au stimulus

$$r = \begin{cases} 0 \; si \; d_0 > d_1 \\ 1 \; sinon \end{cases} \quad 7.10$$

Il est alors possible de réécrire les équations précédentes sous la forme de l'observateur linéaire présenté plus haut :

$$r = \begin{cases} 0 \; si \; d_0 - d_1 > 0 \\ 1 \; sinon \end{cases} \quad 7.11$$

En posant $c = 0$ et

$$\begin{aligned} d &= d_0 - d_1 \\ &= \left(\underline{S} * \underline{t}_0 + \varepsilon_0\right) - \left(\underline{S} * \underline{t}_1 + \varepsilon_1\right) \\ &= \underline{S} * (\underline{t}_0 - \underline{t}_1) + \varepsilon \end{aligned} \quad 7.12$$

L'observateur idéal correspond donc à un observateur linéaire de paramètres $\left\{\underline{\beta} = \underline{t}_0 - \underline{t}_1, \; c = 0\right\}$. Il est non biaisé ($c = 0$) et prend en compte uniquement les parties du stimulus qui permettent de différencier les deux signaux (gabarit idéal $\underline{\beta} = \underline{t}_0 - \underline{t}_1$). Autrement dit, il affecte un poids nul aux composantes j pour lesquelles les cibles sont identiques $(t_0)_j = (t_1)_j$.

7.4.2. La CI comme estimateur de l'observateur linéaire

Comme évoqué plus haut, notre but est de déterminer l'observateur linéaire dont les réponses prédisent au mieux celles d'un observateur réel, car la connaissance du gabarit associé à ce modèle nous renseignera sur la stratégie de catégorisation employée par le participant, en nous indiquant notamment les parties du signal qui influent sur la décision. Abbey et Eckstein démontrèrent que, dans ce cas, la méthode de la CI par corrélation inversée fournit un estimateur non biaisé $\underline{\hat{\beta}}$ du gabarit $\underline{\beta}$ à partir des réponses de l'observateur linéaire, à condition que $\underline{N}$ soit un bruit gaussien indépendant multivarié (Abbey & Eckstein, 2001)[14]. Il est également possible d'obtenir un estimateur du critère de décision $\hat{c}$, comme indiqué dans (Eckstein & Ahumada, 2002). Néanmoins, le biais de l'observateur est le plus souvent exprimé en termes de pourcentage de réponses (les participants ne totalisant pas un nombre égal de réponses « 0 » et « 1 » sont considérés comme biaisés en faveur d'une des cibles).

Les calculs présentés ci-dessus justifient l'utilisation de la technique de corrélation inversée pour l'estimation d'CIs visuelles. Néanmoins cette approche présente certaines limitations :

1) Comme son nom l'indique, la corrélation inversée « inverse » le rôle des facteurs habituellement utilisés en psychophysique (Knoblauch & Maloney, 2008, 2012) :

[14] Le lecteur pourra trouver une représentation graphique de cette démonstration dans (Murray, 2011)

ici, la réponse du participant est considérée comme la variable indépendante (la condition), tandis que le bruit tient lieu de variable dépendante. De ce point de vue, la régression linéaire (équation 7.2) se révèle d'avantage intuitive car le bruit permet de prédire la réponse, et non l'inverse.

2) Le nombre de catégorisations de stimuli bruités nécessaires pour obtenir une estimation fiable est très important : jusqu'à 8000 essais pour (Ahumada, 2002), 9600 essais pour (Gold et al., 2004), 10000 essais pour (Gold et al., 2000; Murray et al., 2002). Une telle durée d'expérience peut s'avérer problématique pour des raisons de fatigue mentale.

3) Comme mentionné ci-dessus, la corrélation inversée suppose que $\underline{N}$ possède une distribution gaussienne. Ceci est vrai dans le cas visuel où la CI est dérivée directement du bruit blanc $\underline{n}$ présenté au participant ($\underline{N} = \underline{n}$). Dans la perspective du calcul de CIs auditives, en revanche, cette hypothèse est très contraignante car la transformation φ reliant $\underline{n}$ et $\underline{N}$ ne correspond pas à l'identité, comme nous le verrons plus loin.

Ces trois limitations fortes de la corrélation inversée motivèrent la recherche d'une nouvelle méthode pour l'obtention de CIs, fondée sur le Modèle Linéaire Généralisé.

7.5. Le Modèle Linéaire Généralisé

7.5.1. Définition et ajustement du Modèle Linéaire Généralisé

En 2008, Knoblauch et Maloney proposèrent une nouvelle approche théorique des CIs en démontrant que celles-ci s'inscrivent naturellement dans le cadre d'un Modèle Linéaire Généralisé (*Generalized Linear Model***,** GLM) (Knoblauch & Maloney, 2008). Le GLM est une extension de la régression linéaire (comme celle présentée équation 7.2) dans le cas où la variable dépendante n'est pas normalement distribuée (Fox, 2008; Knoblauch & Maloney, 2012; Wood, 2006). Effectivement, l'équation linéaire 7.2 présente un problème majeur : les prédicteurs (les composantes de $\underline{N_i}$) sont des variables continues suivant une distribution normale, tandis que la variable dépendante (r_i) est une variable discrète à valeurs dans {0,1}. Il est donc fondamentalement impossible de trouver une combinaison linéaire qui puisse prédire la variable dépendante de manière réaliste, car il existera toujours des valeurs de $\underline{N_i}$ prédisant une espérance de r_i supérieure à 1 ou inférieure à 0. Ceci viole l'hypothèse d'exogénéité du modèle linéaire selon laquelle le terme d'erreur est indépendant des variables explicatives. Le GLM corrige ce problème en introduisant une fonction de lien f non linéaire entre les deux membres de l'équation linéaire classique :

$$E[\mathrm{Y}] = f\left(\underline{X} * \underline{B} + b_0\right) \tag{7.13}$$

avec $E[\mathrm{Y}]$ suivant une distribution normale, binomiale, poisson, gamma, ou inverse-gaussienne (famille des distributions exponentielles) (Fox, 2008).

Knoblauch et Maloney montrèrent que le modèle de l'observateur linéaire se ramène assez directement à une équation de la forme ci-dessus. En effet,

$$\begin{aligned} E[r_i] &= P(r_i = 1) \\ &= P(d_i \leq c) \\ &= P\left(\underline{S_i} * \underline{\beta} + \varepsilon \leq c\right) \\ &= P\left(\varepsilon \leq c - \underline{S_i} * \underline{\beta}\right) \\ &= P\left(-\varepsilon \geq \underline{S_i} * \underline{\beta} - c\right) \end{aligned} \tag{7.14}$$

ε étant une variable aléatoire normale centrée de variance σ_ε^2, la probabilité qu'elle soit supérieure à une certaine valeur x est donnée par la loi normale cumulative notée $\Phi(x, 0, \sigma_\varepsilon)$, soit

$$\begin{aligned} P(r_i = 1) &= \Phi\left(\underline{S_i} * \underline{\beta} - c, 0, \sigma_\varepsilon\right) \\ &= \Phi\left(\underline{S_i} * \underline{\beta}/\sigma_\varepsilon - c/\sigma_\varepsilon, 0, 1\right) \end{aligned} \tag{7.15}$$

Au final, en omettant le facteur de normalisation $1/\sigma_\varepsilon$ (car nous sommes uniquement intéressés par les valeurs relatives des poids)

$$E[r_i] = \Phi\left(\underline{S_i} * \underline{\beta} - c, 0, 1\right) = \Phi\left(\underline{S_i} * \underline{\beta} - c\right) \tag{7.16}$$

en adoptant la notation $\Phi(x) = \Phi(x, 0{,}1)$.

Cette dernière équation suit la forme générale de l'équation 7.13 et, par ailleurs, $E[r_i]$ admet une distribution binomiale ; il s'agit donc bien d'un GLM. Ce modèle présente l'avantage d'offrir un cadre théorique très développé, et un certain nombre d'outils statistiques dédiés. Par ailleurs, des algorithmes pour l'ajustement des paramètres du GLM ont été implémentés sur les principaux logiciels statistiques (fonction 'glmfit' sous MATLAB, fonction 'glm' sous R). Comme dans le cas linéaire, les paramètres du modèle $\underline{\theta} = \left\{\underline{\beta}, c\right\}$ qui correspondent le mieux aux données empiriques sont déterminés alors par une simple maximisation de la vraisemblance (Eliason, 1993) :

$$L(\underline{\theta}) = P\left(\underline{r}\middle|\underline{\theta}, \underline{\underline{S}}\right) \tag{7.17}$$

puis, les réponses r_i étant supposées indépendantes,

$$L(\underline{\theta}) = \prod_i P(r_i|\underline{\theta}, \underline{S_i}) \tag{7.18}$$

En pratique, pour simplifier le calcul, on cherche plutôt le minimum de l'opposé du logarithme de cette fonction (*negative log-likelihood*):

$$\begin{aligned} l(\underline{\theta}) &= -\log\left(L(\underline{\theta})\right) \\ &= -\sum_i \log\left(P(r_i|\underline{\theta}, \underline{S_i})\right) \\ &= -\sum_i \log\left(\Phi\left(\underline{S_i} * \underline{\beta} - c\right)\right) \end{aligned} \qquad 7.19$$

La fonction ci-dessus possède un minimum unique qui peut être trouvé par les algorithmes classique d'optimisation (Nickisch, 2012), le plus commun dans ce cas étant l'algorithme de Newton-Raphson appelé ici *Iteratively Re-weighted Least Squares* (Nelder & Wedderburn, 1972; Woodford & Phillips, 2011; Wood, 2006). Les valeurs des paramètres correspondant au minimum, $\underline{\hat{\theta}} = \text{argmin}_{\underline{\theta}}\left(l(\underline{\theta})\right)$, permettent d'obtenir le modèle $\underline{\hat{\theta}} = \left\{\underline{\hat{\beta}}, \hat{c}\right\}$ s'approchant au plus près des données empiriques. On obtient ainsi une estimation de la CI ($\underline{\hat{\beta}}$) et du critère de decision ($\hat{c}$).

Cette approche basée sur l'optimisation d'un GLM présente un certain nombre d'atouts par rapport aux techniques développées précédemment.

1) Le GLM est plus efficace relativement au nombre d'essais nécessaires pour obtenir une estimation avec une précision donnée. Plus précisément, en simulant un observateur linéaire sans bruit interne, Knoblauch et Maloney ont démontré qu'il suffisait de 1000 essais au GLM contre 10000 pour la corrélation inversée pour obtenir la même variance résiduelle entre la CI et le gabarit réel, c'est-à-dire pour une qualité d'estimation équivalente (Figure 22A). L'ajout de bruit interne fait progressivement perdre cet avantage (Figure 22B) ; cependant les auteurs notent la qualité d'estimation du GLM n'est jamais inférieure à celle de la corrélation inversée, et peut éventuellement être meilleure lorsque le bruit interne reste limité. Ce dernier point sera confirmé par la suite par Kurki et Eckstein (Kurki & Eckstein, 2014).

2) Ce modèle offre une formulation du problème non seulement plus correcte statistiquement, mais également plus intuitive que la précédente. En effet, il prédit la réponse de l'observateur en fonction des caractéristiques du stimulus présenté, et non l'inverse (Knoblauch & Maloney, 2008). Pour ces deux raisons, le GLM est une méthode largement utilisée en psychophysique. Dans sa variante la plus simple, la régression logistique, il est employé entre autres pour le calcul des fonctions psychométriques (Kingdom and Prins, 2010; Treutwein and Strasburger, 1999; Wichmann and Hill, 2001) ou des courbes de réponses fréquentielles et temporelles (Oberfeld et al., 2012; Pedersen and Ellermeier, 2008), deux outils très proches des CIs.

3) Le GLM fournit un cadre théorique très étendu, permettant de relier le calcul des CIs et celui des champs récepteurs des neurones visuels – et donc potentiellement d'établir un pont entre le niveau comportemental et le niveau neurobiologique. En effet, certains auteurs ont proposé de modéliser le taux décharge Y d'un neurone au moyen d'un GLM, la principale différence avec le calcul ci-dessus étant que la distribution de $E[Y]$ suit dans ce cas une loi de Poisson et non plus une loi binomiale (Paninski et al., 2007; Pillow, 2007; Pillow et al., 2008; Vidne et al., 2012). Nous reviendrons sur ce parallèle dans la Discussion de cette thèse (paragraphe 14.2.5).

4) Bien que la complexité de l'algorithme du GLM soit supérieure, en raison de l'étape d'optimisation, son coût en temps de calcul reste cependant dérisoire (Knoblauch & Maloney, 2008).

5) Le GLM n'impose pas d'hypothèse sur la distribution des prédicteurs, contrairement à la corrélation inversée (cf. paragraphe 7.4). En pratique, cela signifie que le bruit $\underline{N}$ ne doit pas nécessairement obéir à une loi gaussienne. Ceci permet non seulement de calculer des CIs dans des cas où le masque n'est pas un bruit blanc, mais également de baser le calcul sur des représentations φ plus sophistiquées puisqu'aucune distribution particulière des $\underline{S_i}$ n'est imposée. Il s'agit là d'un atout majeur dans la perspective d'un calcul de CIs auditives car il nous autorise à utiliser une représentation temps-fréquence du son, plutôt que la seule représentation en forme d'onde.

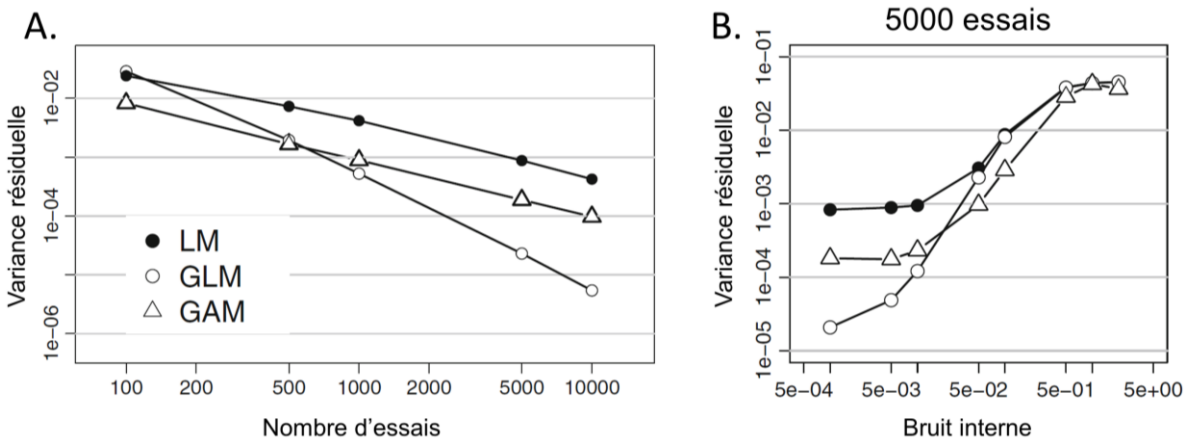


Figure 22 – Effets du nombre d'essais et du bruit interne sur la qualité de l'estimation, exprimée en variance résiduelle (distance entre la CI et le gabarit réel), pour un observateur linéaire simulé. Trois méthodes sont représentées : la corrélation inversée (cercle noir), le GLM (cercle blanc), et le GLM pénalisé ou GAM (triangle blanc). A. Évolution de la variance résiduelle en fonction du nombre d'essais pour un observateur linéaire sans bruit interne. B. Évolution de la variance résiduelle en fonction du bruit interne pour une passation de 5000 essais. Adapté de (Knoblauch & Maloney, 2012).

7.5.2. Transformations dans l'espace des stimuli

Dans ce qui précède, la transformation φ, qui permet d'obtenir les prédicteurs du modèle à partir des stimuli ($\underline{S_i} = \varphi(\underline{s_i})$, $\underline{T_i} = \varphi(\underline{t_i})$ et $\underline{N_i} = \varphi(\underline{n_i})$), a toujours été considérée égale à l'identité, c'est-à-dire que les CIs sont dérivées directement des stimuli présentés à l'observateur. Néanmoins quelques études ont envisagé la possibilité d'appliquer une transformation non linéaire aux stimuli avant de calculer la CI. Cette phase de pré-traitement ne change rien du point de vue statistique, le GLM étant indifférent à la distribution des prédicteurs. En revanche, du point de vue psychophysique, elle revient à considérer un type d'observateur se basant sur des propriétés différentes du signal (par exemple l'énergie au lieu de l'amplitude).

La fonction φ la plus utilisée dans ce contexte est la transformée de Fourier (Thomas & Knoblauch, 2005) ou d'autres représentations en fréquence (Ahumada et al., 1975), ce qui revient à considérer que l'observateur fonde sa décision non sur les échantillons temporels mais sur les propriétés spectrales du stimulus.

7.5.3. Qualité de la prédiction

Une fois ajusté, le GLM constitue un modèle qui permet de prédire la réponse d'un participant à la présentation d'un stimulus particulier. Pour un observateur dont la CI $\underline{\hat{\beta}}$ a été estimée au préalable, la réponse r_i' prédite par le GLM pour le stimulus $\underline{S_i}$ est donnée par

$$r_i' = \begin{cases} 1 \; si \; \Phi\left(\underline{S_i} * \underline{\beta} - c\right) > 0.5 \\ 0 \; sinon \end{cases} \qquad 7.20$$

En effet, $\Phi\left(\underline{S_i} * \underline{\beta} - c\right)$ correspond à la probabilité que le participant donne la réponse 1, $P(r_i = 1)$. Cette prédiction en essai-par-essai est le plus souvent imparfaite, et il est donc possible d'évaluer objectivement la qualité du modèle en calculant son erreur de prédiction.

Une première mesure classique est l'erreur quadratique moyenne (*Mean Squared Error*, MSE) calculée à partir des prédictions du modèle sur un ensemble de N stimuli :

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(r_i' - r_i)^2 \qquad 7.21$$

r_i' et r_i ne prenant leurs valeurs que dans {0,1}, la MSE correspond donc dans ce cas au taux de prédictions incorrectes.

En pratique, la MSE peut être calculée pour des prédictions provenant d'une source quelconque, ce qui en fait un bon outil pour comparer différents modèles statistiques. En revanche, pour un GLM, cette mesure n'est pas la mieux adaptée, car elle entraîne une perte conséquente d'information : une probabilité $\Phi\left(\underline{S_i} * \underline{\beta} - c\right)$ proche de 0 ou de 1 représente une prédiction plus forte sur le comportement du participant qu'une probabilité légèrement supérieure ou légèrement inférieure à 0.5. Pour pallier ce problème, une autre mesure a été développée dans le cadre statistique du GLM : la déviance D (Wood, 2006). Elle est définie simplement par :

$$D = 2 \cdot \left(l(\underline{\theta}) - l_{sat}\right) \qquad 7.22$$

avec $l(\underline{\theta})$ le *negative log-likelihood* du modèle (cf. équation 7.19) et l_{sat} sa valeur minimale atteignable pour les données considérées, obtenue pour un modèle prédisant au mieux les réponses du participant (par exemple, un modèle saturé possédant autant de paramètres que d'essais à prédire). D'après l'équation 7.19, cette déviance est effectivement calculée dans l'espace des probabilités de réponse, et non plus des réponses, ce qui en fait une mesure plus précise de la qualité de la prédiction d'un GLM que la MSE.

7.6. Réduction du bruit d'estimation

En pratique, le calcul d'une CI est toujours associé à une certaine erreur d'estimation, qui se traduit par la superposition d'un bruit sur le profil de poids $\underline{\hat{\beta}}$ par rapport au gabarit réel recherché $\underline{\beta}$. Cette erreur peut-être aggravée par le non-respect des hypothèses de la SDT ou la présence d'un important bruit interne ε (Figure 22B). Naturellement, il n'est pas possible d'influer sur ces facteurs dépendant exclusivement du participant. En revanche, d'autres solutions ont été proposées pour réduire le bruit d'estimation.

Plus précisément, il s'agit en fait de trouver un compromis entre le nombre d'essais nécessaires au calcul et la qualité attendue de l'image. En effet, comme pour tous les estimateurs statistiques, l'augmentation de la taille de l'échantillon accroît la qualité de l'estimation : quelle que soit la méthode choisie, un plus grand nombre d'essais permet d'obtenir une CI plus proche du gabarit réel (Figure 22A). Cependant, la durée de la passation étant déjà extrêmement longue dans les expériences décrites précédemment, il n'est pas pensable de l'accroître encore. Une solution envisageable consisterait éventuellement à combiner les données de plusieurs participants, une possibilité qui sera examinée au paragraphe 7.8.2. Dans l'immédiat, nous nous intéresserons uniquement aux procédés visant à améliorer la qualité de l'estimation à nombre d'essais constant.

7.6.1. Diminution de la résolution

Le nombre extrêmement élevé de prédicteurs pris en compte (i.e. le nombre de pixels de l'image) est l'une des causes de l'erreur d'estimation. De manière générale, les modèles statistiques comprenant un nombre de degrés de liberté très important sont sujets au surapprentissage (*overfitting*) : le modèle peut refléter fidèlement les données sur lesquelles il a été entraîné, mais se montrer incapable de prédire des données inédites engendrées par le même processus.[15] Dans le cadre de cette étude, cela signifie que les paramètres de la CI pourraient être déterminés davantage par la distribution particulière du bruit dans le set de stimuli présentés que par le mécanisme de catégorisation sous-jacent. Ainsi diminuer le bruit d'estimation nécessite de réduire la nombre de paramètres indépendants du modèle (Murray et al., 2002).

Une première solution évidente consisterait à limiter la dimension de la matrice de bruit à une zone plus restreinte autour de la cible. Dans l'expérience d'Ahumada décrite au paragraphe 7.3, par exemple, il semble évident que l'observateur ne s'appuiera pas sur la luminance du coin supérieur gauche pour prendre une décision concernant l'alignement des deux segments au centre de l'image. Ainsi ce pixel contribue au bruit d'estimation mais non à l'estimation du gabarit, et il est donc tentant de le laisser de côté en restreignant le calcul à la zone centrale autour des cibles. Néanmoins, ceci équivaut à introduire dans le processus d'estimation un a priori fort quant au résultat attendu, un compromis que nous cherchons à éviter autant que possible. Aussi les psychophysiciens choisirent-ils, le plus souvent, de diminuer globalement la résolution du bruit en le sous-échantillonnant, soit en regroupant les pixels par régions (Nagai et al., 2008), soit en n'affichant à chaque essai qu'une portion restreinte des pixels (Nagai et al., 2013).

7.6.2. Bruit dimensionnel

Une manière plus élégante de procéder consiste à appliquer le bruit non à la luminance de l'image mais à une autre dimension du stimulus. Pour la question de l'acuité de Vernier, cette approche, appelée bruit dimensionnel, a été exemplifiée par Li et ses collaborateurs (Li et al., 2006). Dans cette étude, comme pour celle d'Ahumada, tous les signaux possédaient une moitié gauche identique composée d'un segment à une position fixe. La moitié droite quant à elle comportait un segment divisé en 5, soit globalement aligné avec le premier, soit légèrement au-dessus ou au-dessous. Le bruit correspondant à chaque essai s'appliquait à la hauteur relative de chacune des 5 portions de segment, en le décalant verticalement par rapport à la position moyenne du

[15] Un modèle suffisamment complexe pourra, par exemple, stocker en mémoire l'ensemble des appariements stimulus-réponse qui lui sont présentés durant l'entraînement. Dans ce cas, il sera en mesure de restituer les données d'apprentissage mais ne pourra en aucun cas généraliser de manière fiable sa prédiction à de nouvelles données.

segment. L'observateur avait pour tâche de juger de l'alignement des deux segments. Contrairement à l'expérience d'Ahumada qui requérait l'estimation de 16384 paramètres (128 x 128 pixels) sur un total de 11398 essais, ce nouveau paradigme n'implique plus que 5 prédicteurs, et l'ajustement est réalisé en seulement 100 essais. En contrepartie, la flexibilité de la CI est réduite et les conclusions sont clairement plus ciblées : tandis que la première étude parvenait à identifier des indices visuels impliqués dans la tâche sans a priori sur leur localisation, le but de la seconde étude est uniquement de comparer l'influence des différentes portions du segment sur la décision, dans différentes conditions. D'autres chercheurs ont mis cette tactique du bruit dimensionnel en œuvre pour calculer des CIs à partir d'un nombre limité d'essais : pour leur expérience sur la reconnaissance de contours de formes (rondes ou carrées), Kurki et ses collaborateurs ont choisi de ne considérer que la distance au centre comme dimension variable (Kurki et al., 2014). Dans une autre étude sur la perception de particules se déplaçant à l'avant-plan et à l'arrière-plan le bruit porte sur les dimensions « direction du mouvement » et « distance à l'observateur » (Neri & Levi, 2008). De même, pour explorer la reconnaissance des mouvements complexes d'un personnage (marche et course), van Boxtel et Lu appliquent le bruit à la position des articulations (van Boxtel & Lu, 2015). Enfin, pour une étude sur la catégorisation des objets, Vondrick et son équipe utilisèrent un bruit gaussien dans l' « espace des objets » qui, une fois inversé dans l'espace des pixels, produit des images pouvant évoquer des représentations de voitures, de balles, etc… (Vondrick et al., 2014).

7.6.3. Prise en compte des dépendances

L'un des points communs des techniques évoquées jusqu'ici est le fait qu'elles ne considèrent les éléments du stimulus qu'indépendamment les uns des autres : le poids associé à un prédicteur est indépendant de celui associé aux prédicteurs adjacents. Autrement dit, ces techniques ne prennent pas en compte les dépendances existant entre les pixels adjacents de $\underline{S_i}$ ou de $\underline{N_j}$. Pourtant de telles relations existent, que ce soit par la nature même du stimulus (par exemple pour un son de parole, selon la représentation choisie, un événement acoustique couvre le plus souvent plusieurs bandes de fréquence ou plusieurs segments temporels) ou par les mécanismes perceptifs (par exemple, des canaux fréquentiels proches stimulent la même bande critique de la cochlée). Il est donc improbable que des poids adjacents de la CI aient des valeurs très différentes. La prise en compte de ces dépendances entre paramètres permettrait de réduire le nombre de degrés de liberté du modèle, et donc le phénomène de surapprentissage évoqué ci-dessus.

En pratique, cette contrainte a été implémentée, le plus souvent, grâce à un simple filtrage passe-bas de l'image par une matrice de convolution plus ou moins large (Ahumada, 1996; Barth et al., 1999; Gold et al., 2000; Neri & Levi, 2008; Sekuler et al., 2004; Thomas & Knoblauch, 2005). Cette approche reconnaît implicitement l'existence

d'interdépendances entre les valeurs des pixels adjacents de l'image, et les introduit explicitement dans le modèle par le choix de la valeur de la fréquence de coupure du filtre passe-bas (Knoblauch & Maloney, 2008). Elle permet de réduire considérablement le bruit haute-fréquence et ainsi d'amplifier relativement l'information située dans les basses fréquences.

Néanmoins, l'importance et le type de filtrage sont généralement déterminés de manière arbitraire par l'expérimentateur. Cette pratique n'est donc pas compatible avec le projet d'établir une méthode objective pour l'étude de la stratégie de catégorisation employée par l'observateur.[16]

7.7. GLM pénalisé

La description de la CI dans le cadre d'un GLM a permis de rationaliser cette phase de lissage en l'intégrant au processus d'optimisation du modèle. Cette généralisation – appelée selon les auteurs Modèle Additif Généralisé (*Generalized Additive Model*), ou GLM pénalisé (*penalized GLM)* – a l'avantage de ne pas fixer arbitrairement le degré de filtrage à appliquer, en l'ajustant plutôt sur la base des données objectives. La technique est issue du champ d'étude de l'apprentissage automatique, et a été appliquée avec succès au domaine de la restauration d'image (Nuyts & Fessler, 2003) et au calcul des STRFs (Sahani & Linden, 2003; Willmore & Smyth, 2003; Wu et al., 2006). Son principe réside dans la formulation explicite des connaissances a priori concernant la forme finale de la CI, puis l'optimisation du modèle pour atteindre un équilibre entre la qualité de l'ajustement aux données observées et la plausibilité de la CI résultante. Cette approche, appelée régression pénalisée, est plus générale que la régression « classique » qui considère uniquement les données observées à l'exclusion des autres connaissances que nous pouvons posséder au sujet du processus étudié.

7.7.1. Maximum A Posteriori (MAP)

Adoptant les conventions de l'inférence bayésienne, nos connaissances concernant le gabarit $\underline{\theta} = \left\{\underline{\beta}, c\right\}$ utilisé par l'utilisateur sont exprimées sous la forme d'une probabilité $P\left(\underline{\theta}\middle|\lambda\right)$ d'observer ce gabarit, indépendamment des données recueillies. L'a priori est donc entièrement défini par une distribution statistique (dépendant ici d'un hyperparamètre λ, qui n'est pas un paramètre du modèle mais de l'estimation elle-même). Alors que la méthode employée précédemment (cf. paragraphe

[16] Il existe toutefois une exception qui mérite attention : Pour fixer la fréquence de coupure, Neri et Levi choisirent d'ajuster une gaussienne bidimensionnelle sur le pic le plus important de l'image, et en déduisirent les paramètres de filtrage nécessaires pour obtenir un pic similaire dans la CI filtrée (Neri & Levi, 2008).

7.5) consistait à maximiser la vraisemblance du modèle $P\left(\underline{r}\middle|\underline{\theta},\underline{\underline{S}}\right)$, c'est-à-dire la probabilité d'obtenir les données observées étant donné ce modèle particulier, nous nous intéressons maintenant à maximiser la probabilité a posteriori $P\left(\underline{\theta}\middle|\underline{r},\underline{\underline{S}},\lambda\right)$, c'est-à-dire la probabilité du modèle étant données nos observations et nos connaissances a priori (Mineault et al., 2009). D'après la formule de Bayes,

$$P\left(\underline{\theta}\middle|\underline{r},\underline{\underline{S}},\lambda\right)=\frac{P\left(\underline{r}\middle|\underline{\theta},\underline{\underline{S}}\right)\cdot P(\underline{\theta}|\lambda)}{P\left(\underline{r}\middle|\underline{\underline{S}}\right)} \tag{7.23}$$

Comme au paragraphe 7.5.1, maximiser cette probabilité a posteriori revient à déterminer

$$\begin{aligned}\underline{\hat{\theta}} &= \underset{\underline{\theta}}{\operatorname{argmin}}\left(-log\left(P\left(\underline{\theta}\middle|\underline{r},\underline{\underline{S}},\lambda\right)\right)\right)\\ &= \underset{\underline{\theta}}{\operatorname{argmin}}\left(-log\left(P\left(\underline{r}\middle|\underline{\theta},\underline{\underline{S}}\right)\cdot P(\underline{\theta}|\lambda)\right)\right)\\ &= \underset{\underline{\theta}}{\operatorname{argmin}}\left(-log\left(P\left(\underline{r}\middle|\underline{\theta},\underline{\underline{S}}\right)\right)-log\left(P(\underline{\theta}|\lambda)\right)\right)\\ &= \underset{\underline{\theta}}{\operatorname{argmin}}\left(l(\underline{\theta})+r(\underline{\theta})\right)\end{aligned} \tag{7.24}$$

avec $l(\underline{\theta})$ le *negative log-likelihood* défini par l'équation 7.19 et $r(\underline{\theta})=-log\left(P(\underline{\theta}|\lambda)\right)$ *le negative log-prior* ou régularisateur. Cette équation qui définit le Maximum A Posteriori (MAP) est très proche de celle du maximum de vraisemblance, hormis l'ajout du terme $r(\underline{\theta})$ qui biaise l'estimation en faveur des paramètres les plus probables d'après nos connaissances a priori sur le gabarit utilisé par le participant. Lorsque ce régularisateur est constant, c'est-à-dire qu'aucune solution n'est jugée plus plausible qu'une autre, l'équation se ramène simplement à celle du maximum de vraisemblance.

7.7.2. Régularisation par lissage

Différents types de régularisations (i.e. différentes fonctions $r(\underline{\theta})$) ont été employées selon les situations (Mineault et al., 2009; Wu et al., 2006). L'une des plus couramment employée est la régularisation par lissage, correspondant à

$$r(\underline{\theta})=\lambda\cdot\underline{\theta}^T\underline{\underline{L}}\,\underline{\theta} \tag{7.25}$$

Dans cette équation, $\underline{\underline{L}}$ correspond à la matrice laplacienne, qui encode l'adjacence des pixels dans la CI. Plus exactement, si i et j sont les indices de deux coefficients de $\underline{\beta}$ correspondant à des pixels adjacents de $\underline{\underline{\beta}}$, $L_{i,j}=L_{j,i}=1$. Tous les autres coefficients de la matrice laplacienne sont nuls (Willmore & Smyth, 2003). $\underline{\theta}^T\underline{\underline{L}}\,\underline{\theta}$

est donc une forme quadratique mesurant le « lissage » de la CI : elle prend des valeurs plus faibles pour les CIs pour lesquelles les pixels adjacents ont des valeurs proches. L'estimation par MAP pénalise les $\underline{\theta}$ associés à des valeurs importantes de $r(\underline{\theta})$, d'après l'équation 7.24, ce qui revient à favoriser les CIs les plus lisses.

Pour une valeur de λ fixée arbitrairement, il est possible d'obtenir le MAP par des algorithmes en descente de gradient similaires à celui de l'estimation par maximum de vraisemblance, notamment le *penalized iteratively re-weighted least squares* (Friedman et al., 2010; Nickisch, 2012; Wood, 2006, 2011). La solution obtenue est un équilibre entre un bon ajustement aux données observées (minimisation de $l(\underline{\theta})$) et une CI satisfaisant notre a priori de lissage (minimisation de $r(\underline{\theta})$), l'importance relative de ces deux facteurs dans l'estimation étant déterminée par λ. Les valeurs élevées de cet hyperparamètre conduisent à des estimations exagérément déformées par le lissage (importante pénalité associée aux CIs « abruptes »), tandis que la solution tend vers celle du Maximum de Vraisemblance lorsque λ se rapproche de zéro (CI présentant un bruit d'estimation important dû au surapprentissage). Entre ces deux extrêmes, on peut trouver des valeurs intermédiaires de λ correspondant à un lissage plausible de la CI. L'exemple de régression pénalisée appliquée au calcul d'un STRF représenté Figure 23 illustre ce raisonnement. Des valeurs trop basses ou trop élevées de l'hyperparamètre conduisent respectivement à des estimations insuffisamment ou trop lissées, et donc à de faibles capacités de prédiction pour le modèle résultant. Le niveau intermédiaire de λ représenté sur la figure correspond à un STRF réaliste permettant une meilleure prédiction des données.

Comme précédemment avec la méthode du filtrage passe-bas, cette approche diminue le nombre de degrés de liberté de l'estimation en prenant en compte les dépendances entre coefficients adjacents, ce qui permet de limiter efficacement le phénomène de surapprentissage. Par rapport au résultat par Maximum de Vraisemblance, l'estimation par Maximum A Posteriori abandonne les coefficients isolés (i.e. possédant une valeur très différente de leur entourage) et ne contribuant pas suffisamment à expliquer les réponses de l'observateur, pour ne garder que les caractéristiques les plus basses-fréquences de l'image, étendues sur un grand nombre de pixels.

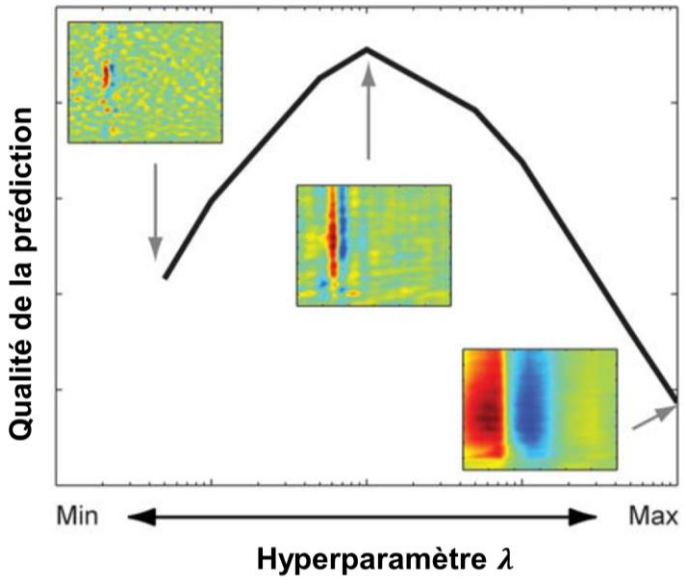


Figure 23 – Exemple de calcul d'un STRF par Régression Pénalisée. Qualité de la prédiction en fonction de la valeur de l'hyperparamètre λ. Les estimations correspondant à trois valeurs de λ sont représentées. Adapté de (Wu et al., 2006).

7.7.3. Sélection de l'hyperparamètre

Un problème essentiel subsiste néanmoins avec cette technique de régularisation : comme pour la technique du filtrage demandant de fixer arbitrairement une fréquence de coupure, l'expérimentateur doit, ici aussi, imposer une valeur de λ définissant l'importance de la pénalité de lissage relativement à la vraisemblance du modèle. Comme évoqué précédemment (voir le paragraphe 7.6.3), notre but est, au contraire, de parvenir à déterminer sans ambiguïté un degré de lissage « optimal » sur la base des données observées. Pour cela, il est nécessaire de définir un critère objectif permettant d'évaluer la qualité des CIs.

Ainsi que nous l'avons déjà souligné, pour un modèle comprenant un grand nombre de paramètres comme les CIs, l'erreur d'estimation peut provenir d'un phénomène de surapprentissage (le modèle possède suffisamment de degrés de liberté pour s'ajuster à la distribution spécifique des données sur lesquelles il a été entraîné, plutôt que de refléter le mécanisme réel sous-jacent). En pratique, ceci se traduit en pratique par une incapacité du modèle à généraliser correctement ses prédictions à de nouvelles données. Une façon directe de mettre en évidence le surapprentisage consiste donc à entraîner le modèle sur une partie des données observées (« ensemble d'entraînement »), puis à le tester sur des données différentes (« ensemble de validation »). Cette division des observations entre un ensemble d'entraînement et un ensemble de validation constitue le principe du protocole de validation croisée. Dans la pratique, on utilise le plus souvent une validation croisée à 10 plis (*10-fold cross-validation*) : les observations sont divisées en 10 sous-échantillons disjoints de taille

égale. L'un de ces sous-échantillons est utilisé comme ensemble de validation, les 9 autres constituant l'ensemble d'entraînement. La qualité de la prédiction du modèle est évaluée au moyen de l'erreur quadratique moyenne ou de la déviance. La même procédure est répétée à 9 reprises pour que tous les sous-échantillons soient utilisés une fois en tant qu'ensemble de validation. Les mesures obtenues pour ces 10 ajustements du modèle sont ensuite moyennées pour obtenir l'erreur de validation croisée ou la déviance de validation croisée.

La validation croisée à 10 plis fournit donc un critère objectif pour évaluer la généralisabilité d'un modèle. Comparée à la qualité d'ajustement (*goodness of fit*) – autre critère souvent employé pour évaluer un modèle statistique – elle permet de vérifier que le GLM réalise des prédictions correctes, non du fait d'un surapprentissage mais parce qu'il reflète fidèlement le mécanisme que nous cherchons à décrire. Il devient alors possible de sélectionner un hyperparamètre optimal selon ce critère : pour chaque valeur de λ on ajuste un GLM pénalisé, puis on évalue sa généralisabilité par validation croisée à 10 plis. Cette procédure permet alors d'obtenir la qualité de prédiction du modèle considéré. Parmi toutes les valeurs possibles de λ, on sélectionne celle qui maximise ce critère (Mineault et al., 2009; Pillow, 2007; Wood, 2006; Wu et al., 2006).

Typiquement, la qualité de prédiction est faible pour les petites valeurs de λ, car le surapprentissage entraîne dans ce cas une mauvaise généralisabilité du modèle. Pour des valeurs de λ trop importantes la pénalité de lissage devient prépondérante et la qualité de prédiction chute (Figure 23). Autrement dit, une faible pénalité de lissage (λ petit) revient à appliquer un filtrage passe bas à l'Image de Classification avec une fréquence de coupure trop élevée pour supprimer le bruit haute-fréquence dû au surapprentissage. Lorsque λ augmente, ce bruit est progressivement estompé par le filtrage. À partir d'une certaine valeur optimale de l'hyperparamètre, le filtrage devient trop important et déforme la CI qui n'a alors plus de rapport avec le gabarit réel utilisé par l'observateur. En conséquence, le modèle n'est pas en capacité de prédire correctement ses réponses, conduisant à nouveau à une qualité de prédiction réduite. Cette approche est souvent envisagée comme une forme de rasoir d'Occam puisqu'elle permet de sélectionner un modèle suffisamment complexe pour prédire les données, mais néanmoins assez simple pour ne pas subir de surapprentissage (Mineault et al., 2009).

Les différentes étapes de la procédure de calcul sont récapitulées Figure 24.

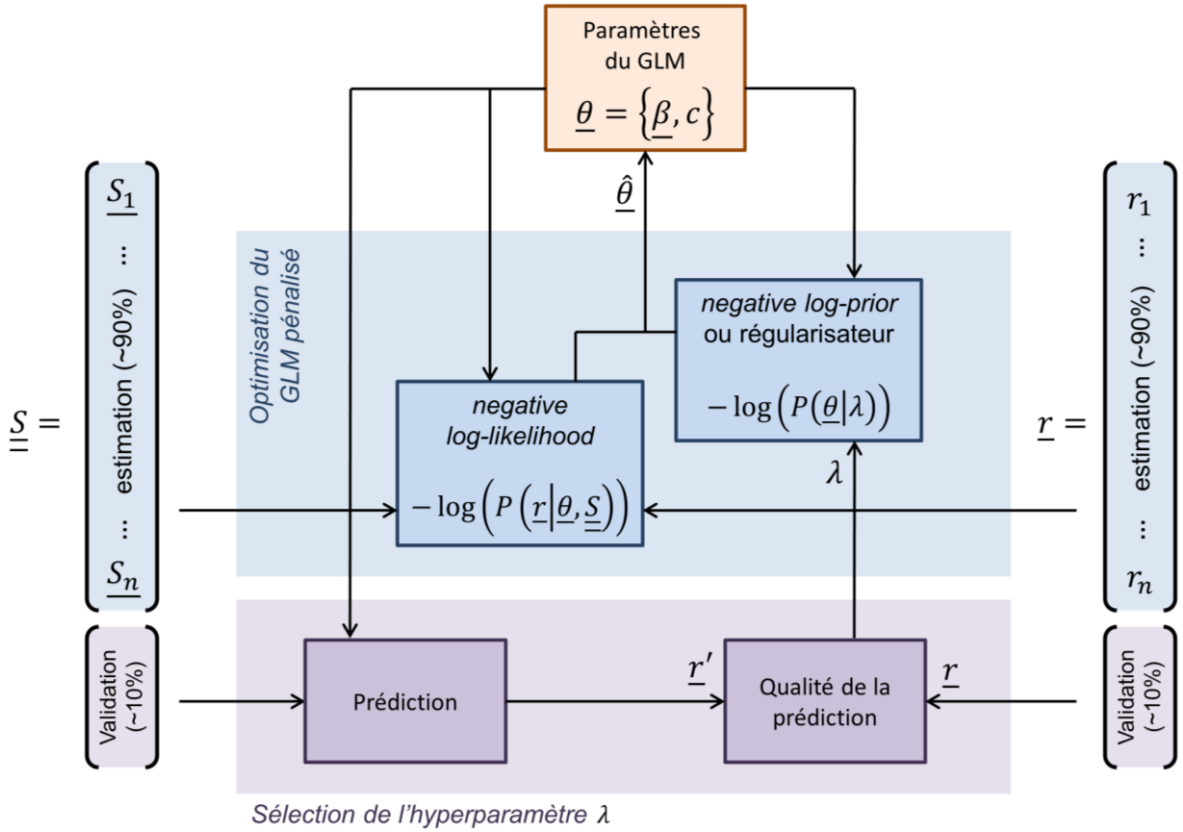


Figure 24 – Représentation schématique de la procédure de l'estimation de CI visuelles par GLM pénalisé.

En résumé, le GLM pénalisé fournit donc un cadre théorique parfaitement adapté pour les images de classification. En tant que GLM, il s'inscrit naturellement dans la cadre de l'observateur linéaire décrit par la SDT et n'impose pas d'hypothèse sur le type de bruit utilisé. Grâce à l'adjonction d'un régularisateur, il peut, en outre, comporter un grand nombre de prédicteurs sans rencontrer de problèmes de surapprentissage. Enfin, l'utilisation d'un GLM pénalisé offre dans certains cas un gain en termes de nombre d'essais et/ou une meilleure robustesse en présence d'un bruit interne important (Figure 22). Ces différentes raisons plaident en faveur de l'usage du GLM pénalisé dans le cadre d'une application à des stimuli auditifs.

7.8. Tests statistiques sur les CIs

Quel que soit le cadre statistique employé pour le calcul des CIs, ces dernières n'offrent qu'une approximation du gabarit « réel » utilisé par l'observateur. Elles comportent en effet une part plus ou moins importante de bruit d'estimation qui complique l'identification des indices utilisés. Nous avons envisagé au chapitre 7.6 différentes méthodes permettant de réduire l'importance de ce bruit. Dans le contexte

d'une étude expérimentale, il est néanmoins impératif d'effectuer des tests d'hypothèse sur les images obtenues (Chauvin et al., 2005; Murray, 2011), notamment pour déterminer objectivement dans chaque CI quelles zones reflètent la stratégie de l'observateur et quelles zones sont composées uniquement de bruit (tests statistiques individuels), ou bien pour rechercher des similarités significatives entre les CIs d'un même groupe de participants (tests statistiques multi-sujet). Dans les deux cas, on se heurte au problème des comparaisons multiples (*multiple comparisons problem*) (Bennett et al., 2010; Brett et al., 2003; Maris & Oostenveld, 2007).

7.8.1. Tests statistiques individuels

L'une des premières tentatives pour localiser les poids significatifs d'une CI consista à tester l'un après l'autre chacun des paramètres de $\underline{\hat{\beta}}$ afin de décider s'il est possible, ou non, de rejeter l'hypothèse nulle, selon laquelle la valeur de ce paramètre serait seulement attribuable au hasard. Ce procédé permet d'obtenir une distinction objective entre les pixels significatifs de l'image (correspondant à des régions contenant effectivement des indices acoustiques utilisés par le participant) et les pixels non significatifs (pour lesquels il est impossible de conclure). Ainsi, certains chercheurs employant la méthode de la corrélation inversée décidèrent de ne représenter que les pixels de l'image qui avaient une valeur de corrélation significativement différente de zéro (Barth et al., 1999; Gold et al., 2000; Sekuler et al., 2004).

Cependant, cette procédure rencontre inévitablement un écueil théorique. En effet, le nombre élevé de paramètres dans la CI implique un nombre tout aussi élevé de tests statistiques. Or, il est bien connu que la probabilité globale d'obtention d'un faux positif (ou erreur de première espèce) est démultipliée lorsqu'un même test est utilisé de manière répétée. Plus précisément, si l'on considère un test possédant une erreur de première espèce α et réitéré k fois de manière indépendante, la probabilité d'obtenir au moins un faux positif, ou risque global d'erreur de type 1 (*Family-Wise Error Rate*, FWER) vaut

$$FWER = 1 - (1 - \alpha)^k \qquad 7.26$$

À titre indicatif, pour l'étude de Gold et ses collaborateurs, la CI comportait $100 \times 100 = 10000$ pixels. En considérant uniquement les corrélations significatives à $p < 0.001$ il y a donc plus de 99.99% de chances d'obtenir au moins un faux positif.

La résolution de ce problème nécessite alors d'appliquer une correction pour les comparaisons multiples. La plus simple est appelée correction de Bonferroni et consiste à fixer α de manière à obtenir la valeur désirée de FWER, grâce à l'équation 7.26. Cette solution, envisageable en théorie, s'avère cependant trop conservative en pratique (Chauvin et al., 2005) : dans l'exemple précédent, pour obtenir un FWER de 0.05 il faudrait imposer $\alpha = 5 \cdot 10^{-6}$, un seuil difficile à atteindre avec des données réelles.

Gold envisage donc un autre type de correction en constatant simplement que le nombre de poids dépassant la signifiance (327 en moyenne) est largement supérieur au nombre de faux positifs attendu (10 pour 10000 tests à $p < 0.001$). Il en conclue donc qu'une part de ces pixels significatifs correspond à des effets réels. Ce type de raisonnement pourrait être considéré comme une forme de correction pour le taux de faux positif (*false discovery rate*, FDR) puisqu'il s'intéresse à la proportion d'erreurs de type 1 plutôt qu'à la probabilité d'obtenir au moins une erreur de type 1.

Les dépendances entre pixels adjacents, déjà envisagées ci-dessus, rendent encore plus impraticable l'usage de tests répétés, car elles violent l'hypothèse d'indépendance des tests statistiques. Une alternative consiste à intégrer dans le test d'hypothèse nos connaissances concernant la structure hautement corrélée de l'image. Ce type d'approche, destiné à être appliqué à des images « naturelles », a déjà été largement développé auparavant dans le domaine de la neuro-imagerie et implémenté dans les logiciels d'analyse de données EEG comme Fieldtrip (Oostenveld et al., 2011), ou IRMf comme SPM (Friston, 2003). Chauvin et ses collaborateurs (Chauvin et al., 2005) proposèrent donc naturellement de transposer une correction pour les comparaisons multiples issue de SPM, la *Random Field Theory* (Brett et al., 2003), afin de construire un test statistique opérant sur les CIs au niveau des pixels ou des ensembles de pixels adjacents (appelés *clusters*).

En outre, l'utilisation du GLM permet d'employer les tests de modèles développés dans ce cadre théorique, en particulier les comparaisons de modèles imbriqués (*nested models*) sur la base de la qualité d'ajustement, ou le calcul d'intervalles de confiance pour l'estimation (Knoblauch & Maloney, 2008; Kurki et al., 2014).

Les tests statistiques décrits jusqu'ici ne s'appliquent qu'au niveau d'une CI unique. Leurs conclusions portent donc seulement sur la stratégie employée par un individu et elles ne sont pas directement généralisables à la population. L'intérêt de ces analyses, souvent pratiquées en psychophysique visuelle, est donc davantage pratique que théorique : elles offrent un critère pour « seuiller » l'image et, ainsi, améliorer sa lisibilité.

7.8.2. Tests statistiques multi-sujets

Les sciences cognitives visent le plus souvent à tirer des conclusions à l'échelle d'un groupe trop large pour tester l'ensemble des individus qui le composent. Une seconde famille de méthodes statistiques rassemble les approches portant sur les données d'un échantillon restreint d'individus afin de tester une hypothèse généralisable aux propriétés du groupe dans son ensemble.

Dans le développement de la technique des CIs, les expérimentateurs ont très rapidement tenté de combiner les images de plusieurs participants, de manière à réduire le bruit d'estimation sans augmenter le nombre d'essais passés par chaque participant. Les données furent d'abord agrégées directement, comme si elles provenaient d'un seul et même observateur (Barth et al., 1999; Neri & Levi, 2008). Les chercheurs remarquèrent toutefois que cette procédure n'était pas satisfaisante car elle pouvait être biaisée, favorisant les participants dont le bruit interne est le plus faible. Pour s'assurer que la CI globale ne reflétait pas seulement les indices d'un petit nombre de participants, ils décidèrent donc de normaliser toutes les CIs en amplitude maximum (Kurki & Eckstein, 2014) ou en puissance (Neri & Levi, 2008) avant d'effectuer une moyenne.

Ces CIs « moyennes » obtenues sur un groupe de participants constituent la première étape vers un test d'hypothèse multi-sujets. Cependant, à notre connaissance, aucune étude visuelle basée sur les CIs ne s'est encore intéressée à tenter d'obtenir des conclusions généralisables à l'échelle du groupe. À ce jour, seule une étude dans le domaine auditif s'est penchée sur ce problème.

7.9. CIs auditives, tentatives récentes

Comme nous l'avons décrit au début de ce chapitre, la méthode des CIs visait initialement à déterminer si, chez un auditeur donné, la détection de ton dans le bruit est réalisée par un Détecteur Énergétique, un Détecteur Spectral, ou un Détecteur d'Enveloppe (Ahumada & Lovell, 1971; Ahumada et al., 1975; Gilkey & Robinson, 1986). La méthode fut ensuite rapidement adaptée à la modalité visuelle et les applications à des tâches auditives momentanément abandonnées, faute d'une résolution spectro-temporelle suffisante. Le développement des CIs visuelles donna naissance à de nouvelles techniques de calcul, plus puissantes, parmi lesquelles la corrélation inversée. Ces développements pouvaient dès lors être réinvestis dans le domaine auditif. Près de 40 ans après les premières études d'Ahumada, plusieurs groupes de recherche tentèrent à nouveau d'utiliser les CIs pour répondre à la question de la détection d'un ton dans le bruit.

En 2007, une première tentative, basée sur la méthode de corrélation inversée, appliquée au spectrogramme du bruit, a été proposée par Ardoint et ses collaborateurs (Ardoint et al., 2007). Elle était basée sur un paradigme très similaire à celui d'Ahumada, mais la cible à détecter était ici une sinusoïde porteuse (à 1 kHz) modulée en amplitude à la fréquence de 4 Hz. Chacun des 10 participants exécuta 10000 essais, divisés en 100 sessions. En effectuant la moyenne des spectrogrammes des bruits conduisant à une réponse positive (« la cible est présente ») et en lui soustrayant la moyenne des spectrogrammes des bruits conduisant à une réponse négative (« la cible est absente »), les chercheurs établirent les CIs auditives des participants sur cette tâche de détection.

Sur les 10 CIs, 4 montraient une composante modulée à 4 Hz dans la bande correspondant à la porteuse. L'absence de cette composante chez les 6 autres participants fut interprétée par les auteurs comme un signe de l'incertitude du système quant à la phase des signaux auditifs: un stimulus présentant une composante modulée similaire à la cible mais légèrement décalée en phase aura de fortes probabilités d'être détecté en tant que cible. Or la moyenne de toutes ces sinusoïdes déphasées est globalement nulle.

Une étude comparable fut menée en 2012, pour une détection de ton simple, comme dans les expériences d'Ahumada, mais en appliquant un paradigme de choix forcé : à chaque essai, 2 stimuli étaient présentés, l'un contenant la cible dans le bruit et l'autre le bruit seul ou avec un ton non-cible (Joosten & Neri, 2012). Comme précédemment, la technique de corrélation inversée a été employée sur les spectrogrammes des sons, donnant une image possédant une résolution faible mais néanmoins supérieure aux précédentes. Elle était divisée en 7 segments temporels de 40 ms chacun et 30 échantillons fréquentiels répartis selon une échelle logarithmique entre 1247 Hz et 8818 Hz. Chaque participant réalisa environ 7400 choix forcés. La CI résultante, bien que très bruitée, laissait apparaitre un pic positif à la position temps-fréquence de la cible, entouré de pondérations négatives pour les fréquences et les instants adjacents. Ce résultat, cohérent avec celui de l'expérience d'Ahumada (Ahumada et al., 1975), confirmait donc que le système est capable de se focaliser sur une région très restreinte de l'espace temps-fréquence et procède par comparaison avec l'énergie des régions adjacentes du spectrogramme.

En 2009, Shub et Richards utilisèrent un paradigme de détection à la volée (*free-running*), inspiré des STRFs, pour étudier la détection de ton à une fréquence cible (1000 Hz ou 2431 Hz) dans un bruit diffusé en continu et composé, lui aussi, de tons répartis aléatoirement en temps et en fréquences (Shub & Richards, 2009). Le signal était présenté en 10 blocs de 5 minutes chacun, séparés par des pauses, et les trois participants avaient pour instruction de presser une touche le plus rapidement possible lorsqu'ils entendaient le ton cible. Le calcul de la moyenne des spectrogrammes précédant chaque réponse faisait là encore ressortir la fréquence cible, avec des poids négatifs associés aux fréquences adjacentes. En revanche la variabilité des délais de réponse entraînait une perte importante de résolution temporelle, ce qui conduisit à un abandon de la technique.

Ces trois tentatives récentes d'applications des CIs à des tâches de catégorisation ou de détection auditives se basent sur une transposition directe de la technique de corrélation inversée au spectrogramme du bruit, ou à des versions dérivées du spectrogramme. Leur résolution en temps et en fréquence et leur précision sont intrinsèquement limitées par la quantité de données nécessaires pour ajuster un modèle linéaire contenant un nombre très élevé de paramètres. Il est donc tentant de leur appliquer les développements théoriques plus récents dans le domaine des CIs,

comme le GLM. Schönfelder et Wichmann démontrèrent par simulation l'efficacité de la régression pénalisée dans le cadre de la détection de ton dans le bruit (Schönfelder & Wichmann, 2012). Ils employèrent pour cela une pénalité de parcimonie, légèrement différente de la pénalité de lissage présentée précédemment. Néanmoins le principe reste inchangé : cette régression pénalisée favorise les modèles simples (ici les modèles contenant peu de paramètres non nuls) par rapport aux modèles complexes (contenant un grand nombre de paramètres non nuls), prévenant ainsi tout risque de surapprentissage. Après avoir simulé le comportement de différents observateurs théoriques (Détecteur Energétique, Détecteur Spectral, et Détecteur d'Enveloppe) pour la tâche de détection de ton dans le bruit, les auteurs montrèrent de manière probante que la régression pénalisée permettait de déterminer clairement sur quel caractéristique du son (énergie, spectre ou enveloppe) s'appuyait chaque participant simulé, d'après ses réponses, et ainsi d'identifier le type de stratégie employée. Cette approche est par ailleurs robuste pour une large gamme de SNRs et de bruits internes, contrairement à la corrélation inversée et au GLM non pénalisé.

Bien que ces derniers résultats soient particulièrement encourageants, l'application du GLM pénalisé pour l'identification de la stratégie employée par un participant « réel » lors d'une tâche auditive n'a, à notre connaissance, encore jamais fait l'objet d'une publication.

8. Synthèse de la partie théorique

Comme nous l'avons vu dans les trois premiers chapitres de cette thèse, la compréhension de la parole s'appuie sur l'extraction de primitives acoustiques. Parmi toutes les informations – parfois redondantes – contenues dans le signal de parole, certaines sont effectivement utilisées par l'auditeur pour différencier les phonèmes, tandis que d'autres ne sont pas pertinentes pour la compréhension. Ainsi le signal de parole, multidimensionnel et extrêmement variable, peut-il être réduit à un nombre restreint de caractéristiques acoustiques encodant de manière stable les unités linguistiques. Ce système d'extraction des indices acoustiques présente, en outre, une certaine variabilité selon les situations d'écoute et les individus (chapitre 5). Les données neurophysiologiques présentées au chapitre 6 viennent étayer et compléter ces observations comportementales concernant l'interface acoustico-phonétique.

La question de l'identité de ces primitives de parole est récurrente dans la littérature, et plusieurs approches ont été mises en œuvre pour tenter d'y répondre (chapitre 4) : le paradigme du continuum de parole synthétique, la méthode du signal progressivement dégradé, l'analyse des matrices de confusion (directe ou par le truchement des profils de confusion), la 3DDS, les fonctions de pondération, et la méthode des bulles auditives. Cependant, ces approches restent limitées, non seulement du fait qu'elles imposent une tâche (par exemple, la catégorisation de phonèmes en choix forcé) et un type de stimuli particuliers (par exemple, des stimuli de parole synthétique) mais, surtout, parce qu'elles n'aboutissent pas à une description détaillée des indices acoustiques mis en jeu. En effet, la première se ramène plutôt à une vérification a posteriori du rôle d'une caractéristique acoustique supposée critique pour la catégorisation de phonèmes. Les autres possèdent pour la plupart des résolutions spectro-temporelles relativement contraignantes et ne révèlent que les indices acoustiques principaux, laissant de côté la possibilité d'éventuelles stratégies plus complexes.

Le principe des Images de Classification décrit au chapitre 7 laisse entrevoir une nouvelle manière d'aborder ce problème. La méthode des CIs, développée principalement pour l'étude de la modalité visuelle, offre une visualisation directe des portions du stimulus utilisées dans une tâche de catégorisation dans le bruit. Dans sa version la plus récente, elle emploie un outil statistique – appelé GLM pénalisé – pour relier la réponse de l'observateur aux propriétés physiques du stimulus présenté. Cette méthode permet de calculer une « carte » représentant le poids de chaque pixel dans la décision, tout en réduisant le bruit d'estimation lié au surapprentissage.

Si le parallèle avec notre question première est assez évident, l'application des Images de Classification à l'identification des primitives de la parole nécessite cependant une transposition de la méthode depuis le domaine visuel dans le domaine auditif. Au

cours de cette thèse, nous avons réalisé 4 études concernant le développement méthodologique et la mise en œuvre des ACI dans des tâches de catégorisation de parole dans le bruit. Notre première étude (chapitre 9) établit un cadre théorique pour le calcul d'Images de Classification Auditives. Dans la deuxième étude (chapitre 10) nous proposons un outillage statistique permettant de tirer des conclusions quant à la significativité des ACIs au niveau d'un groupe. Ces recherches nous ont amené à appliquer cette méthode dans les deux études suivantes afin d'examiner deux questions actuellement discutées dans le milieu scientifique : d'une part, l'origine des meilleures performances des musiciens experts lors de tâches de compréhension de la parole dans le bruit, comparativement aux non-musiciens (chapitre 11) et, d'autre part, le déficit des représentations phonologiques chez les personnes dyslexiques et les stratégies compensatoires qu'ils mettent en place (chapitre 12).

Partie expérimentale

9. Étude 1 : Mise en place de la méthode des Images de Classification Auditives pour l'identification des indices acoustiques utilisés dans une tâche de catégorisation aba/ada.

9.1. Présentation de l'étude 1

Dans le domaine de la compréhension de la parole, les primitives acoustiques servant à la catégorisation des sons en unités perceptives (phonèmes ou syllabes) restent encore très mal connues aujourd'hui. L'étude de ces indices acoustiques nécessite de pouvoir « visualiser » quels éléments du signal de parole sont utilisés par les auditeurs. L'absence d'une méthode satisfaisante qui répondrait spécifiquement à cette tâche représente un frein majeur aux avancées dans le domaine. Dans le domaine de la perception visuelle, un questionnement assez proche a conduit au développement des Images de Classification (CI) permettant de révéler les indices visuels employés par un observateur pour réaliser une catégorisation d'images dans le bruit.

Le but de cette première étude est donc d'adapter puis de rendre exploitable et opérationnelle la technique des CIs visuelles pour la modalité auditive, afin de l'appliquer ensuite à l'exploration de la perception de la parole. Ce projet demandait, dans un premier temps, de déterminer les obstacles pratiques entravant ce transfert, puis de mettre en œuvre un modèle statistique – semblable à celui des Images de Classification mais correspondant à l'interface acoustico-phonétique – et, enfin, d'illustrer son application à travers un exemple.

Nous avons choisi ici de démontrer le bon fonctionnement de la méthode en identifiant les indices acoustiques impliqués dans la catégorisation /b/-/d/ dans le bruit. À cette fin, nous avons collecté les réponses de trois participants dans une tâche de catégorisation des logatomes /aba/ et /ada/ présentés dans un bruit blanc. Chaque auditeur réalisa ainsi 10000 essais, avec un niveau de bruit évoluant au cours de l'expérience en fonction de ses performances. Cette démarche expérimentale visait à isoler ainsi les régions de l'espace temps-fréquence essentielles pour la catégorisation et ainsi déterminer les indices acoustiques utilisés par l'auditeur.

Cette étude a été publiée en tant qu'article de méthode dans la revue en accès ouvert Frontiers in Neuroscience (Varnet et al., 2013b). Une version préliminaire de ces résultats avait été présentée dans un congrès international (Interspeech 2013, Lyon), donnant lieu à la publication d'un article court dans les actes de la conférence (Varnet et al., 2013a). Cette première démonstration de la méthode a également fait l'objet d'une communication orale lors de la journée de l'École Doctorale NSCo 2015.

9.2. Article 1 : Using auditory classification images for the identification of fine acoustic cues used in speech perception.

frontiers in
HUMAN NEUROSCIENCE

METHODS ARTICLE
published: 16 December 2013
doi: 10.3389/fnhum.2013.00865

Using auditory classification images for the identification of fine acoustic cues used in speech perception

Léo Varnet[1,2]*, Kenneth Knoblauch[3], Fanny Meunier[1,2,4] and Michel Hoen[1,2]

[1] *Neuroscience Research Centre, Brain Dynamics and Cognition Team, INSERM U1028, CNRS UMR5292, Lyon, France*
[2] *Ecole Doctorale Neurosciences et Cognition, Université de Lyon, Université Lyon 1, Lyon, France*
[3] *Integrative Neuroscience Department, Stem Cell and Brain Research Institute, INSERM U846, Bron, France*
[4] *Laboratoire sur le Langage le Cerveau et la Cognition, CNRS UMR5304, Lyon, France*

Edited by:
Srikantan S. Nagarajan, University of California, San Francisco, USA

Reviewed by:
Shanqing Cai, Boston University, USA
Nima Mesgarani, Columbia University, USA

****Correspondence:***
Léo Varnet, Institute for Cognitive Science, 67 boulevard Pinel, 69675 Bron, France
e-mail: leo.varnet@isc.cnrs.fr

An essential step in understanding the processes underlying the general mechanism of perceptual categorization is to identify which portions of a physical stimulation modulate the behavior of our perceptual system. More specifically, in the context of speech comprehension, it is still a major open challenge to understand which information is used to categorize a speech stimulus as one phoneme or another, the auditory primitives relevant for the categorical perception of speech being still unknown. Here we propose to adapt a method relying on a Generalized Linear Model with smoothness priors, already used in the visual domain for the estimation of so-called classification images, to auditory experiments. This statistical model offers a rigorous framework for dealing with non-Gaussian noise, as it is often the case in the auditory modality, and limits the amount of noise in the estimated template by enforcing smoother solutions. By applying this technique to a specific two-alternative forced choice experiment between stimuli "aba" and "ada" in noise with an adaptive SNR, we confirm that the second formantic transition is key for classifying phonemes into /b/ or /d/ in noise, and that its estimation by the auditory system is a relative measurement across spectral bands and in relation to the perceived height of the second formant in the preceding syllable. Through this example, we show how the GLM with smoothness priors approach can be applied to the identification of fine functional acoustic cues in speech perception. Finally we discuss some assumptions of the model in the specific case of speech perception.

Keywords: classification images, GLM, phoneme recognition, speech perception, acoustic cues, phonetics

INTRODUCTION

A major challenge in psychophysics is to establish what exact parts of a complex physical stimulation modulate its percept by an observer and constrain his/her behavior toward that stimulus. In the specific field of speech perception, identifying the information in the acoustic signal used by our neurocognitive system is crucial in order to understand the human language faculty and how it ultimately developed in human primates (Kiggins et al., 2012). In this context, questions of speech segmentation, i.e., which acoustical cues are used to isolate word units in the continuous acoustic speech stream; or phonemic categorization, i.e., which among the auditory primitives that are encoded at the neural acoustic/phonetic interface are actually used by our perceptual system to recognize and categorize phonemes, still constitute an important open debate (see Cutler, 2012 for a review). As a consequence, today there is no universal model of speech recognition that can work directly on the acoustic stream. Models of speech recognition, even the most efficient and well developed ones, usually avoid the acoustic/phonetic step (e.g., Luce and Pisoni, 1998; Norris and McQueen, 2008) or rely on systems that are not based on realistic human behaviors (Scharenborg et al., 2005).

In this paper we propose a method and procedure allowing direct estimation of which parts of the signal are effectively used by our neurocognitive system while processing natural speech. Of course one way that was used in previous work to identify relevant acoustic cues in speech is to proceed by progressive signal reductions, i.e., eliminating certain cues from the speech signal in order to demonstrate which ones are mandatory. In the 1950's, phoneme recognition was extensively studied by Liberman and colleagues for example, using the systematic variation of a limited number of features in the time-frequency domain (usually one or two) along a continuum of synthetic speech (Liberman et al., 1952, 1954, 1957). More recent work conducted on this topic has involved artificially degraded speech, such as noise-vocoded (Xu et al., 2005), sine-wave (Loizou et al., 1999), or band-pass filtered speech (Apoux and Healy, 2009). These approaches can, however, only offer a very limited account of the problem, as it is known that the speech comprehension system shows very fast and efficient functional plasticity. Once shaped by linguistic experience, our speech perception system can rapidly modify the cues that are relevant for phonemic categorization in response to drastic signal reductions or even stronger manipulations (see for example: Shannon et al., 1995). This resistance of speech perception to drastic signal impoverishment was attributed to the redundancy of information in speech: no single acoustic feature in speech is absolutely crucial for its comprehension (Saffran and Estes, 2006).

The signal reduction approach can therefore not account for the many possible acoustic dimensions used by listeners in a single categorization task, or for their evolution with listening situations. While signal reduction paradigms are appropriate to study the functional plasticity triggered in our speech perception system by signal reductions, they can hardly inform us on the way the system reacts in more natural perception situations.

An alternate way to proceed would be to develop a method allowing experimenters to directly "see" where humans listen inside natural speech signals, without having to modify them. In the following, we show how a methodological solution to this issue can be provided by new developments in the domain of so-called *classification images (CIm)*. We demonstrate how this method can now be adapted to auditory experiments and how this method can further be developed to study the identification of functional fine acoustic cues in speech perception. We will also discuss how this method could be adapted to other domains of studies both in perceptual and cognitive neuroscience.

Since Ahumada and Lovell (1971) first developed a correlational technique to estimate the frequency weighting-function of observers detecting a 500-Hz tone-in-noise, much has been done for establishing a robust theoretical framework in which to describe and analyze the set of techniques gathered under the name of CIm (see Murray (2011) for an in-depth review). The basic idea underlying the classification image approach is that, faced with any kind of perceptual decision, our neurocognitive system will sometimes generate correct perceptions and sometimes errors, which could be informative on the computational mechanisms occurring in perceptual systems. If one could have access to the physical conditions of the stimulation that favor either perceptual failure or success, then one can derive the relevant parts of any stimulation that impact the perceptual decision process. As a consequence, the tasks used to generate CIm are categorization tasks. The typical paradigm used in classification image experiments is an identification or detection experiment, in which each trial consists in the presentation of one of two possible signals and the participant is instructed to classify the stimuli between the two options (t_0 or t_1). In order to derive a classification image, stimuli are systematically masked by a certain amount of random background-noise. For each trial, the response given by the participant, the signal actually presented and the trial-specific configuration of the noise field are recorded. The classification image aims at showing the precise influence of the noise field on the observer's response, for a given signal.

The best known (and maybe the most intuitive) method for calculating a classification image, first used by Ahumada (1996) and termed reversed-correlation, derives from the idea of establishing the correlation map between the noise and the observer's responses. In practical terms, this is done by averaging all of the noise fields eliciting response t_0 and subtracting the average of all of the noise fields eliciting response t_1. The idea is that if one can determine how the presence of background-noise at each point inside the space of a stimulus interferes with the decision of the observer, one can derive a map of the perceptual cues relevant to achieve a specific categorization task. By showing which components influence the recognition performance, this method gives us insights into the observer's internal decision template for this specific task. Although it has been primarily conceived as an answer to a question raised in the auditory domain (Ahumada and Lovell, 1971; Ahumada et al., 1975), and although the method could have easily been further developed to study auditory processes, this powerful tool has been mostly exploited up to now in studies on visual psychophysics. This technique has been used to investigate a variety of visual tasks, including the ability to perceive two segments as aligned or not (i.e., Vernier acuity, Ahumada, 1996), the detection of Gaussian contrast modulation (Abbey and Eckstein, 2002), the processing of illusory contours (Gold et al., 2000), visual perceptual learning (Gold et al., 2004), and luminance (Thomas and Knoblauch, 2005) and chromatic (Bouet and Knoblauch, 2004) modulation.

In the auditory domain, the classification image is a promising approach for determining which "aspects" of the acoustic signal (formant position or dynamic, energy burst, etc.) are crucial cues for a broad variety of psychoacoustic tasks (i.e., tonal or pitch discrimination, intensity perception or streaming, etc.) and particularly in the context of speech comprehension. However, to our knowledge, attempts to adapt this methodology to the auditory modality have until now produced limited results. Among noteworthy attempts, Ardoint et al. (2007) have adapted the reversed correlational method to study the perception of amplitude modulations and very similar correlational procedures have been used to determine spectral weighting functions of speech stimuli (see for example: Doherty and Turner (1996); Apoux and Bacon (2004) or Calandruccio and Doherty (2008)). Two severe limitations can, at least partly, explain the limited development of the technique. Firstly, several thousands of trials are typically needed to compute a classification image accurately. The minimum number of trials theoretically required is equal to the number of free-parameters, but many more are needed to be able to estimate the classification image with an adequate signal-to-noise ratio (up to 11400 trials, in Barth et al., 1999). This problem has been overcome in the visual domain by reducing the number of random variables under consideration, for example by averaging along irrelevant dimensions (Abbey and Eckstein, 2002, 2006), or by using a "dimensional" noise (Li et al., 2006). Unfortunately, none of these methods can be used with such complex and time-varying stimulus as speech. Furthermore, mental and physiological fatigue occurs rapidly when listening to very noisy stimuli. The second factor restricting the use of reverse-correlation for estimating auditory CIm is the strong assumption about the statistical distribution of the noise imposed by the statistical theory. Since its theoretical background has mostly been developed assuming additive Gaussian-noise, methods such as reverse-correlation are not the most suitable statistical framework to deal with non-Gaussian noise-fields. In the visual domain, CIm can be based on Gaussian noise, for example in the case of luminance noise which will modify the observer's perception in a symmetric fashion, adding or subtracting luminance to pixels in a picture. The interest of using CIm for the study of speech signals, however, imposes the use of acoustic stimuli which will have complex spectro-temporal composition and the calculation of an auditory classification image should therefore not be based on the amplitude of the noise samples, but rather on the power of

the time-frequency bins of their power spectrum. These unfortunately generally have non-Gaussian distributions. These two limitations make it difficult to calculate auditory CIm using the standard reverse-correlation method.

A major advance in the comprehension and computation of CIm was achieved by Knoblauch and Maloney (2008) who proposed to fit the data with a Generalized Linear Model (GLM), which provide a more accurate and comprehensive statistical framework for calculating CIm. This initial work was followed by Mineault et al. (2009) and Murray (2011, 2012). Interestingly, this appealing approach offers a way to overcome the two pitfalls mentioned above. Firstly, GLMs naturally allow the addition of prior knowledge on the smoothness of the expected classification image, resulting in Generalized Linear Models (GLMs) (Hastie and Tibshirani, 1990; Wood, 2006). By exploiting the dependencies between adjacent noise values, one can significantly reduce the number of trials required. In fact, GLMs with priors are widely used for describing the stimulus-response properties of single neurons (Pillow, 2007; Pillow et al., 2008), in particular in the auditory system (in terms of Spectrotemporal Receptive Field, STRF, Calabrese et al., 2011). Secondly, unlike the reverse-correlation method, the GLM does not require the noise to be normally distributed. Accordingly, it can measure CIm using noise fields from non-Gaussian distributions, such as the power spectrum of an acoustic noise, in a similar way to the calculations of second-order CIm using GLM by Knoblauch and Maloney (2012). Therefore, Generalized Linear and Additive Models provide suitable and powerful tools to investigate the way in which the human system achieves fast and efficient categorization of phonemes in noise. In this paper we applied the GLM with smoothness priors technique to the identification of acoustic cues used in an identification task involving two VCV speech sequences: ABA (/aba/) and ADA (/ada/). In this particular case a strong hypothesis, formulated in Liberman et al. (1954), is that the second formant transition would be a key for classifying the stimulus into [ABA] or [ADA]. Under this assumption, the classification image would be focused on the time-frequency localization of the second formant transition.

MATERIALS AND METHODS

In the following sections we use the convention of underlined symbols to indicate vectors, double underlined symbols to indicate matrices, and non-underlined symbols to indicate scalars.

EXPERIMENTAL PROCEDURE

Three native French-speaking listeners took part into this study: the first two Léo Varnet and Michel Hoen are co-authors on the paper and were not naïve regarding details of the study, a third participant was thus added, S.B. who was completely naïve toward the task. They were 24, 25, and 35 year old, males, right handed and native French speakers, without known language or hearing deficits.

Targets sounds, hereafter denoted $\underline{t}_0$ and $\underline{t}_1$,were two natural-speech signals (ABA /aba/ and ADA /ada/ respectively) obtained by concatenating the same utterance of /a/ with an utterance of /ba/ or /da/. Original sounds were recorded in a soundproof chamber by the same female speaker and digitized at a sample rate of 44.1 kHz. The sound samples were 680 ms long, and their average power was normalized. Each stimulus $\underline{s}_i$ consisted of one target-sound, embedded in a Gaussian additive-noise using Equation (1):

$$\underline{s}_i = \alpha_i \cdot \underline{t}_{k_i} + \underline{n}_i \tag{1}$$

In (1), i is the trial number, k_i the target number (0 or 1) associated with this trial, $\underline{n}_i$ the noise field drawn from a normal distribution, and α_i a factor allowing the adjustment of signal-to-noise ratio (SNR) as a function of the participant's behavior, see Adaptive stimulus-delivery procedure below. Each stimulus was normalized in intensity level using the root mean square and preceded with a Gaussian fade-in of 75 ms convolved with a Gaussian-noise, in order to avoid clicks or abrupt attacks. The sample rate was the same as for the original sounds.

The experiment consists in the presentation of a list of $N = 10{,}000$ noisy stimuli (5000 for each target) presented in a completely random order. Participants were instructed to listen carefully to the stimulus and then indicate by a button press whether the masked signal was $\underline{t}_0$ or $\underline{t}_1$, a response denoted by r_i (= 0 or 1). The following trial began after 200 ms. Listeners could complete the task over a period of 1 week, at their own pace, depending on individual fatigue and availability, for a total duration of approximately 3h. Each participant divided the experiment in sessions of approximately 1000 stimuli, on their own initiative. The experiment was run in a quiet experimental room and stimuli were delivered using Sennheiser's HD 448 headphones.

ADAPTIVE STIMULUS-DELIVERY PROCEDURE

During the course of the experiment, the signal level was constantly adjusted to ensure a correct response rate around 75%, as in several previous classification image experiments (e.g., Gold et al., 2000). Signal contrast must be strong enough to ensure that the SNR will not severely affect the decision rule, but sufficiently low so that noise influences the decision of the observer. That is to say, that noise must be misleading on a sufficient number of stimuli, without leading the observer to reply randomly on the task. For this purpose, the SNR was varied from trial-to-trial on the basis of a local rate of correct responses calculated on a 10-trial window, with an adaptation of 0.2, 0.4, 0.6 or 0.8 dB for differences of 5, 10, 15, or 20% between intended and actual scores (variations of the SNR were limited to the range -20 dB to -0 dB; we systematically record the final SNR value for one session and use it as starting value for the next session before the adaptive algorithm takes over in adjusting the SNR). However, in the following we assume that the SNR does not affect the observer's strategy for categorization, a point that will be discussed in Discussion.

DERIVING AUDITORY CLASSIFICATION IMAGES

Each stimulus noise $\underline{n}_i$ is characterized by its power spectrogram, whose components are entered as predictor variables in the model. Power spectrograms were calculated with Matlab function *spectrogram*, using a Short-Time Fourier Transform with a moving 512 points Hamming window and no overlap, resulting in 86.13 Hz frequency resolution and 11.6 ms temporal resolution. Since the last 340 ms of the signal were almost silent, we limited

our analysis to a time range of 0–0.34 s and a frequency range of 0–4048 Hz, thus ensuring that the size of the data-set would not exceed computational limits. The resulting 46-by-30 matrix (frequency bins by time bins) is reshaped into a 1380-by-1 vector of time-frequency bins, labeled $\underline{X}_i$. A similar treatment is applied to both targets, resulting in vectorized power spectrograms $\underline{T}_0$ and $\underline{T}_1$ (**Figure 1**).

More biologically-inspired time-frequency representations of the noise, as a cochleagram, can replace the spectrogram for deriving the CIm, in order to obtain a "higher-level" representation of the functional acoustic cues. Or more simply, we could apply a logarithmic scale to the frequency axis to account for the non-linear spacing of filter center frequencies, as it is done in the STRF calculation. Nevertheless, in our case the aim was only to replicate a known property of the speech comprehension system, which is more intuitive on a simple time-frequency representation.

In general agreement with the literature on classification image (for example Ahumada, 2002) we assume that the observer performs the detection of acoustic cues linearly by template matching, a longstanding model for decision making. First, an internal decision variable d_i is computed by taking the scalar product of the input with a weighting function $\underline{w}$ referred to as the observer's template, and adding a random variable ε_i representing the internal noise of the system (accounting for the fact that the observer does not necessarily give the same response when presented with the same stimulus twice). In (2), the errors ε_i are assumed to have a zero mean symmetric distribution and to be independent from trial to trial.

$$d_i = \left(\underline{X}_i + \underline{T}_{k_i}\right)^T \cdot \underline{w} + \varepsilon_i \tag{2}$$

Then the response variable is given by (3):

$$r_i = \begin{cases} 1 \text{ if } \mathrm{d}_i > c \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

c is a fixed criterion that determines the bias of the observer toward one alternative. Knoblauch and Maloney (2008) reformulated this very simple model in terms of a GLM, by expressing the probability that the observer gave the response $r_i = 1$, given the data $\underline{X}_i$, in the case of presentation of the target number k_i:

$$P(r_i = 1) = \Phi(\underline{X}_i{}^T \cdot \underline{\beta} + \mathrm{s}_{k_i}) \tag{4}$$

with Φ cumulative distribution function associated with ε, $\underline{\beta}$ the classification Image, and $\underline{s}$ a two-level factor reflecting the influence of the target actually presented on the response. In line with the psychophysics literature, we could assume that ε is taken from a logistic distribution (a common choice for modeling binomial data), and therefore the associated psychometric function Φ will be the inverse of the logit function. It would still be possible to use other assumptions, as a Gaussian distribution for ε and as a consequence, the inverse of the probit function as Φ. Such changes might have an impact on the model though it would be presumably small.

The structure of Equation (4), with a linear combination of parameters linked to the dependent variable via a psychometric function, is the typical form of a GLM (Fox, 2008; Knoblauch and Maloney, 2012). At this stage we could thus determine the values of the model parameters $\underline{\theta} = \left\{\underline{\beta}, \underline{s}\right\}$ that best fit the empirical data, by simply maximizing the log-likelihood:

$$\begin{aligned} L\left(\underline{\theta}\right) &= \log\left(P\left(\underline{r}\,\middle|\,\underline{\theta}, \underline{k}, \underline{\underline{X}}\right)\right) \\ &= \log\left(\prod_i P\left(r_i\,|\underline{\theta}, k_i, \underline{X}_i\right)\right) \end{aligned} \tag{5}$$

that is a natural measure of match between data and fit, assuming statistical independence between responses r_i. Thus, calculating:

$$\underline{\hat{\theta}}_{ML} = \underset{\underline{\theta}}{\operatorname{argmax}}\, L\left(\underline{\theta}\right) \tag{6}$$

by a standard maximization algorithm (e.g., the built-in Matlab function *glmfit*) would provide us maximum likelihood estimates of the CIm, $\underline{\hat{\beta}}$ and of the stimulus factor $\underline{s}$.

Unfortunately, these estimates would be presumably too noisy to be decipherable. Indeed GLMs, as well as reverse-correlation, when comprising a large number of predictors (1382 in this example), are prone to overfitting, which means that the model will describe the trial-dependent noise as well as the underlying classification mechanism. Estimates of the observer's template by GLM can therefore be quite noisy, and the model will not be able to generalize to novel data. For proper predictions of previously unseen data, templates should not be determined by the specific distribution of noise in trials used to fit the model, but rather reflect the decision process of the observer.

One solution has been developed in the GLM framework under the name "Penalized Likelihood," which has been widely used for estimating the receptive fields of single neurons (Wu et al., 2006; Calabrese et al., 2011; Park and Pillow, 2011) and adapted to CIm by Knoblauch and Maloney (2008) and later by Mineault et al. (2009). Another example of using a similar method for an application in the auditory domain can be found in Schönfelder and Wichmann (2012), who modeled results from a classical auditory tone-in-noise

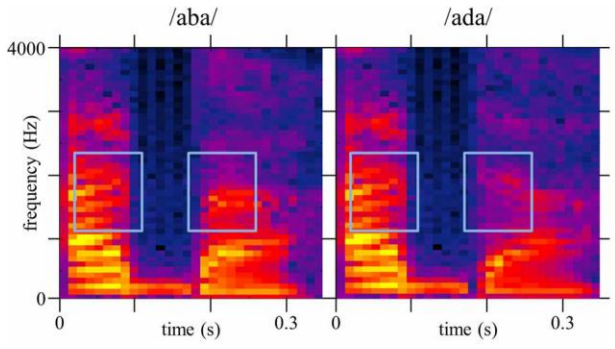


FIGURE 1 | Spectrograms of target-signals t_0 (/aba/) and t_1 (/ada/) used for the vectorized spectrograms *T0* and *T1*, on a logarithmic scale (dB). Blue boxes indicate the second formantic transition (F2).

detection task using this approach. Among the various R or Matlab toolboxes available, we decided to use Mineault's function glmfitqp (http://www.mathworks.com/matlabcentral/fileexchange/31661-fit-glm-with-quadratic-penalty) that allows optimizing a GLM with quadratic penalty. The aim of this method is to incorporate prior knowledge about the smoothness of the intended classification image (which is equivalent to introduce dependencies between adjacent predictors). To do so, we associate with each value of the model parameters $\underline{\theta}$ a probability $P(\underline{\theta}|\underline{\lambda})$ representing our a priori beliefs about the true underlying template (in our case, a smoother classification image will be more expected, and therefore have a high prior probability). This prior is defined by a distribution and a set of hyperparameters $\underline{\lambda}$, as explained later. Then, instead of maximizing the likelihood as before, we maximize the log of the posterior $P\left(\underline{\theta}\left|\underline{r}, \underline{k}, \underline{\underline{X}}, \underline{\lambda}\right.\right)$ that takes into account the likelihood and prior information, as evidenced with Bayes' rule:

$$P\left(\underline{\theta}\left|\underline{r}, \underline{k}, \underline{\underline{X}}, \underline{\lambda}\right.\right) = \frac{P\left(\underline{r}\left|\underline{\theta}, \underline{k}, \underline{\underline{X}}\right.\right) \cdot P\left(\underline{\theta}\left|\underline{\lambda}\right.\right)}{P(\underline{r}|\underline{k}, \underline{\underline{X}})} \tag{7}$$

Therefore the Maximum A Posteriori (MAP) estimate of the model parameters is given by:

$$\begin{aligned}\widehat{\underline{\theta}}_{\mathrm{MAP}} &= \mathrm{argmax}_{\underline{\theta}} \log\left(P\left(\underline{\theta}\left|\underline{r}, \underline{k}, \underline{\underline{X}}, \underline{\lambda}\right.\right)\right) \\ &= \underset{\underline{\theta}}{\mathrm{argmax}} \log\left(P\left(\underline{r}\left|\underline{\theta}, \underline{k}, \underline{\underline{X}}\right.\right) \cdot P\left(\underline{\theta}\left|\underline{\lambda}\right.\right)\right) \\ &= \underset{\underline{\theta}}{\mathrm{argmax}} \left[\log\left(P\left(\underline{r}\left|\underline{\theta}, \underline{k}, \underline{\underline{X}}\right.\right)\right) + \log\left(P\left(\underline{\theta}\left|\underline{\lambda}\right.\right)\right)\right] \\ &= \underset{\underline{\theta}}{\mathrm{argmax}} \left[L\left(\underline{\theta}\right) + R\left(\underline{\theta}\right)\right]\end{aligned} \tag{8}$$

The last equation can be seen as the same maximization of the log-likelihood as before, with an additional penalty term, $R\left(\underline{\theta}\right)$, that biases our estimate toward model parameters with higher a-priori probability. The optimal estimate is a tradeoff between fitting the data well and satisfying the constraints of the penalty term. Therefore, a prior on smoothness will favor CIm with slow variations in time and frequency (but note that other types of priors exist (Wu et al., 2006), e.g., implying assumptions on independence (Machens et al., 2004), sparseness (Mineault et al., 2009; Schönfelder and Wichmann, 2012), or locality (Park and Pillow, 2011) of the model parameters).

In agreement with the Matlab function we use, we chose our smoothness prior to be a sum of two quadratic forms:

$$P\left(\underline{\theta}|\lambda_1, \lambda_2\right) = \lambda_1 \underline{\theta}^T \underline{\underline{L}}_1 \underline{\theta} + \lambda_2 \underline{\theta}^T \underline{\underline{L}}_2 \underline{\theta} \tag{9}$$

where $\underline{\underline{L}}_1$ is the Laplacian matrix along dimension 1 (time), $\underline{\underline{L}}_2$ the Laplacian matrix along dimension 2 (frequency) in the time-frequency space (Wu et al., 2006). Thus, the quadratic form $\underline{\theta}^T \underline{\underline{L}}_D \underline{\theta}$ provides a measure of the smoothness of $\underline{\theta}$ over dimension D. As we do not know the appropriate importance of smoothness along the time and frequency axis, we introduce two hyperparameters (indeed the scale of smoothness in the spectral and temporal dimensions are presumably unrelated) $\underline{\lambda} = \{\lambda_1, \lambda_2\}$ that control the prior distribution on $\underline{\theta}$, and therefore the strength of penalization. The absolute values of the hyperparameters (also called "regularization parameters") have no clear interpretation, as they represent the relative importance of quality of fit and smoothness. For large (>1) values of λ_1 and λ_2 we put a strong disadvantage on sharp CIm, and for $\lambda_1 = \lambda_2 = 0$ we recover the initial maximum likelihood solution.

The standard method for setting the value of the hyperparameters is cross-validation (e.g., in Wu et al., 2006; Schönfelder and Wichmann, 2012). This approach involves a partition of the data between a "training" and a "test" set. For each given couple of hyperparameters, we can estimate the model parameters on the training-set by maximum a posteriori, as explained previously. It thus becomes possible to assess how the model parameters would generalize to an independent dataset, by comparing the predicted responses on the test-set to the actual responses. When the model predicts the most accurately unseen data, the strength of priors is considered as optimal.

We determined one single couple of optimal hyperparameters $\left\{\lambda_{1,\,opt},\ \lambda_{2,\,opt}\right\}$ by participant. More precisely, the selection of $\lambda_{1,\mathrm{opt}}$ and $\lambda_{2,\mathrm{opt}}$ is performed on a model gathering together trials on which signal 0 or 1 was presented, following the equation:

$$P\left(r_i = 1\right) = \Phi(\underline{X}_i^T \cdot \underline{\beta} + b) \tag{10}$$

This GLM relates strongly to that derived from Equation (4), except that it does not take into account information about the target signals that was actually presented at each trial (the two level factor $\underline{s}$ being replaced with a constant term b). In particular this simple linear model cannot account for the fact that when presented with a masked target $\underline{t}_i$ the observer is more likely to respond r_i and as a consequence, it yields less accurate predictions. Nevertheless, because the estimated template $\widehat{\underline{\beta}}$ is very close to that derived from Equation (4), this model provides a good basis for estimating a common set of optimal hyperparameters, which will then be applied in all estimations of CIm for this subject. To do so, we plotted the 10-fold cross-validation rate of the model as a function of the hyperparameters $\{\lambda_1, \lambda_2\}$ used for fitting this model. The optimal hyperparameters $\left\{\lambda_{1,\,\mathrm{opt}}, \lambda_{2,\,\mathrm{opt}}\right\}$ are found by choosing the models that yielded the best prediction of responses to a new data set i.e., that correspond to a maximum of cross-validation rate. When the function exhibits several peaks, the values are chosen to favor smooth weights along the two dimensions. The same procedure was repeated for both participants. In more simple terms, this technique yields to a form of Automatic Smoothness Determination (Sahani and Linden, 2003) allowing us to apply smoothing in a principled fashion.

We assessed the statistical significance of the weights in the resulting CIm by a simple permutation test. "Resampled" estimates of the CIm were computed from 500 random reassignment of the responses to the trials (i.e., random permutation of the response vector r). We therefore obtained estimates of the distribution of weights at each time-frequency bin, under the null hypothesis of no effect of noise at this time-frequency bin.

We used these estimated distributions to highlight weights significantly different from 0 ($p < 0.005$, two-tailed) in the actual CIm.

RESULTS

The SNR was manipulated across trials via an adaptive procedure, in order to maintain the percentage of correct answers roughly equal to 75% during the course of the entire experiment. In consequence, variations of SNR provide an overview of observers' performances in the phoneme categorization task. **Figure 2A** plots the evolution of SNR during the experiment and the mean SNR for each participant, showing that there is no strong effect of perceptual learning as a decline of SNR for the same performance level over the course of the experiment. The psychometric functions are therefore estimated on all available data for each participant (**Figure 2B**). As noted by Eckstein et al. (1997), linear observers such as the one described in Equations (2) and (3) produce a linear relationship between signal contrast and detectability index (defined here as $d' = \Phi^{-1}(P_H) - \Phi^{-1}(P_{FA})$ with P_H the proportion of response 1 when signal t_1 was presented and P_{FA} the proportion of response 1 when signal t_0 was presented). For the real data, such a linear relationship can be observed in the range 2–10% signal contrast, supporting our assumption of a (at least locally) linear model for the observers. Furthermore, the small number of trials corresponding to very high or very low contrast could also account for the non-linearity.

As explained in the Methods section, an optimal set of hyperparameters is chosen by plotting the cross-validation rate of the model derived from Equation (4) and fitted by MAP estimate as a function of $\{\lambda_1, \lambda_2\}$. An example of resulting surface for subject MH is shown on **Figure 3**, with a clear maximum at $\lambda_1 = 3.16\text{e-}08$, $\lambda_2 = 1\text{e-}08$ (a similar pattern of cross-validation rate is seen for the four other subjects). The low values of cross-validation rate, ranging from 0.49 to 0.53, are explained by the fact that the very simple model described in Equation (10) does not take into account information about the target signal presented, which is critical for an accurate prediction of observer's responses. Nevertheless, it allows us to track how the calculated template generalizes to new data sets, excluding predictors other than noise. For low values of hyperparameters, the model is overfitted and cannot generalize to the "test" dataset, resulting in prediction performances around chance level (50%). For high values of hyperparameters, the classification image is flat and the model always gives the same answer, which corresponds to the response bias of the observer (in this case 52% of Michel Hoen's answers were "aba"). In between a couple of hyperparameters may be found that maximizes prediction performances.

Figure 4A shows the CIm $\widehat{\beta}$ obtained by the GLM method with smoothness priors, as well as the optimal values of λ_1 and λ_2 for the three listeners. For each participants, the classification image provides a measure of the strength of the relation between the noise at different time-frequency locations and the speech identification scores. In that sense, the classification image may be regarded as a measure of the contribution of each time-frequency bin to categorization, with high absolute values for locations at which the power of the noise influences the decision of the

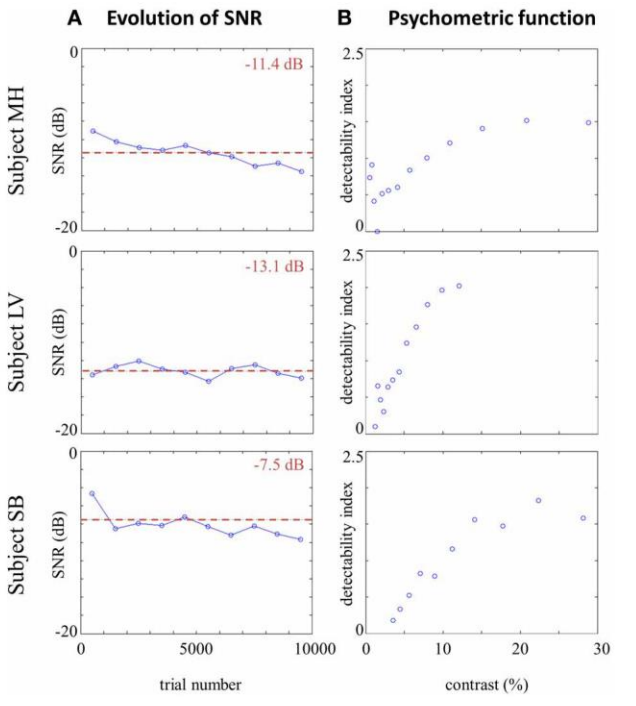


FIGURE 2 | (A) Evolution of SNR across trials (mean SNR by blocks of 1000 trials) for each participant, and overall mean SNR (red dotted line). **(B)** Psychometric function of each participant: detectability index d' (defined as $d' = \Phi^{-1}(P_H) - \Phi^{-1}(P_{FA})$ as a function of signal contrast (values calculated on less than 20 observations are not included).

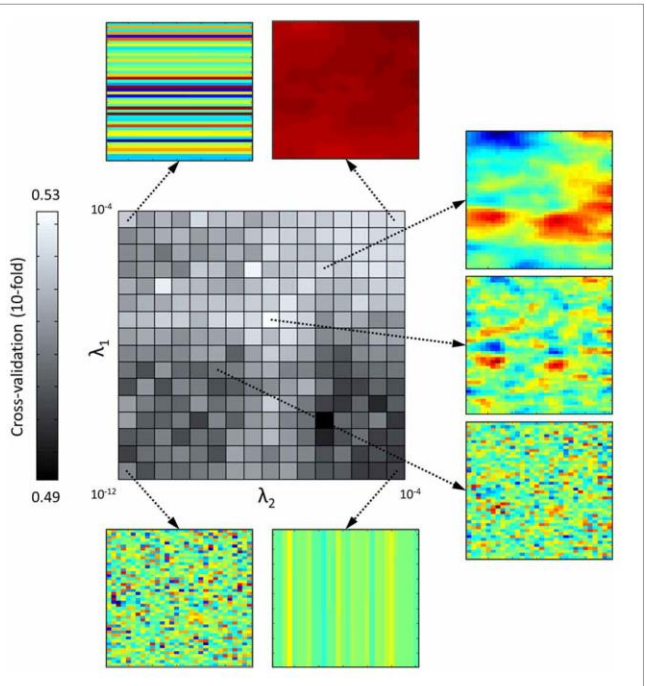


FIGURE 3 | Prediction accuracy of the model (in terms of 10-fold cross-validation rate) as a function of regularization parameters λ_1 (x-axis) and λ_2 (y-axis) in logarithmic scale, for one participant (MH). Around are shown classification images obtained with different pairs of regularization parameters (λ_1, λ_2) ($n = 10000$ trials for each estimate).

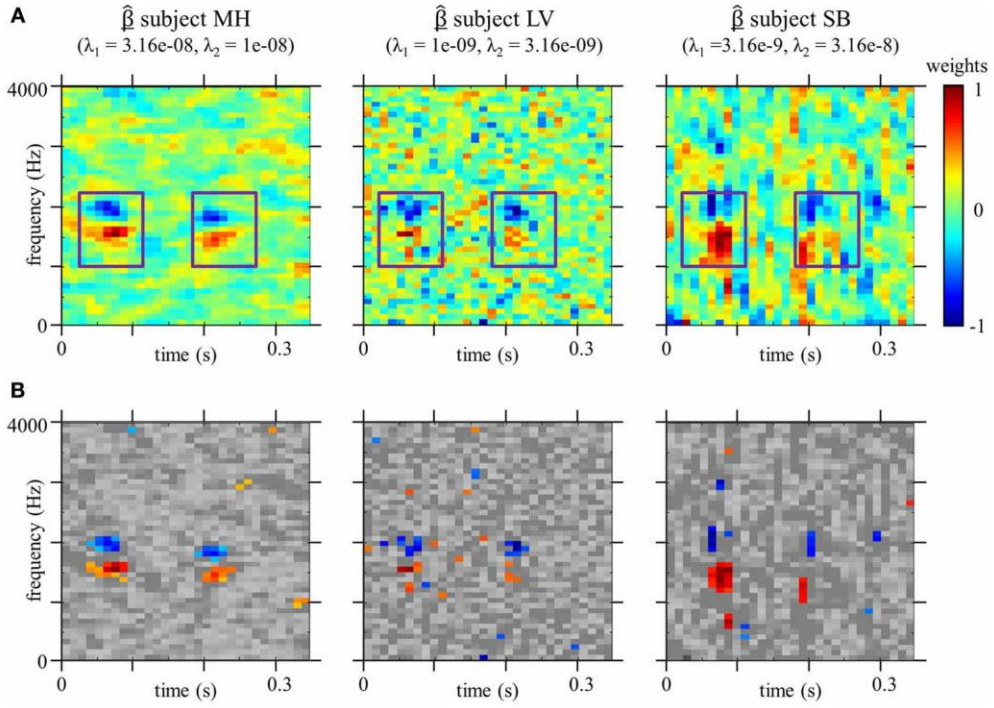


FIGURE 4 | (A) Classification Image $\hat{\beta}$ for each participants, estimated with optimal smoothness hyperparameters λ_1 and λ_2 ($n = 10000$ trials for each estimate). Weights are divided by their maximum absolute values. Boxes corresponds to the position of the second formantic transition (F2) in the original stimuli spectrograms. **(B)** Same as above except that non-significant weights are shown in gray scale ($p < 0.005$, permutation test).

observer. As can be seen from **Figure 4**, CIm often exhibit both positive (red) and negative (blue) weights corresponding to areas where the presence of noise, respectively, increases or decreases the probability of stimulus to be identified as signal $\underline{t}_0$ (/aba/) (weights are divided by their maximum absolute value to provide a common basis for comparison). **Figure 4B** shows the same classification image, but non-significant weights are represented in gray tones ($p < 0.005$, permutation test), as explained in the method section.

For a better understanding of these CIm, we ran a similar test performed by an ideal template-matcher (**Figure 5**). This classifier is the optimal observer for the linear model presented in Equation (2) and (3) and is defined by taking $\underline{w} = (\underline{T}_1 - \underline{T}_0)/K$ with K a normalization constant (difference template shown on **Figure 5A**), and $c = 0$. Note that as it is used by the template-matcher as a linear weighting function, we represented it with a linear scale, whereas speech spectrograms on **Figure 1** are classically represented using a logarithmic scale (dB). Since the difference template corresponds to the difference between the two target spectrograms, the template-matcher observer bases its classification strategy on the time-frequency bins where the spectrograms of the two signals differ most in terms of power (corresponding in this case to the region of the onset of the first formant, which appears in red on the difference template). As the performances of the algorithm do not vary over time, the SNR for stimulus presentation was set to −25 dB, for a resulting percentage of correct answers of 68%. The difference template and the obtained CIm are plotted on **Figure 5**.

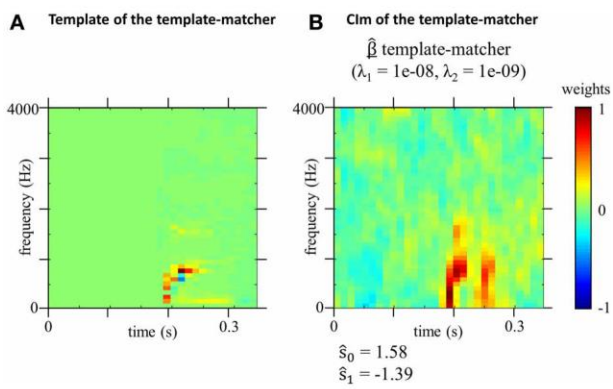


FIGURE 5 | (A) Difference template $\underline{w}$ used by the template matcher (difference between spectrograms of the targets). **(B)** Estimated model parameters for the template-matcher optimal hyperparameters λ_1 and λ_2 ($n = 10000$ trials). Weights are divided by their maximum absolute values.

The classification image obtained from the optimal observer (**Figure 5**) is very different that those obtained from human listeners (**Figure 4**), suggesting that the usage of acoustic cues by the human speech perception system is suboptimal in this particular example.

DISCUSSION

By providing insight into how a given noise distribution affects speech identification, the GLM may help to better understand the

perceptual mechanisms behind speech-noise interferences. With the present study, we demonstrated that CIm obtained from the categorization of natural speech signals, i.e., the phonemes /b/ and /d/ embedded in /aba/ and /ada/ logatomes, can offer insight into the way in which the human speech perception system achieves fast and efficient categorization of phonemes in noise. By adapting the GLM with smoothness priors to an adaptive identification task performed on speech stimuli, we have shown that CIm are applicable to studies in the auditory modality and can be used to identify relevant portions of speech.

AUDITORY CLASSIFICATION IMAGES FROM A PHONEME CATEGORIZATION TASK

Because the optimal values obtained for the smoothing parameters λ_1 and λ_2 are not the same for all participants, the calculation yields smoother CIm for MH than LV, and SB (left vs. middle and right panel on **Figure 4**). Nevertheless, a similar pattern of weights is observed for both participants. If we map the CIm obtained from our human listeners onto the original stimuli spectrograms (**Figure 1**), we can observe two main foci of high- and low-value weights, located in the time-frequency domain exactly over the second formant F2 (blue frames in **Figure 4**). More precisely, our preliminary observations suggest that, unlike the template-matcher, which bases its decisions on main energy differences between the two signals, the human observers used for categorizing /aba/ and /ada/ speech specific cues, namely the end of F2 in the first vowel and the onset of F2, on the consonant, just following the occlusion. This is in agreement with previous findings by Liberman et al. (1954): they showed that the second formantic transition can serve as a cue for classifying phonemes into /b/ or /d/, by using synthetic speech and by modulating the direction and extent of the second formantic transition. However, they did not test all possible cues and limited their study to manipulations of F2, leaving open the possibility that other portions of the signal could also be identified as functional cues for this categorization task. Our approach conversely takes into account all possible acoustic cues which might be used in the categorization and thus the results provides stronger support for the hypothesis that the second formantic transition is the only crucial characteristic for performing the task.

The pattern consistently observed at each time-frequency locations of the second formantic transition, composed of a cluster of positive weights below a cluster of negative weights, supports the assumption that frequency information is coded in terms of relative difference across frequency bands (Loizou et al., 1999). When the energy of the noise is concentrated around 2000 Hz during the formantic transition, the second formant sounds higher than it actually is, and therefore the consonant is more likely to be perceived as /d/. On the contrary, a high noise power around 1500 Hz biases the decision toward /b/. In both cases, this phenomenon is strengthened by a similar distribution of the noise at the end of F2 in the preceding vowel /a/. Indeed, the strong absolute values of weights around 0.075 s for frequencies between 1500 Hz and 2200 Hz in **Figure 4**, indicates that the decision depends on this region, even though it contains no useful information for performing the task (in our experiment the first syllable was the same for both stimuli because it was obtained by concatenating the same utterance of /a/). A very similar pattern of weights has been observed for a Vernier acuity task in the visual domain (Ahumada, 1996), highlighting the fact that our phoneme categorization task could be seen as the detection of the alignment of formants in time. In addition, the obtained CIm evidence the fact that the estimation of the second formant by the auditory system is a relative measurement, since the presence of noise masking the position of the second formant in the preceding vowel influences the decision of the observer. This is in-line with theories postulating phonemic perception as an interpretation of phonetic movements and trajectories. Further work will be dedicated to studying in details the relationship between classification image and phonetic discriminations.

This simple example illustrates the fact that our method is suitable for studying the processing of fine-acoustic cues during speech categorization by the human speech perception system. Indeed, the use of a GLM with smoothness priors as a statistical method for the estimation of CIm in the auditory modality is a reasonable way of overcoming traditional limitations of this methodological approach in the auditory modality.

First, this method allows the addition of prior knowledge about the smoothness of the expected image. By exploiting the dependencies between adjacent noise values, one can significantly reduce the number of trials required to obtain a reliable classification image. Since our goal here was to explore the possibilities of the method, participants completed a very large set 10,000 trials, in order to gather sufficient amount of information and data to be able to run accurate simulations. Nevertheless, there is in fact no need for so many trials to calculate a classification image. To get an idea of the appropriate amount of data, we estimated the model parameters at various stage of completion of the experiment for participant Michel Hoen, and calculated their correlation with the "overall" CIm (calculated on 10,000 trials) as a measure of accuracy (**Figure 6**). It can be seen from this graphs that we reached the level of $r = 0.8\%$ with approximately 6000 trials, and therefore this amount of data can be considered as sufficient to calculate a reliable estimate of the underlying template. On the other hand, below 6000 trials the optimal set of hyperparameters becomes very difficult to identify because the cross-validation rate exhibits several peaks and a less typical profile.

Second, unlike the reverse-correlation method, the GLM does not require the stimulus or the noise to be normally distributed. Accordingly, it can efficiently measure CIm using noise-fields with non-Gaussian distributions, such as the power spectrum of an acoustic-noise, in a similar way to the calculations of second-order CIm using GLMs (Barth et al., 1999; Knoblauch and Maloney, 2012). It should be noted that we could also rely on the Central Limit Theorem and assume that images are normally distributed, as long as the noise is not heavy-tailed, but this approximation leads to less precise estimation and to far less smooth CIm. In this experiment we used white noise in order to mask equally acoustic cues at low and high frequencies; however, it is known that the human auditory system does not perceive all frequency octaves with equal sensitivity (Robinson and Dadson, 1956; Suzuki and Takeshima, 2004). One option could be to use another spectral distribution that compensates for the weighting function of the auditory system, like pink noise in which

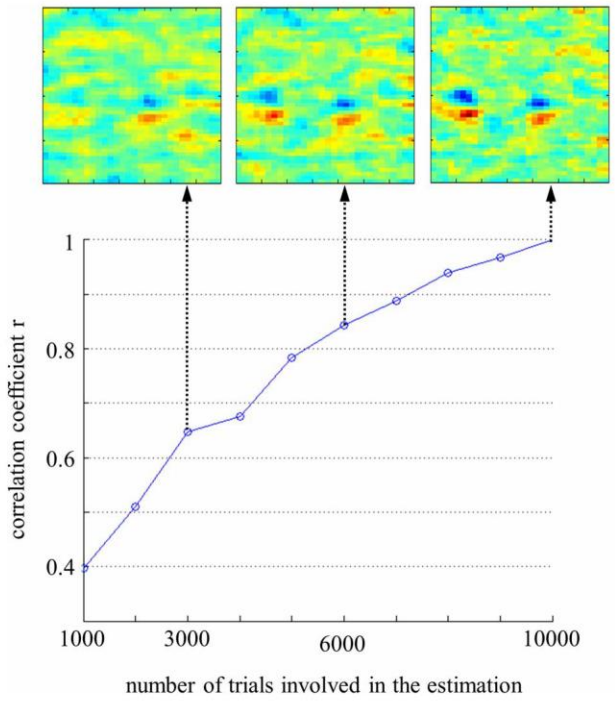


FIGURE 6 | Correlation between coefficients of the Classification Images estimated on *n* trials and the "overall" Classification Image, for participant MH. Examples of Classification Images are shown at 3000, 6000, and 10,000 trials.

all frequency octaves are assumed to be perceived with an equal loudness.

As mentioned earlier, our approach based on GLM with smoothness priors has the advantage that it does not make any assumptions about the distribution of noise in the stimulus, unlike the reverse correlation approach. Nevertheless, other strong assumptions about how observers perform speech categorization tasks in noise have been made or maintained and must be discussed.

SPECTROTEMPORAL ALIGNMENT OF TARGETS

The first simple requirement for observing functional cues involved in our identification task is the precise spectrotemporal alignment of the two targets. As we want to know in which time-frequency bins the listener is focusing, the acoustic cues of interest must be at the same time-frequency locations on the spectrogram of the stimuli on which the CIm is based. If this is not the case, the resulting CIm would probably exhibit two clusters corresponding to the same acoustic cue, instead of one. If not addressed, this issue could put into question the method, as this alignment is not trivial for natural speech. Of course, we also do not know in advance the functional cues which have to be matched between the two targets. Two possible practical solutions can be considered:

1) Forcing the temporal alignment of the targets by using synthetic speech or by cross-splicing the constant parts of the stimuli. In the above example the first syllable /a/ was the same

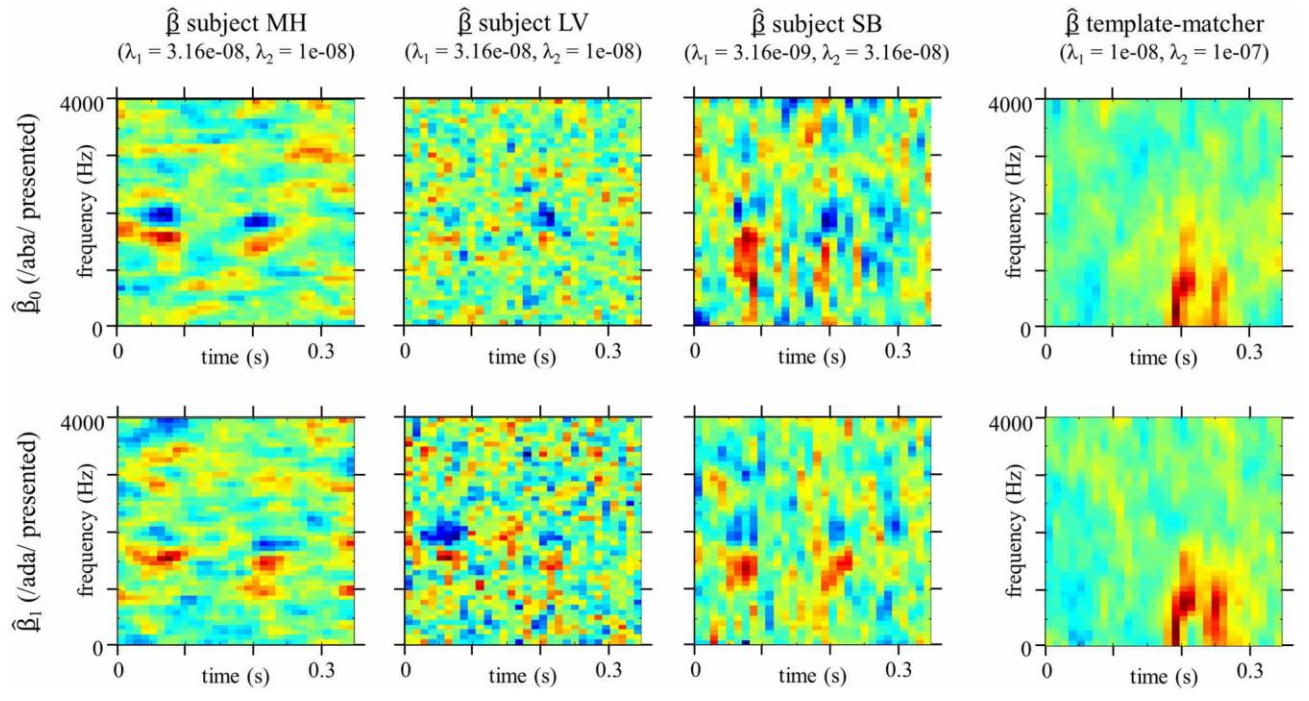


FIGURE 7 | Classification Images $\hat{\beta}_0$ and $\hat{\beta}_1$ estimated on the trials where t_0 (/aba/) or t_1 (/ada/) was presented respectively ($n = 5000$ trials for each estimate). Hyperparameters values are the same as for the "overall" Classification Images **Figure 4**. Weights are divided by their maximum absolute values.

in the two targets. This is a very convenient solution as it also ensures that participants would not rely on trivial non-speech cues to perform the task, such as a delay in the beginning of the second syllable of one target compared to the other. However in some cases we do not want to manipulate the natural utterances of the targets, and we will have to go with a second option.

2) Calculating two separate "target-specific" CIm, based only on trials where one target was presented. Therefore, we ensure that for all stimuli considered the acoustic cues are at the same time-frequency locations. This is done simply by optimizing the GLM parameters on a subset of our data, the 5000 trials where t_0 or t_1 are presented, with the same regularization parameters as for the "overall" CIm (**Figure 7**). The resulting CIm $\widehat{\underline{\beta}}_0$ and $\widehat{\underline{\beta}}_1$ are of course noisier than previous ones, because they each rely on the half of the data, but they are helpful in checking that the position of functional cues does not differ when participants are presented with one target signal or the other. The "target-specific" CIm will be discussed in more detail in the next section. This last point raises the broader issue of non-linearities in the processing of the input stimuli.

NON-LINEARITY OF THE AUDITORY SYSTEM

Our model is derived from Equation (2) defining the decision rule for a linear observer. As for all studies involving any classification image technique, we modeled the real observer performing the identification task as a template-matcher linearly combining the input sound and a decision template to calculate a decision variable. It should be noted, however, that information processing throughout the human auditory system is obviously non-linear (Goldstein, 1967; Moore, 2002).

A first type of non-linearity already mentioned occurs when the listener's strategy is not identical when targets t_0 or t_1 are presented. This phenomenon can be revealed by estimating two separate CIm $\widehat{\underline{\beta}}_0$ and $\widehat{\underline{\beta}}_1$ based on the trials where the target signals $\underline{t}_0$ or $\underline{t}_1$ were presented, respectively (**Figure 7**). Differences between the two estimates for a given observer are generally interpreted as evidence for non-linearities in the auditory system, the template used for detection depending on the input signal [Abbey and Eckstein (2006)]. For all participants the critical patterns of positive and negative weights show up at the same time-frequency locations, although sometimes less clearly because they are estimated with only 5000 trials. As expected, for the ideal template-matcher case, $\widehat{\underline{\beta}}_0$ and $\widehat{\underline{\beta}}_1$ are very close because this simulated observer is actually implemented as a linear algorithm involving a single template. Similarly, for real listeners, differences between the estimated templates appear to be less visible than in other studies involving a discrimination signal-present vs. signal-absent task (Ahumada, 2002; Thomas and Knoblauch, 2005). Note that in our experiment the amount of phase uncertainty is reduced by the presentation of a signal in both conditions.

We additionally assumed here that non-linearities in the auditory system may be locally approximated by a linear function

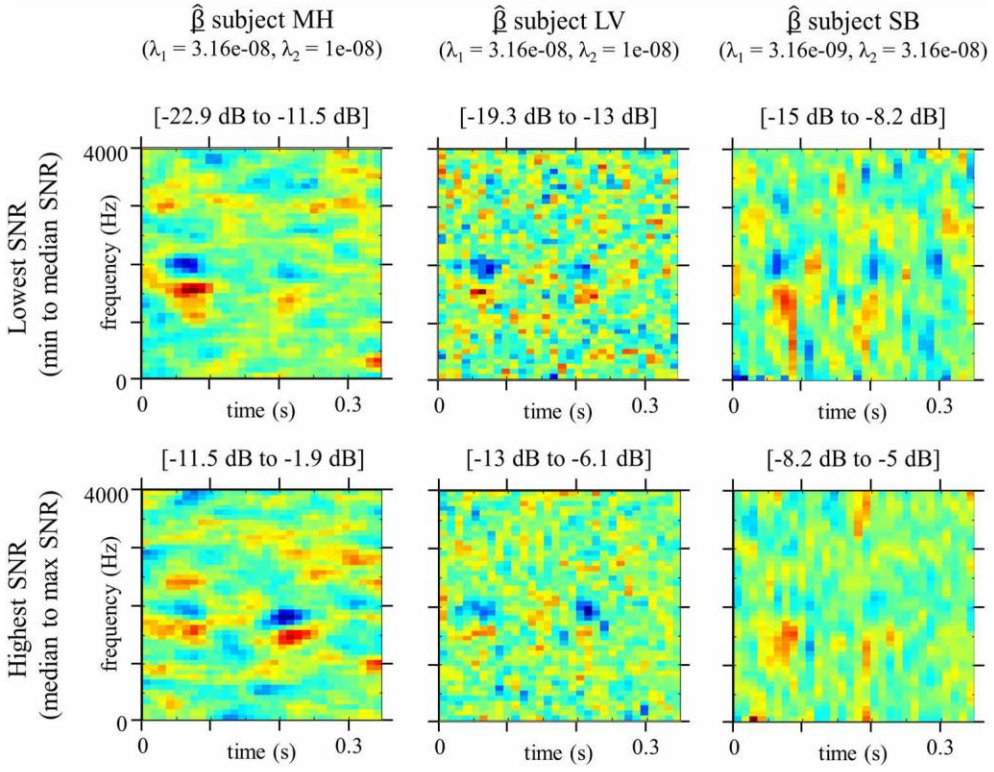


FIGURE 8 | Classification Images $\widehat{\beta}$ for conditions lowest SNR (min to median SNR) or highest SNR (median to max SNR), estimated using GLM approach with smoothness priors (n = 5000 trials for each estimate). Hyperparameters values are the same as for the "overall" Classification Images **Figure 4**. Weights are divided by their maximum absolute values.

within the SNR range studied, a hypothesis supported by the local linearity of psychometric functions (Abbey and Eckstein, 2006). To explore this assumption empirically, we estimated the model parameters for all participants by taking into account only trials with signal contrast in the linear part of their psychometric function. The resulting CIm are very similar to those obtained previously on the whole dataset.

Furthermore, even if higher-order computations are involved, the actual mechanisms of phoneme categorization are very likely to rely on time-frequency regions highlighted by our CIm because, to some extent, noise in these regions predicts the response of the participant. In that sense, the literature on visual tasks suggests that even when the strategy used by observers is clearly non-linear, CIm may still be informative about the time-frequency location of the cues involved in the categorization mechanism. As a second step, some of these studies investigate specific non-linear effects to account for the observed divergence from linearity in their results, such as spatial or phase uncertainty (Barth et al., 1999; Murray et al., 2005; Abbey and Eckstein, 2006).

ADAPTIVE SNR

Another related theoretical issue of interest here relates to the use of an adaptive-SNR method. When gathering together data from the whole experiment in order to calculate a classification image, we assume that the observer's strategy for phoneme categorization in noise does not change drastically with SNR. However, this is somewhat unlikely as a number of neurophysiological studies have highlighted significant changes in cortical activity with the level of acoustic degradation of speech sounds (Obleser et al., 2008; Miettinen et al., 2010; Obleser and Kotz, 2011; Wild et al., 2012). This led us to investigate the effect of SNR on the classification image, by estimating the model parameters only on the half of the dataset with the lowest SNR (from min to median for each participant) and on the half of the dataset with the lowest SNR (from median to max) separately (**Figure 8**). The result seems to indicate a difference in processing within the categorization mechanism: while for low SNR the estimated templates exhibits stronger weights in absolute value on the first cue, for high SNR participants appear to rely equivalently on both cues, maybe even more on the second F2 transition. A simple explanation for this phenomenon could be that, when the noise fully masks the signal, the only remaining indicator to temporally track the relevant cues is the onset of the stimulus. Therefore, temporal uncertainty is stronger on the latest cue, resulting in more dispersed weights on the second F2 transition. This example underlines that categorization mechanisms in noise do depend on SNR level and that a lower signal contrast can bias estimated weights toward earliest cues. Further developments and studies will thus be dedicated to studying the evolution of functional fine acoustic cues with SNR value and also to adapting the methods in order to account for this influence (limiting the allowed SNR range for example).

CONCLUSIONS

We have shown how an adaptation of a GLM with smoothness priors provides a suitable and powerful framework to investigate the way in which the human speech perception system achieves fast and efficient categorization of phonemes in noise and to estimate how human observers differ from ideal template matchers. Further developments and improvements of this method can be derived from the visual classification image literature (i.e., generalizing to multiple response alternatives and rating scales, see Dai and Micheyl (2010) and Murray et al. (2002)). Additionally, the possibility of calculating CIm in non-Gaussian noise makes it feasible to extend our method to more ecological situations as complex as speech-in-speech listening situations for example (Hoen et al., 2007; Boulenger et al., 2010); a situation that is well known to cause particular challenges in certain speech-development pathological conditions, for example dyslexia (Ziegler et al., 2009; Dole et al., 2012). Further developments should also deal with the issues of realizing and analyzing group studies.

ACKNOWLEDGMENTS

The first author (Léo Varnet) is funded by a PhD grant from the Ecole Doctorale Neurosciences et Cognition (EDNSCo), Lyon-1 University, France. This research was supported by a European Research Council grant to the SpiN project (no. 209234) attributed to Fanny Meunier.

REFERENCES

Abbey, C. K., and Eckstein, M. P. (2002). Classification image analysis: estimation and statistical inference for two-alternative forced-choice experiments. *J. Vis.* 2, 66–78. doi: 10.1167/2.1.5

Abbey, C. K., and Eckstein, M. P. (2006). Classification images for detection, contrast discrimination, and identification tasks with a common ideal observer. *J. Vis.* 6, 335–355. doi: 10.1167/6.4.4

Ahumada, A. J. (1996). Perceptual classification images from vernier acuity masked by noise. *Perception* 25(ECVP Suppl.), 18.

Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *J. Vis.* 2, 121–131. doi: 10.1167/2.1.8

Ahumada, A., and Lovell, J. (1971). Stimulus features in signal detection. *J. Acoust. Soc. Am.* 49, 1751–1756. doi: 10.1121/1.1912577

Ahumada, A., Marken, R., and Sandusky, A. (1975). Time and frequency analyses of auditory signal detection. *J. Acoust. Soc. Am.* 57, 385–390. doi: 10.1121/1.380453

Apoux, F., and Bacon, S. P. (2004). Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise. *J. Acoust. Soc. Am.* 116, 1671–1680. doi: 10.1121/1.1781329

Apoux, F., and Healy, E. W. (2009). On the number of auditory filter outputs needed to understand speech: further evidence for auditory channel independence. *Hear. Res.* 255, 99–108. doi: 10.1016/j.heares.2009.06.005

Ardoint, M., Mamassian, P., and Lorenzi, C. (2007). "*Internal representation of amplitude modulation revealed by reverse correlation*," in *Abstracts of the 30th ARO Midwinter Meeting, Feb* 10–15 (Denver, CO).

Barth, E., Beard, B. L., and Ahumada, A. J. (1999). "Nonlinear features in vernier acuity," in *Human Vision and Electronic Imaging III*, eds B. E. Rogowitz and T. N. Pappas (San Jose, CA: SPIE Proceedings), 3644, 88–96.

Bouet, R. and Knoblauch, K. (2004). Perceptual classification of chromatic modulation. *Vis. Neurosci.* 21, 283–289. doi: 10.1017/S0952523804213141

Boulenger, V., Hoen, M., Ferragne, E., Pellegrino, F., and Meunier, F. (2010). Real-time lexical competitions during speech-in-speech comprehension. *Speech Commun.* 52, 246–253. doi: 10.1016/j.specom.2009.11.002

Calabrese, A., Schumacher, J. W., Schneider, D. M., Paninski, L., and Woolley, S. M. N. (2011). A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. *PLoS ONE* 6:e16104. doi: 10.1371/journal.pone.0016104

Calandruccio, L., and Doherty, K. A. (2008). Spectral weighting strategies for hearing-impaired listeners measured using a correlaional method. *J. Acoust. Soc. Am.* 123, 2367–2378. doi: 10.1121/1.2887857

Cutler, A. (2012). *Native Listening: Language Experience and the Recognition of Spoken Words*. Cambridge, MA: MIT Press.

Dai, H., and Micheyl, C. (2010). Psychophysical correlation with multiple response alternatives. *J. Exp. Psychol.* 36, 976–993. doi: 10.1037/a0017171

Doherty, K. A., and Turner, C. W. (1996). Use of a correlational method to estimate a listener's weighting function for speech. *J. Acoust. Soc. Am.* 100, 3769–3773. doi: 10.1121/1.417336

Dole, M., Hoen, M. and Meunier, F. (2012). Speech-in-noise perception deficit in adults with dyslexia: effects of background type and listening configuration. *Neuropsychologia* 50, 1543–52. doi: 10.1016/j.neuropsychologia.2012.03.007

Eckstein, M. P., Ahumada, A. J., and Watson, A. B. (1997). Visual signal detection in structured backgrounds. II. Effects of contrast gain control, background variations, and white noise. *J. Opt. Soc. Am.* 14, 2406–2419. doi: 10.1364/JOSAA.14.002406

Fox, J. (2008). "Generalized linear models," in *Applied Regression Analysis and Generalized Linear Models*, *2nd Edn., Chapter* 15 (Thousand Oaks, CA: SAGE Publications), 379–424.

Gold, J. M., Sekuler, A. B., and Bennett, P. J. (2004). Characterizing perceptual learning with external noise. *Cogn. Sci.* 28, 167–207. doi: 10.1207/s15516709cog2802_3

Gold, J. M., Murray, R. F., Bennett P. J., and Sekuler A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Curr. Biol.* 10, 663–666. doi: 10.1016/S0960-9822(00)00523-6

Goldstein, J. L. (1967). Auditory nonlinearity. *J. Acoust. Soc. Am.* 41, 676–89. doi: 10.1121/1.1910396

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.

Hoen, M., Meunier, F., Grataloup, C., Pellegrino, F., Grimault, N., Perrin, F., et al. (2007). Phonetic and lexical interferences in informational masking during speech-in-speech comprehension. *Speech Commun.* 49, 905–916. doi: 10.1016/j.specom.2007.05.008

Kiggins, J. T., Comins, J. A., and Gentner, T. Q. (2012). Targets for a comparative neurobiology of language. *Front. Evol. Neurosci.* 4:6. doi: 10.3389/fnevo.2012.00006

Knoblauch, K., and Maloney, L. T. (2008). Estimating classification images with generalized linear and additive models. *J. Vis.* 8, 1–19. doi: 10.1167/8.16.10

Knoblauch, K., and Maloney, L. T. (2012). "Classification images" in *Modeling Psychophysical Data in R. Chap. 6* (New York, NY: Springer), 173–202. doi: 10.1007/978-1-4614-4475-6

Li, R. W., Klein, S. A. and Levi, D. M. (2006). The receptive field and internal noise for position acuity change with feature separation. *J. Vis.* 6, 311–321. doi: 10.1167/6.4.2

Liberman, A. M., Delattre, P. C., and Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *Am. J. Psychol.* 65, 497–516. doi: 10.2307/1418032

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychol. Monogr. Gen. Appl.* 68, 1–13. doi: 10.1037/h0093673

Liberman, A. M., Safford Harris, K., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* 54, 358–368. doi: 10.1037/h0044417

Loizou, P., Dorman, M., and Tu, Z. (1999). On the number of channels needed to understand speech. *J. Acoust. Soc. Am.* 106, 2097–2103. doi: 10.1121/1.427954

Luce, P. A., and Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear Hear.* 19, 1–36. doi: 10.1097/00003446-199802000-00001

Machens, C. K., Wehr, M. S., and Zador, A. M. (2004). Linearity of cortical receptive fields measured with natural sounds. *J. Neurosci.* 24, 1089–1100. doi: 10.1523/JNEUROSCI.4445-03.2004

Miettinen, I., Tiitinen, H., Alku, P., and May, P. (2010). Sensitivity of the human auditory cortex to acoustic degradation of speech and non-speech sounds. *BMC Neurosci.* 11:24. doi: 10.1186/1471-2202-11-24

Mineault, P. J., Barthelmé, S., and Pack, C. C. (2009). Improved classification images with sparse priors in a smooth basis. *J. Vis.* 9, 1–24. doi: 10.1167/9.10.17

Moore, B. C. (2002). Psychoacoustics of normal and impaired hearing. *Br. Med. Bull.* 63, 121–34. doi: 10.1093/bmb/63.1.121

Murray, R. F. (2011). Classification images: a review. *J. Vis.* 11, 1–25. doi: 10.1167/11.5.2

Murray, R. F. (2012). Classification images and bubbles images in the generalized linear model. *J. Vis.* 12, 1–8. doi: 10.1167/12.7.2

Murray, R. F., Bennett, P. J., and Sekuler, A. B. (2002). Optimal methods for calculating classification images: weighted sums. *J. Vis.* 2, 79–104. doi: 10.1167/2.1.6

Murray, R. F., Bennett, P. J., and Sekuler, A. B. (2005). Classification images predict absolute efficiency. *J. Vis.* 5, 139–149. doi: 10.1167/5.2.5

Norris, D., and McQueen, J. M. (2008). Shortlist B: a bayesian model of continuous speech recognition. *Psychol. Rev.* 115, 357–395. doi: 10.1037/0033-295X.115.2.357

Obleser, J., Eisner, F., and Kotz, S. A. (2008). Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. *J. Neurosci.* 8, 8116–8124. doi: 10.1523/JNEUROSCI.1290-08.2008

Obleser, J., and Kotz, S. (2011). Multiple brain signatures of integration in the comprehension of degraded stimuli. *Neuroimage* 55, 713–723. doi: 10.1016/j.neuroimage.2010.12.020

Park, M., and Pillow, J. W. (2011). Receptive field inference with localized priors. *PLoS Comput. Biol.* 7:e1002219. doi: 10.1371/journal.pcbi.1002219

Pillow, J. W. (2007). "Likelihood-based approaches to modeling the neural code," in *Bayesian Brain: Probabilistic Approaches to Neural Coding*, *Chapter 3*, eds K. Doya, S. Ishii, A. Pouget and R. Rao (MIT press), 53–70.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., et al. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* 424, 995–999. doi: 10.1038/nature07140

Robinson, D. W., and Dadson, R. S. (1956). A re-determination of the equal-loudness relations for pure tones. *Br. J. Appl. Phys.* 7, 166–181. doi: 10.1088/0508-3443/7/5/302

Saffran, J. R., and Estes, K. G. (2006). Mapping sound to meaning: connections between learning about sounds and learning about words. *Adv. Child. Dev. Behav.* 34, 1–38. doi: 10.1016/S0065-2407(06)80003-0

Sahani, M., and Linden, J. F. (2003). "Evidence optimization techniques for estimating stimulus-response functions," in *Advances in Neural Information Processing Systems*, eds S. Becker, S. Thrun, and K. Obermayer (Cambridge, MA: MIT Press), 301–308.

Scharenborg, O., Norris, D., ten Bosch, L., and McQueen, J. (2005). How should a speech recognizer work? *Cogn. Sci.* 29, 867–918 doi: 10.1207/s15516709cog0000_37

Schönfelder, V. H., and Wichmann, F. A. (2012). Sparse regularized regression identifies behaviorally-relevant stimulus features from psychophysical data. *J. Acoust. Soc. Am.* 131, 3953–3969. doi: 10.1121/1.3701832

Shannon, R. V., Zeng, F. G., Wygonski, J., Kamath, V., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science* 270, 303–304. doi: 10.1126/science.270.5234.303

Suzuki, Y., and Takeshima, H. (2004). Equal-loudness-level contours for pure tones. *J. Acoust. Soc. Am.* 116, 918–33. doi: 10.1121/1.1763601

Thomas, J. P., and Knoblauch, K. (2005). Frequency and phase contributions to the detection of temporal luminance modulation. *J. Opt. Soc. Am.* 22, 2257–2261. doi: 10.1364/JOSAA.22.002257

Wild, C. J., Davis, M. H., and Johnsrude, I. S. (2012). Human auditory cortex is sensitive to the perceived clarity of speech. *Neuroimage* 60, 1490–1502. doi: 10.1016/j.neuroimage.2012.01.035

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Boca Raton, FL: *Chapman* & Hall/CRC.

Wu, M. C. K., David, S. V., and Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.* 29, 477–505. doi: 10.1146/annurev.neuro.29.051605.113024

Xu, L., Thompson, C. S., and Pfingst, B. E. (2005). Relative contributions of spectral and temporal cues for phoneme recognition. *J. Acoust. Soc. Am.* 117, 3255–3267. doi: 10.1121/1.1886405

Ziegler, J. C., Pech-Georgel, C., George, F., and Lorenzi, C. (2009). Speech-perception-in-noise deficits in dyslexia. *Dev. Sci.* 12, 732–745. doi: 10.1111/j.1467-7687.2009.00817.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 07 June 2013; accepted: 27 November 2013; published online: 16 December 2013.

Citation: Varnet L, Knoblauch K, Meunier F and Hoen M (2013) Using auditory classification images for the identification of fine acoustic cues used in speech perception. Front. Hum. Neurosci. 7:865. doi: 10.3389/fnhum.2013.00865

This article was submitted to the journal Frontiers in Human Neuroscience.

9.3. Résumé de l'étude 1

Dans cette première étude, nous avons identifié les contraintes pour la transposition de la méthode des Images de Classification à la modalité auditive. Les deux principaux obstacles à surmonter étaient, d'une part, le nombre très important d'essais nécessaires pour le calcul, entraînant des expériences extrêmement longues et, d'autre part, la non-normalité des prédicteurs du modèle (ici, le spectrogramme du bruit). Nous avons montré, de manière théorique, que l'utilisation d'un GLM combiné avec une régularisation par lissage permet de s'affranchir de ces deux contraintes. Nous avons ainsi pu proposer une nouvelle méthode pour l'identification des primitives de la parole : les Images de Classification Auditives (ACIs).

L'application de cette méthode aux réponses de 3 participants, pour une tâche de catégorisation de phonèmes dans le bruit, a aboutit à des résultats dépassant nos attentes. Non seulement les ACIs individuelles pointent toutes les trois vers l'attaque du 2nd formant comme indice acoustique impliqué dans cette tâche, conformément aux conclusions des études menées précédemment par d'autres équipes sur des tâches similaires, mais elles indiquent également, et de manière plus inattendue, que la fin de ce même formant dans la syllabe précédente influe, lui aussi, sur la décision du participant.

10. Étude 2 : Analyse statistique des images de classification auditives obtenues pour un groupe de participants : l'exemple de l'expérience de Mann.

10.1. Présentation de l'étude 2

L'article précédent montre qu'il est possible de calculer une CI auditive pour visualiser les indices acoustiques utilisés par un auditeur donné, à travers l'exemple de 3 participants engagés dans une tâche de catégorisation dans le bruit. Néanmoins, en vue de l'application de cette méthode dans le cadre d'études expérimentales, il est nécessaire de pouvoir augmenter le nombre de participants et de disposer d'une forme de validation statistique sur les données du groupe pour espérer tirer des conclusions à l'échelle d'une population. Il s'agit donc d'intégrer la procédure de l'ACI à la démarche classique du test d'hypothèse. À cette fin, nous avons donc réalisé une seconde étude, également comportementale, mais impliquant cette fois-ci 17 auditeurs.

Comme nous l'avons déjà souligné dans la partie théorique, les tests statistiques sur les CIs sont relativement peu développés. Nous sommes confrontés ici à deux problèmes qui rendent impossible l'emploi des tests « standards » : d'une part, la multiplicité des comparaisons à effectuer (une par pixel temps-fréquence) et, d'autre part, l'existence de dépendances entre les pixels adjacents. À ce stade, il devient indispensable de proposer des procédures complémentaires afin de pouvoir mener à bien une étude expérimentale complète basée sur les ACIs. L'article qui suit propose l'étude d'un cas de catégorisation dans le bruit pour un groupe de participants dont les résultats seront soumis à une analyse statistique détaillée. À cette fin, nous utiliserons des outils provenant du champ de la neuro-imagerie, où les chercheurs sont confrontés à des problèmes similaires (multiplicité et dépendance des tests statistiques).

De manière générale, le protocole expérimental est identique à celui de l'étude précédente. La catégorisation phonémique porte ici sur l'opposition /da/-/ga/. Pour tenter d'évaluer l'impact des effets de coarticulation sur la stratégie d'écoute, deux exemplaires de chaque cible étaient présentés, précédés de /al/ ou de /aʁ/. Comme nous l'avons déjà signalé dans la partie théorique, des tâches de catégorisation semblables avaient été réalisées auparavant – essentiellement en parole synthétique – nous conduisant à formuler deux hypothèses :

(1) les indices acoustiques impliqués sont les attaques des 2ème et 3ème formants.

(2) le contexte a une forte influence sur la position des frontières phonémiques.

Le but poursuivi ici est donc de parvenir à tester et quantifier ces effets grâce à l'analyse statistique détaillée des ACIs.

Cette étude a fait l'objet d'une publication dans la revue en accès ouvert PLoS One (Varnet et al., 2015a). Ces résultats ont également été présentés publiquement lors de communications orales au Knight's Lab (Berkley) et au SpiN Workshop 2014 à Marseille (Varnet et al., 2014b)

10.2. Article 2 : A Psychophysical Imaging Method evidencing Auditory Cue Extraction during Speech Perception: A Group Analysis of Auditory Classification Images

PLOS | ONE

RESEARCH ARTICLE

A Psychophysical Imaging Method Evidencing Auditory Cue Extraction during Speech Perception: A Group Analysis of Auditory Classification Images

Léo Varnet[1,3,5]*, **Kenneth Knoblauch**[2], **Willy Serniclaes**[6], **Fanny Meunier**[3,5], **Michel Hoen**[1,4,5]

1 Lyon Neuroscience Research Center, CNRS UMR 5292, Auditory Language Processing (ALP) research group, Lyon, France, **2** Stem Cell and Brain Research Institute, INSERM U 846, Integrative Neuroscience Department, Bron, France, **3** Laboratoire sur le Langage le Cerveau et la Cognition, CNRS UMR 5304, Auditory Language Processing (ALP) research group, Lyon, France, **4** INSERM U1028, Lyon Neuroscience Research Center, Brain Dynamics and Cognition Team, Lyon, France, **5** Université de Lyon, Université Lyon 1, Lyon, France, **6** Université Libre de Bruxelles, UNESCOG, CP191, Bruxelles, Belgique

* leo.varnet@isc.cnrs.fr

Citation: Varnet L, Knoblauch K, Serniclaes W, Meunier F, Hoen M (2015) A Psychophysical Imaging Method Evidencing Auditory Cue Extraction during Speech Perception: A Group Analysis of Auditory Classification Images. PLoS ONE 10(3): e0118009. doi:10.1371/journal.pone.0118009

Academic Editor: Christian Friedrich Altmann, Kyoto University, JAPAN

Received: October 24, 2014

Accepted: January 5, 2015

Published: March 17, 2015

Data Availability Statement: All relevant results are within the paper and its Supporting Information files. All raw data are available from Zenodo, including speech stimuli (doi: 10.5281/zenodo.12300), noise stimuli (doi: 10.5281/zenodo.12374), responses from all participants (doi: 10.5281/zenodo.12303) and results of preliminary screening tests (doi: 10.5281/zenodo.12347).

Funding: LV is funded by a PhD grant from the Ecole Doctorale Neurosciences et Cognition (http://nsco.universite-lyon.fr/), Lyon-1 University, France. This research was partially supported by a European

Abstract

Although there is a large consensus regarding the involvement of specific acoustic cues in speech perception, the precise mechanisms underlying the transformation from continuous acoustical properties into discrete perceptual units remains undetermined. This gap in knowledge is partially due to the lack of a turnkey solution for isolating critical speech cues from natural stimuli. In this paper, we describe a psychoacoustic imaging method known as the Auditory Classification Image technique that allows experimenters to estimate the relative importance of time-frequency regions in categorizing natural speech utterances in noise. Importantly, this technique enables the testing of hypotheses on the listening strategies of participants at the group level. We exemplify this approach by identifying the acoustic cues involved in da/ga categorization with two phonetic contexts, Al- or Ar-. The application of Auditory Classification Images to our group of 16 participants revealed significant critical regions on the second and third formant onsets, as predicted by the literature, as well as an unexpected temporal cue on the first formant. Finally, through a cluster-based nonparametric test, we demonstrate that this method is sufficiently sensitive to detect fine modifications of the classification strategies between different utterances of the same phoneme.

Introduction

In speech perception, we unconsciously process a continuous auditory stream with a complex time-frequency structure that does not contain fixed, highly reproducible, or evident boundaries between the different perceptual elements that we detect in the stream of speech. Phonemes [1] or syllables [2], the building-blocks of speech, are sophisticated perceptual entities.

Research Council (http://erc.europa.eu/) grant for the SpiN project (No. 209234) attributed to FM and by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Through a long evolutionary process, human brains have learned to extract certain auditory primitives from the speech signal and associate them with different perceptual categories. For example, we perceive the sounds /d/ or /g/ as discrete and distinct elements, without being aware of the underlying perceptual construction causing their distinction [3,4]. Which acoustic features are extracted and used to perceive speech remains unknown, largely because of the lack of an experimental method enabling the direct visualization of auditory cue extraction. The aim of this paper is to propose and demonstrate the validity of adapting the classification image framework to directly visualize auditory functional cues actually used by individual listeners that are processing speech.

Acoustic cues for speech perception

Speech is a continuous waveform comprising an alternation of harmonic and non-harmonic acoustic segments. Periodic sounds are caused by vibrations of the vocal folds and are shaped by resonances of the vocal tract to produce formants in the acoustic signal [5]. Thus, formants correspond to local energy maxima inside the spectral envelope of the signal and are present for vocalic sounds (*e.g.*, vowels such as /a/ or /u/) or voiced consonants (*e.g.*, /v/, /d/, or /g/). The number of formants is typically 4 to 5, depending on the phoneme considered. Formants cover a frequency range from approximately 150 Hz to 4 to 6 kHz, with approximately one formant per kHz, and last approximately 100 ms. Each vowel appears to be loosely tied to a specific formantic structure (essentially determined by the height of the first two formants, F1 and F2). Perturbations of the acoustic flux created by the rapid occlusion or release of the air flow generate silences, hisses, bursts or explosions that constitute the core of consonantal sounds (*e.g.*, consonants such as /t/, /p/ or /sh/). Their presence transitorily inflects the formant trajectories, thus creating brief formantic transitions. The formantic structure and formant dynamics are examples of spectrotemporal acoustic cues that could be exploited at the acoustic/phonetic interface [6]. By studying coarticulated utterances of /alda/, /alga/, /arda/, and /arga/, [7] determined that (1) the identity of the first consonant affected the spectral content of the second syllable, and vice-versa, and that (2) listeners were able to compensate for this coarticulation during perception. Although the first effect is clearly due to the partial overlapping of articulatory commands between adjacent phonemes, the exact nature of the compensation phenomenon remains undetermined [8–11]. Coarticulation introduces internal variations into the system referred to as allophonic variations: a range of different formantic structures will be perceived as the same phoneme. This phenomenon makes the system more resistant to intra- and inter-speaker variations, but it also makes the problem of learning to associate acoustic cues to phonemic percepts more difficult and the reverse engineering problem of designing automatic speech recognition and automatic speech comprehension systems largely unresolved [12].

Identifying auditory primitives for speech perception

The precise mechanism underlying the transformation from continuous acoustical properties into the presence or absence of some acoustic cues and finally into a discrete perceptual unit remains undetermined. The acoustic-phonetic interface has been studied extensively since 1950. Many studies on this topic have been conducted under experimental conditions, which have involved stimuli that were degraded in a controlled fashion in order to narrow the problem to a small number of possible cues. Among the most well-known attempts is the series of papers published by the Haskins Laboratories on the relationship between second formant transition and stop consonant perception using synthetic speech [13,14]. However, their conclusions are inherently limited by the non-naturalness of the synthesized stimuli: the variations of synthetic stimuli are restricted to a small number of cues, and thus they may not be processed in the

same manner as natural stimuli. Furthermore, participants exposed to this type of stimuli often report them as "unnatural" and typically achieve lower recognition performances, a clear sign that the natural cues are poorly represented in synthesized speech. More recent work has also relied on drastic spectral and/or temporal impoverishment of the speech signal [15,16]. However, in a "real-life" situation, listeners are not typically required to address filtered speech but have access to the entire spectrum. As before, the question remains: Are the evidenced acoustic cues with synthetic speech identical to those for natural speech? The resistance of speech intelligibility to drastic signal reductions, such as those noted above, could rely on secondary perceptual cues not used in natural listening situations. Scientists seeking to address this problem will ultimately be required to use natural speech production as stimuli.

In this context, a recent solution demonstrates the merits of using a masking noise on natural speech utterances to isolate the regions of the spectrogram crucial for identifying a particular phoneme. The technique initially proposed by [17] involves masking natural speech utterances with noise at various signal-to-noise ratios (SNRs). By analyzing the patterns of confusion in a participant's responses with respect to the noise level, researchers were able to identify the point at which noise masks the crucial acoustic cue, thus corresponding to a drop of correct identifications [18,19].

Alternative approaches for determining the mapping of sounds to mental representations of acoustic cues have been enabled by recent statistical developments in neuroimaging, including advances in the multivariate encoding/decoding models of neural activity. By reverse-engineering the processing of speech in the brain, it has become possible to reveal the encoding of subphonological information in the auditory cortex [20,21]. One such solution has been to record the firing rate modulations of individual auditory neurons in response to specific stimuli to derive their spectrotemporal receptive-fields (STRFs), which are a linear approximation of the time-frequency function of the neuron. This technique has been widely used in studying birds, specifically when hearing conspecific birdsongs [22,23]. These studies have demonstrated that auditory neurons are tuned to specific time-frequency regions, surrounded by one or more inhibitory regions. Spectrotemporal filters are assumed to be somewhat similar for human auditory neurons. Electrocorticographical (ECoG) recordings have enabled the estimation of average STRFs for small groups of human auditory neurons in epileptic patients [24], thereby strengthening the idea that the basic auditory cues for humans are also composed of an excitatory region surrounded by inhibitory regions. As a next step, [20] gathered STRFs from clusters of neurons that are functionally similar, e.g., auditory neurons responding preferentially to particular phonemes. They obtained the first images of the encoding of acoustic cues for several features, as well as the tuning of neurons to frequencies corresponding to formant values. Although these results represent a major breakthrough in understanding how speech sounds are primarily decoded along the primary auditory pathway, it is difficult to infer how this information is combined to facilitate the identification of one phoneme rather than another phoneme. Computational models have been proposed [25] that link the STRF with a multiresolution representation of speech sounds in the auditory cortex. This approach could provide a unified model of the transformation of a speech input from the cochlea to the midbrain. However, this account currently remains theoretical, because of the lack of a method allowing the observation of the use of acoustic cues in normal participants and other non-epileptic patients and large-group studies or studies on the individual variations of these processes.

The auditory classification image approach

In a previous paper [26], we demonstrated the feasibility of addressing this gap in the auditory domain by adapting a method designed to identify the primitives of simple perceptual tasks,

the classification image technique. Inspired from an auditory tone-in-noise detection experiment by Ahumada and Lovell [27], classification images have then been developed in the visual domain and successfully used to study Vernier acuity [28], perceptual learning [29,30], the interpolation of illusory contours [31], the detection of luminance [32] and chromatic [33] modulations, and recently face pareidolia [34]. We developed the Auditory Classification Image (ACI) technique by transposing this experimental and statistical framework to an auditory categorization task between two target speech sounds (/aba/ and /ada/). The two signals were presented in an additive Gaussian noise, and participants were asked to indicate whether the target was /aba/ or /ada/. Each participant's response was then linked to the specific noise configuration in the corresponding trial with a Generalized Linear Model (GLM) with smoothness priors. The rationale underlying this technique is that if the time-frequency coordinates at which the noise interferes with the decision of the observer are known, then the regions on which the observer focuses to perform the task would also be known. By fitting the decision weights corresponding to every pixel of the representation, it became possible to draw a time-frequency map of the categorization strategy and directly visualize which parts of the stimulus are crucial for the decision.

In the first report on ACIs, we only reported individual data on three volunteers and used two speech productions as targets, thus leaving the question of the specificity of the obtained ACIs to these particular utterances unanswered. In the present study, we aimed to 1) further develop the method and complete a first group study to extend the feasibility of the method to group studies; 2) apply statistical tests permitting the evaluation of statistical significance inside or between classification images and 3) explore the specificity of the ACI to the utterances used as targets. To this end, we acquired auditory classification images from a group of 16 participants performing 10,000 categorizations of the four /alga/, /alda/, /aʁga/, /aʁda/ speech sounds.

Materials and Methods

Participants

Seventeen native speakers of French with no background knowledge of phonetics and phonology participated in this study. All participants had normal hearing, as confirmed by a standard pure-tone audiometric test (<20 dB HL, 125–8,000 Hz), and reported no history of neurological or developmental disorders. Additionally, participants were administered a series of tests on nonverbal intelligence (Raven's Standard Progressive Matrices) and phonological and reading skills (ECLA-16+). They obtained scores within normal ranges on all tests (S1 Table). The study was approved by the Comité d'évaluation éthique de l'Inserm / Institutional Review Board (IORG0003254, FWA00005831, IRB00003888). Participants provided written informed consent before participating in the experiment, and they received financial compensation (€100) for their participation. One participant's data were excluded from further analyses due to extremely poor performances that suggested the participant had misunderstood the instructions. Thus, the analyses are based on the answers of 16 participants (12 females; mean age: 22.6 years ± 4.6 years S.D).

Stimuli

Four speech samples, i.e., /alda/, /alga/, /aʁda/, and /aʁga/, were recorded from one male speaker in a soundproof chamber at a sample rate of 48 kHz. The 4 stimuli were obtained by removing the silent gap between the two syllables to align the onset of the second syllable at the same temporal position and then equating the 4 sounds in root mean square and in duration (680 ms). The resulting speech signals (hereafter denoted $\underline{t}$) sounded perfectly natural and were perfectly intelligible in a quiet setting.

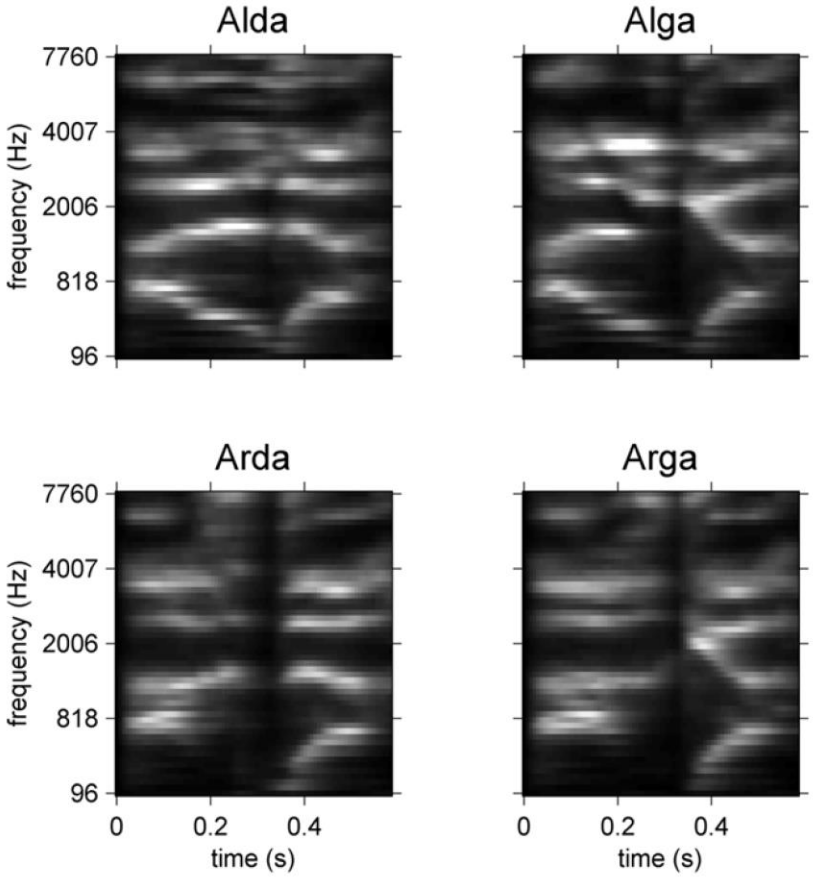


Fig 1. Cochleograms of the four stimuli involved in the experiment. Parameters for spectral and temporal resolution are identical to those used to derive the ACIs (see details in the main text).

doi:10.1371/journal.pone.0118009.g001

Each stimulus $\underline{s}$ in this experiment consisted of one target signal $\underline{t}$ embedded in an additive Gaussian noise $\underline{n}$ at a given SNR using Equation (1).

$$\underline{s}_i = \alpha_i \cdot \underline{t}_{k_i} + \underline{n}_i \quad (1)$$

where i is the trial number; k_i the signal number associated with this trial; and α_i a factor determining the SNR during the course of the experiment ($\alpha_i = 10^{\frac{SNR_i}{20}}$, for $\underline{n}_i$ and $\underline{t}_{k_i}$ both normalized in power and SNR in dB). The sampling rate of the stimuli was set to 48 kHz for the original sounds. All stimuli were root-mean-square normalized and were then preceded by 75 ms of Gaussian-noise with a Gaussian fade-in to avoid abrupt attacks. The cochleograms of the 4 stimuli are shown in Fig. 1.

Experimental procedure

Participants were seated in a sound booth in front of a computer monitor and wore Sennheiser's HD 448 headphones. They completed a set of 10,000 trials consisting of 2,500 noisy presentations of each of the 4 speech signals, presented in random order. For each trial, they were asked to listen carefully and then to indicate, by a button press, whether the final syllable was 'da' or 'ga'. The response to trial i is denoted r_i(= 0 for 'da' and 1 for 'ga'), and the correct answer (corresponding to the target actually presented) is denoted c. Participants were allowed to

replay the stimulus before entering their response. For each trial, the participant's response, his/her response time, the SNR level, and the time-frequency configuration of noise $\underline{n}_i$ were recorded for offline analysis.

Given the long duration of the experiment (approximately 4 h), we divided it into 20 sessions of 500 trials completed over 4 days to avoid mental and auditory fatigue. Sessions were separated by minimum breaks of 3 min. In addition, there was a short practice block before the beginning of the experiment that was similar to the test phase, except that the correct answers were displayed after each trial. Over the course of the experiment, the SNR was adapted from trial-to-trial based on the participant's responses by a 3-down 1-up staircase procedure [35], thereby allowing us to constantly target the 79% correct point on the psychometric function. The SNR was increased by one step after each incorrect response and decreased by one step after three consecutive correct responses from the last change in stimulus intensity. At the beginning of each session, the step size was set to 2 dB to accelerate convergence and then decreased 10% by each step until a step size of 0.2 dB was attained. The initial SNR level was -11 dB, and each session began with the final SNR of the previous session.

Generating Auditory Classification Images

The method used for deriving ACIs has been previously detailed [26]. A summary is provided below, with a focus on several improvements that have been introduced since the publication of the first version.

Cochleograms

The same preprocessing was applied to all noise and speech sounds before analysis. Cochleograms were generated from the waveforms using Lyon's Cochlea Model [36], implemented in the classic Matlab Auditory Toolbox by Malcom Slaney (https://engineering.purdue.edu/~malcolm/interval/1998-010/). This algorithm involves a bank of simulated auditory filters of constant quality factor (Q = 8), spaced quasi-logarithmically and overlapping by 40% (this step factor was chosen to be slightly greater than the default parameter proposed by Slaney to ensure sufficient spectral resolution). The vertical axis of the cochleogram represents the center frequencies of each auditory filter. Two additional processing levels are implemented in this function to mimic the non-linear behavior of the hair cells: a half-wave rectifier followed by an Automatic Gain Control modeling the neural adaptation mechanism and a difference between adjacent filters to improve the frequency response. Finally, the output of the model is decimated in time to a sample rate of 64.1 Hz, with a time step of 15.6 ms. The cochleograms of our 4 stimuli are presented in Fig. 1. The cochleogram of the noise sound at each trial i was calculated and will be hereafter denoted by $\underline{X}_i$ in its vectorized form.

Generalized Linear Model

For each participant, several ACIs were derived from the responses to all or part of the 10,000 trials using a GLM. This statistical model links the probability that the participant responded with 1, $P(r_i = 1)$, with the specific configuration of the noise through the following equation:

$$P(r_i = 1) = \Phi(\underline{X}_i^T \cdot \underline{\beta} + b_{c_i}). \quad (2)$$

where Φ a psychometric function (here, the inverse of the logit function); $\underline{\beta}$ the decision template; and $\underline{b}$ a two-level factor reflecting the influence of the target actually presented on the response. Phoneme categorization is regarded in this context as a simple template-matching process between the input sound and two mental representations of the targets stored in

memory. The decision template corresponds to a particular linear weighting of the noise profile and is specific to the two targets involved in the task. The output of the dot-product $\underline{X}_i^T \cdot \underline{\beta}$ is added to the factor $\underline{b}$ to yield a linear predictor that is eventually transformed nonlinearly through the psychometric function into a probability ranging between 0 and 1. It is important to note that the GLM does not simulate the internal processing of the human speech perception system. However, it is useful for determining which variations of the stimulus affect human perception. Thus, our main goal was to approach the decision template $\underline{\beta}$ with an estimator $\underline{\hat{\beta}}$, the ACI.

Smoothness priors

Rather than directly estimating the model parameters $\underline{\theta} = \{\underline{\beta}, \underline{b}\}$ with a simple maximization of the log-likelihood $L(\underline{\theta}) = \log(P(\underline{r}|\underline{\theta}, \underline{c}, \underline{\underline{X}}))$, we introduced a smoothness prior during the optimization of the GLM. This statistical development, named "Penalized Likelihood," or "Generalized Additive Model" (GAM), has been widely used for estimating receptive fields at the neuron level [22,37] and then adapted to the Classification Images method [38,39]. The main concept is to place constraints on the parameter values during the estimation process. This method has been shown to be efficient for preventing the overfitting problem inherent in maximum likelihood estimation when processing a large number of parameters. In the present case, overfitting would generate a noisy ACI, which would thus describe mainly random noise, not the underlying mechanism involved in the classification. The direct consequence would be that this model would closely fit the data on which it is trained but would not be able to predict responses to novel stimuli.

The introduction of a smoothness prior allows us to reduce noise in the classification image method by applying low-pass smoothing in a principled manner and therefore to minimize overfitting [40,41]. We characterize the smoothness of the ACI with the quadratic form $Q(\underline{\theta}) = Q(\underline{\beta}) = \underline{\beta}^T \underline{\underline{L}} \underline{\beta}$, where $\underline{\underline{L}}$ is the Laplacian matrix, encoding the adjacency between the pixels of the spectrotemporal representation [37,42]. This function assumes higher values when neighboring weights in the ACI markedly differ. The smoothness is assumed to be equal over the two dimensions of the ACI (time and frequency). Note that this assumption can be more or less plausible depending on the sampling rates of the time and frequency axes. We can address the issue by using two separate smoothing priors. However, doing so would dramatically increase the computational cost in our case.

Rather than maximizing the log-likelihood, we maximize the log-posterior, thereby yielding a maximum a posteriori (MAP) estimate:

$$\underline{\hat{\theta}}_{MAP} = \underset{\underline{\theta}}{\operatorname{argmax}}[L(\underline{\theta}) + \lambda \cdot Q(\underline{\theta})] \tag{3}$$

In this context, $\lambda \cdot Q(\underline{\theta})$ is called the "Regularizer" and corresponds to our a priori beliefs regarding the model parameters. In this equation, it acts as a penalty term, assigning a high cost for strong deviations between neighboring parameters and thus enforcing smoother estimates. λ is called the "hyperparameter" as it does not appear in the model Equation (2) but affects the final estimate. It controls the tradeoff between fitting the data well and obtaining a smooth ACI by determining the degree to which roughness should be penalized (higher penalty for increasing values of λ; for $\lambda = 0$, we recover the non-penalized maximum likelihood estimate). Given $\underline{c}$, $\underline{\underline{X}}$, and $\underline{r}$, it is possible to estimate the model parameters $\underline{\hat{\theta}}_{MAP}$ associated with a given hyperparameter value λ using the function *glmfitqp* developed by Mineault in his MATLAB toolbox *fitglmqp*. Examples of such λ-dependent ACIs are shown in Fig. 2.

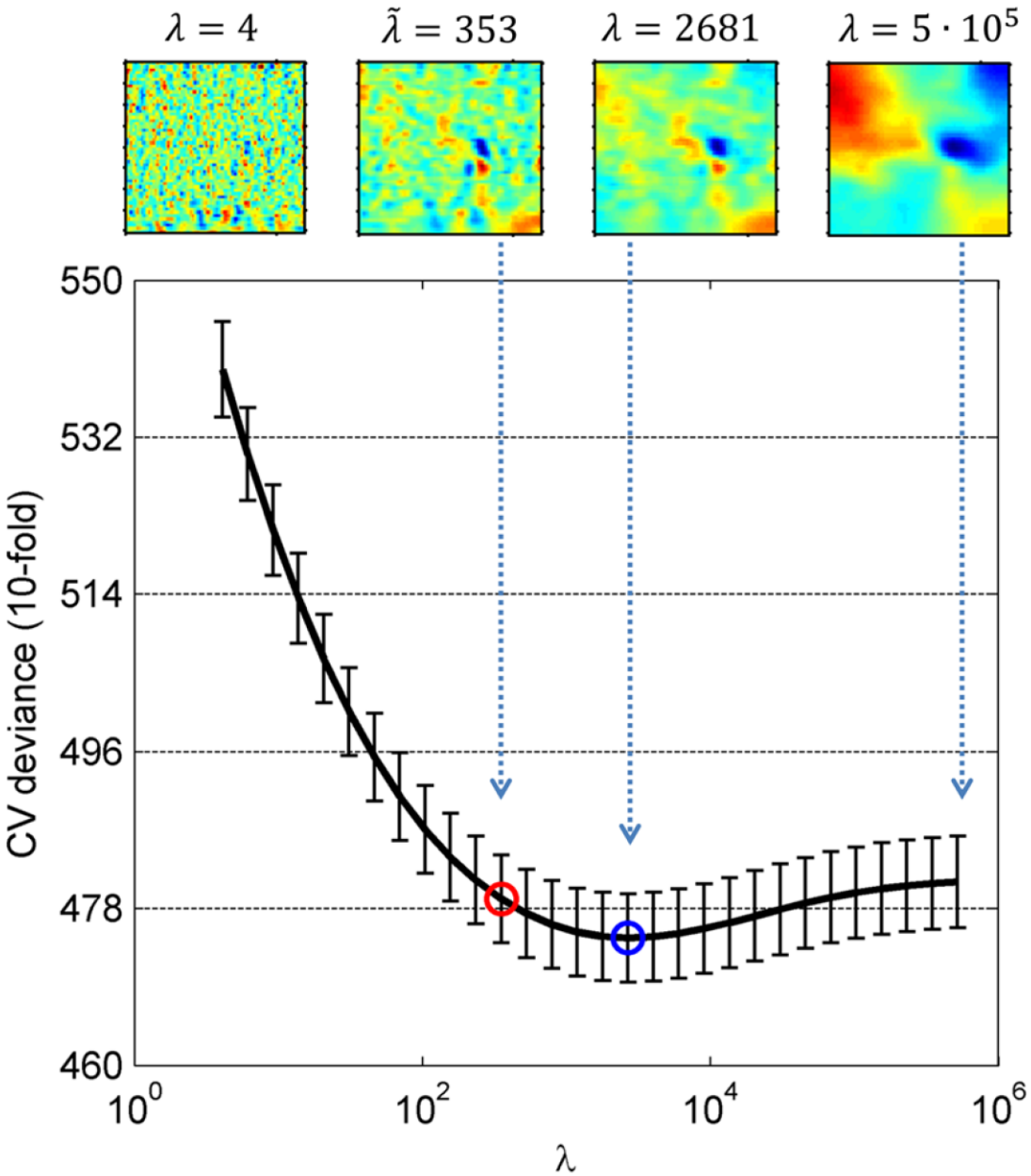


Fig 2. CV deviance of the penalized GLM for participant #17 as a function of regularization parameter *λ*. The minimum value of the CVD function is indicated with a blue circle, and the optimal lambda $\tilde{\lambda}$ is indicated with a red circle. Examples of ACIs obtained with different values for this participant are shown below.

doi:10.1371/journal.pone.0118009.g002

Lambda optimization

However, consistent with the literature on STRF and CI estimations using penalized likelihood [37,38,40], we do not want to presume an a priori value for λ. Instead, we aim to determine how much smoothing is appropriate based on our data. Because we aim for the ACIs to be generalizable to an independent dataset, models corresponding to different smoothing values are

evaluated with a cross-validation approach, and we determine an optimal regularization parameter, denoted $\tilde{\lambda}$, according to this criterion.

Individual optimization. For example, in our previous paper [26], we computed a 10-fold cross-validation for a wide range of λ values by randomly partitioning the dataset each time between a "training" set and "test" set, thereby estimating the model parameters on the training set (9,000 trials) through MAP for all considered λ values, as explained previously, and then comparing the predicted responses on the test set (1,000 trials) to the actual responses of the participant. The same procedure was followed in the present study, except that the match between predicted and observed responses was assessed by computing the deviance of each fitted model applied to the test set. This deviance is a more natural measure of goodness-of-fit in the case of GLMs [41]. In this manner, each λ value is associated with a corresponding cross-validated deviance $CVD(\lambda)$, which is the mean deviance for the 10 cross-validations (see Fig. 2 for an example of one participant). For small values of λ, the estimate is overfitted and unreliable for predicting unseen data, thus generating a high CVD. As λ increases, the true estimate emerges (with a subsequent decrease in CVD) and is finally smoothed out for high λ values. This final step generally corresponds to a slow increase in CVD.

Thus, an optimal hyperparameter could be found by selecting the model that yields a minimum CVD value, that is to say, the degree of smoothness of the ACI that allows the most accurate predictions of unseen data. However, the increase of this function for high lambda values is sometimes relatively small, thus causing an oversmoothing of the estimate. Thus, we rather selected the smallest λ value at which the CVD becomes smaller than the minimum plus one standard deviation, denoted $\tilde{\lambda}$. A similar criterion, the "one-standard-error" rule, is presented in [43] and is implemented with the MATLAB function *lassoglm*.

Conjoint optimization. Gathering data from N participants enables the selection of the lambda values conjointly. Rather than estimating N distinct optimal hyperparameters, we can select a single value $\tilde{\lambda}$ to apply to all models. Considering the sum of the individual CVDs enables the derivation of two optimal lambda values for the group identically as performed above (with the standard deviation estimated over the group of participants' CVDs). The resulting models may not predict the answers of the participants as well as the individual models; however, a major advantage of this method is that it provides increased stability of the hyperparameter selection, even with a limited number of trials. Furthermore, a common degree of smoothing may be required to gather different images in a group analysis. The convergence and stability of the lambda optimization will be investigated below.

Statistical tests

Because the ACI estimate is intrinsically noisy, certain quantitative measures are required to distinguish random or non-significant effects in the images from actual functional cues. In previous works, we performed a bootstrap test to identify significant observations at the individual ACI level. In practice, experimenters will typically be more interested in formulating generalizable conclusions about a population of subjects rather than a simple sum of remarks on the behavior of several single subjects. Group analyses provide a method of testing hypotheses on the probability distribution of weights from which the individual ACIs are drawn. Hence, they will allow us to make generalizations about a population from a sample. Statistical tests were conducted at the group level for two purposes: 1) to identify the significant differences between ACIs calculated under two conditions and 2) to assess the significance of the weights for one condition. Statistical tests were performed on the z-scored individual ACIs. In both cases, the tests involved as many comparisons as there are parameters in one ACI (4,374 in our case); therefore, it is important to correct for multiple comparisons [44].

When comparing ACIs between two conditions, we used a cluster-based non-parametric test. This statistical procedure, originally developed to analyze neuroimaging data [45–47], allows the correlation inherent to the natural images to be taken into account (i.e., each pixel depending on the values of the adjacent pixels). Statistical analyses were conducted using FieldTrip, an open-source MATLAB toolbox developed for processing electrophysiological data [48]. The test is performed at two statistical levels. First, a running paired t-test is performed on all participants and compares weights at each time-frequency bin between the two conditions of interest. Second, the result is corrected for multiple comparisons by thresholding the output of the (two-tailed) t-test at $p<0.01$ and clustering adjacent significant samples in the time-frequency space. The statistic used to describe these clusters is T_{sum}, the sum of all t-values inside the cluster. A permutation-test is performed by randomly re-assigning the ACI of each individual between the two conditions (5,000 iterations in our case) to obtain an estimate of the distribution of T_{sum} under the null hypothesis. It is then possible to compare the experimental value of T_{sum} with this calculated distribution to decide whether to reject the null hypothesis given a specified alpha value.

Because this procedure applies only to comparisons between conditions, the significance of weights in one ACI was corrected using a simple false discovery rate (FDR) correction. As a cluster-based non-parametric test, this statistical technique has been widely used for addressing the problem of multiple comparisons in neuroimaging studies [49]. More precisely, in this case, a t-value was calculated for each pixel corresponding to the hypothesis that the corresponding weight is significantly different from zero across participants. This result was then corrected by keeping the probability of type I error below a threshold of FDR<.01.

Results

Behavioral results

Due to the extreme length of the experiment, particular care was taken to ensure comfortable listening at all times for the participants. They reported no perceived effect of excessive mental fatigue over the course of the experiment, although some participants admitted experiencing occasional and brief attention loss. As expected, participants obtained a mean correct response rate of 78.8%, as determined by the adaptive SNR algorithm. Moreover, it was possible to determine their individual performances by observing the SNR levels (given in Fig. 3A.): except for one low performer, the SNR distributions of all participants were approximately -12 dB (mean = -11.8 ± 0.9 dB), although the individual variations were quite large, with standard deviations ranging from 1.32 to 2.47 dB. Participant 24 did not achieve a stable 79% point and was therefore excluded from analysis, as noted earlier. To characterize the participants' performances more precisely, we estimated their individual psychometric functions (the rate of correct responses as a function of the SNR) by fitting a cumulative normal distribution of unknown threshold and slope [50,51]. The results are shown in Fig. 3B. The similarity between psychometric functions suggests that all participants included in the analysis formed a homogeneous group of listeners, at least in terms of the SNR level at which they correctly categorize 79% of all stimuli.

A slight but significant bias of all participants toward response 'ga' can be observed from the mean rate of 'ga' responses (Fig. 3C, $p<10^{-4}$). An analysis of variance (ANOVA) was conducted, with proportions of 'ga' responses as the dependent variable and context (Al- or Ar-) and the SNR level (high or low SNR) as the within-subject factors, thus revealing a significant main effect of context ($F(1,15) = 78.65$, $p<10^{-10}$): the proportions of 'ga' responses were higher in context Ar- (61.2%) than in context Al- (50.3%). Similarly, we obtained a below-significance trend of SNR level ($F(1,15) = 2.97$, $p = 0.09$), with a low SNR generating a lower bias (54.7%)

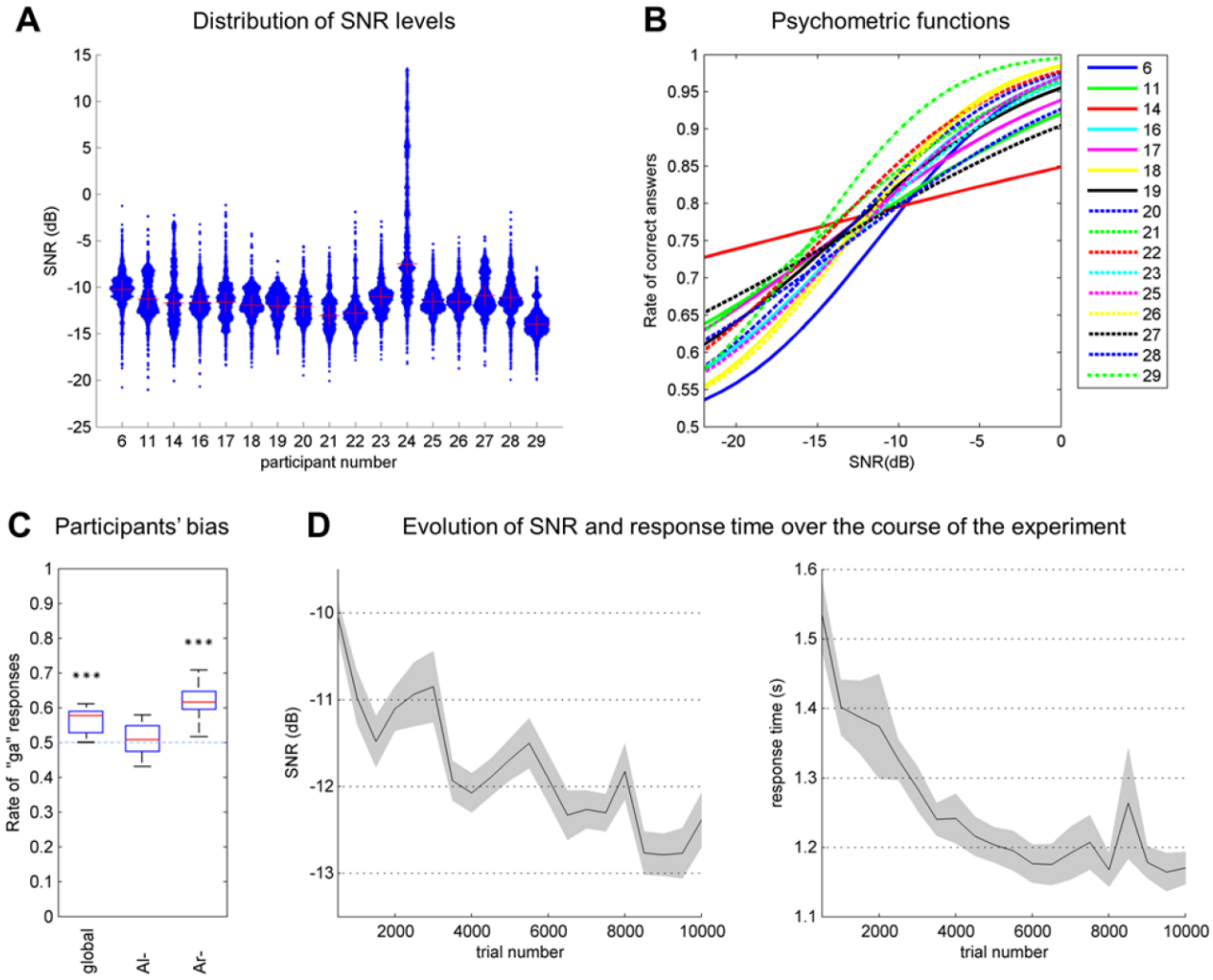


Fig 3. Overview of performance across all participants. A. Distribution of SNR levels for all participants (N = 17), with the mean SNR indicated with crosses. The width of each histogram is proportional to the number of trials ran at the corresponding SNR. B. Psychometric functions for participants included in the analysis (N = 16). C. Participants' bias towards response 'ga' (N = 16) over the entirety of the experiment ("global") and in conditions Al- and Ar-. Conditions with a significant bias (i.e., rate of 'ga' responses differing significantly from the level of 0.5 based on a random distribution, blue line) are indicated by asterisks. D. Evolution of the SNR and response time over the course of the experiment for sessions of 500 trials. Mean for the participants collectively (N = 16) with s.e.m.

doi:10.1371/journal.pone.0118009.g003

compared with a high SNR (56.8%). There was no significant interaction effect between these two factors ($p = 0.62$). The bias toward response 'ga' is linked to the participants' scores, with a higher percentage of correct answers linked to stimulus 'Arga' (89.4%) compared with stimulus 'Arda' (67.2%), whereas the percentages are extremely similar between stimuli 'Alda' (79.1%) and 'Alga' (79.4%).

Additionally, the characteristics and distributions of responses are not time-stationary but evolved over the course of the experiment. Thus, a clear progressive facilitation effect was observed in terms of both the reaction time (decreasing from 1.53 s in the first session to 1.17 s in the final session, $p<10^{-5}$) and SNR level (from -10.0 to -12.4 dB, $p<10^{-5}$) (see Fig. 3D). Thus, at the end of the experiment, each listener was performing the task more rapidly and more efficiently. Similarly, the mean bias tends to disappear over the course of the experiment, from

60.2% of 'ga' responses in the first session to 50.6% in the final session ($p<10^{-6}$). However, this effect can be considered a direct consequence of the decreasing SNR, as low SNR levels have been shown to be associated with lower biases.

Obtained auditory classification images

Seven ACIs were derived for each of the 16 participants: in addition to the "global" statistical model that considered all responses from one participant (10,000 trials), we estimated the model parameters from different subsets of the data (each of 5,000 trials) to attempt to disentangle the effects of several factors on the ACI. Six conditions were defined according to the context (target beginning with Al- or Ar-), the trial number (the first 5,000 trials or the last 5,000 trials), and the SNR (the 5,000 highest SNRs or the 5,000 lowest SNRs). In the "global" condition, one individual hyperparameter $\tilde{\lambda}$ was selected to fit the model parameters ($\underline{b}$ and the ACI $\underline{\beta}$). For proper averaging of the ACIs of multiple participants, we also selected a conjoint hyperparameter ($\tilde{\lambda}$ = 1,191), as explained in the Materials and Methods section. These values and goodness of fit (the minimum of the CVD curve) are systematically reported in the corresponding figures. To enable comparison between participants and conditions, in each ACI the weights are divided by their maximum absolute value.

The "global" ACIs for each participant are shown in Fig. 4A. As expected, there were slight differences in smoothness due to the variation of the regularization parameter,$\tilde{\lambda}$. The difference in terms of contrast is also notable, with some ACIs exhibiting a large number of maxima (as for participant #6), whereas others appear to be more focused (e.g., participant #19). Nevertheless, all participants exhibit a similar pattern of weights in a small region of times ranging from 300 to 470 ms and frequencies ranging from 1,300 to 2,800 Hz. This pattern becomes clearer for the mean ACI over all participants (Fig. 4B.). A statistical analysis revealed that the seven most distinct acoustic cues were all composed of positive or negative weights significantly different from zero (corrected t-test, FDR = 0.01). The significant weights are shown in Fig. 4C.

To further explore this result, we dissociated the effects of the context by estimating the model parameters separately on the 5,000 responses to targets beginning with /al/ ('Alda' or 'Alga') and on the 5,000 responses to targets beginning with /aʁ/ ('Arda' or 'Arga'), with the same value of $\tilde{\lambda}$ = 1,191 as before. Differences between the two resulting ACIs are considered to reflect the non-linearities of the auditory system [52]. The "signal-specific" ACIs are shown in Fig. 5A. Notably, the distribution of weights differs slightly between both conditions.

A cluster-based nonparametric test was performed on the difference between the two ACIs to confirm this result (see Fig. 5B a paired t-test with cluster based correction, $p<0.05$). Indeed, one cluster appears to be significant, corresponding to a difference in the weighting of the main positive cue ($p = 0.045$, Tsum = 139.9).

Conversely, a similar comparison between the first 5,000 trials (condition "firsttrials") and the last 5,000 trials (condition "lasttrials") elicited no significant difference (Fig. 5, $p>0.3$, $|Tsum| = 50.2$). No differences were found between the 5,000 trials with the highest SNR (condition "highSNR") and the 5,000 trials with the lowest SNR (condition "lowSNR") (Fig. 5, $p>0.15$, $|Tsum|<77.7$).

Discussion

In the present experiment, we used a psychoacoustic imaging method to isolate acoustic cues from the natural stimuli in a speech-in-noise categorization task. Participants were asked to perform 10,000 classifications in the presence of Gaussian noise. During each trial, they answered whether they heard /da/ or /ga/, independently of the preceding context. The accuracy

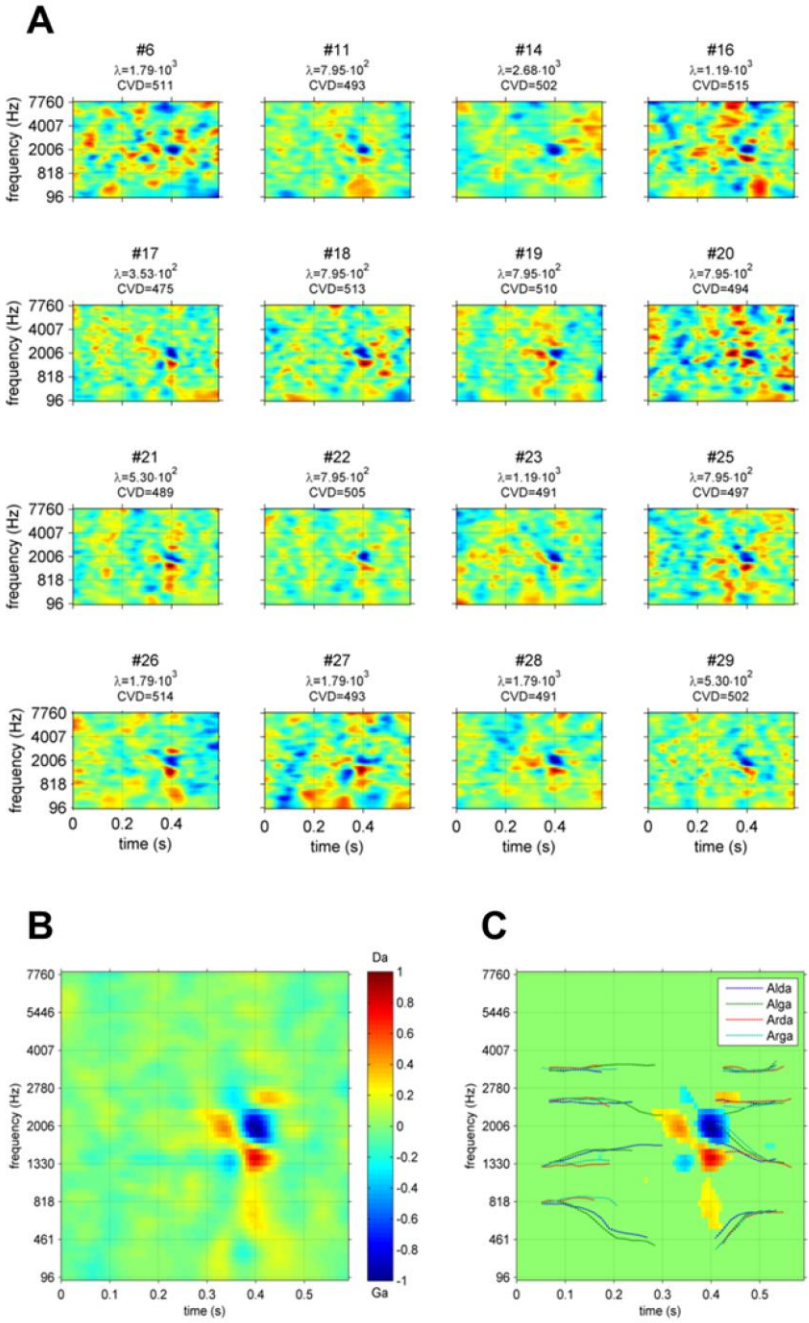


Fig 4. Global ACIs. A. Individual ACI estimated on 10,000 trials for all 16 participants. $\tilde{\lambda}$ and CVD are indicated below each image. B. Mean ACI for the participants collectively (estimated with $\lambda = 1{,}191$). C. Same ACI, with non-significant weights set to 0 (corrected t-test, FDR<0.01) and formant trajectories superimposed. In each ACI, weights are divided by their maximum absolute value.

doi:10.1371/journal.pone.0118009.g004

rate of 78.8% for 16 participants for a SNR range of approximately -11.8 dB and their similar psychometric functions confirmed that they all successfully performed the task. Moreover, all participants included in the study demonstrated a performance improvement over time in terms of both the SNR and response time. Finally, when dissociating the participants' answers

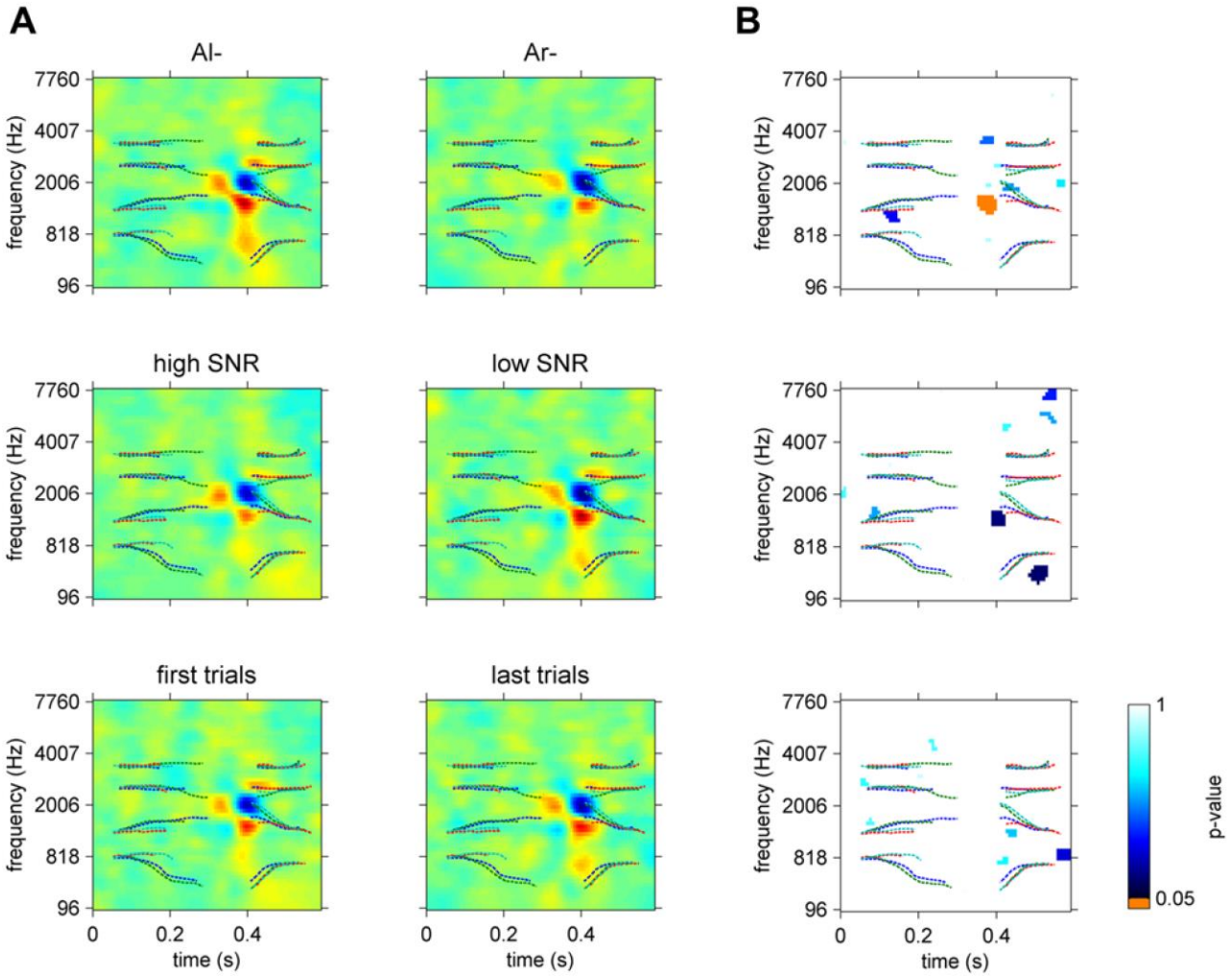


Fig 5. ACIs estimated on subsets of the data. A. Mean ACI for the 16 participants collectively, for conditions Al-, Ar-, high SNR, low SNR, firsttrials and lasttrials. Each individual ACI was estimated on 5,000 trials, with $\bar{\lambda} = 1.191$. In each ACI, the weights are divided by their maximum absolute value, and the formant trajectories are superimposed (same legend as in Fig. 6). B. Position of significant (orange) and non-significant clusters for each comparison (cluster-based non parametric test, $p<0.05$).

doi:10.1371/journal.pone.0118009.g005

with respect to the presented stimulus, the rate of "ga" responses was higher in context "Ar" than in context "Al". This result may seem contradictory to that of Mann [7]. Using a continuum of synthetic "da" and "ga" varying only in the height of F3 onset, preceded by a synthetic context "Al" or "Ar", she demonstrated that participants were more likely to answer "ga" in context "Al" and "da" in context "Ar." This effect was interpreted as direct evidence of "compensation for coarticulation" and was reproduced in several studies. However, in the present experiment using natural stimuli, our particular utterance of "Arga" may simply be produced more distinctly and may therefore be more robust to noise than "Arda", as suggested by the lower percentage of correct answers for the latter. This difference would account for a lower rate of "ga" responses in context "Al". Nevertheless, this slight response bias was not an issue, as a sufficient number of responses were obtained for both types for the ACI estimation.

The calculation of ACI at the group level exhibited well-defined clusters of weights on the onsets of the F2 and F3 transitions. As has been suggested in previous studies [7,53,54], the

main acoustic cues involved in this categorization task are the onsets of these two formants. Here, when there is a large amount of noise in the central negative cluster (approximately 0.4s and 2,000 Hz) corresponding to the junction between the two formants in the syllable 'ga,' the F2 and F3 onsets are perceived as closer than they actually are, and the target is more likely to be categorized as 'ga.' Conversely, when the noise is mainly distributed above or below this cluster, the target is more likely to be categorized as 'da.' This result also confirmed that participants were categorizing stimuli as speech sounds, not by relying on non-phonetic cues, such as prosody or intonation. As the auditory system detects variations in acoustic energy rather than absolute values, all 3 "main" acoustic cues are preceded by an inverse, smaller cue lasting approximately 0.35s, thus demonstrating an effect of temporal masking: perception of stimulus energy in a cochlear band is relative to energy at the previous time instant.

One objective of this study was to examine the specificity of an ACI to the particular utterances used in the categorization: do the positions and weightings of the acoustic cues depend on the production of speech used as targets? This question was not answered in the previous experiment involving only one recording each of "aba" and "ada." In the present experiment, we used two productions of each target phoneme instead. To ensure that these two utterances of the same phoneme were acoustically different in a predictable manner, they were produced and presented in a situation of coarticulation, i.e., preceded by two different contexts /al/ and /aʁ/. Indeed, the production of a stop consonant is influenced by the position of the preceding context. As evidenced by the cochleograms of the 4 stimuli (Fig. 1), the two couples of allophones, although sharing a similar pattern, exhibit slight differences in the relative power and precise position of their formants (e.g., relative onset times between the two 'da,' F3 onset frequency between the two 'ga'). Additionally, the perception of a phoneme can be biased by the preceding context [7,55,56]. One question that arose was the following: are those differences reflected in the ACIs? When splitting the ACIs according to the first syllable, we could reveal significant differences between the ACI in context Al- and in context Ar-. These differences are typically interpreted as nonlinearities in the auditory/speech perception system, with the processing applied to the input signal depending on the signal [52]. More specifically, the significant cluster corresponds to a difference in the weighting of the main cues: in context Ar-, participants relied less heavily on the main positive cue. The cause of this dissimilarity could not be determined with certainty because our comparison involves differences in both the targets and their contexts. A possible explanation may be that this imbalance between the positive cue and adjacent negative cue in context Ar- could correspond to a mechanism of compensation for coarticulation, as both Ar- contexts have F2 and F3 at the frequencies corresponding to those of the positive cues. Thus, the participants could perceptually mask the positive cues at the same frequency, compared with the central negative cue, by a simple spectral contrast effect [9,57].

Two other conditions were tested in addition to the Al- /Ar- contexts. The absence of significant differences between the ACIs calculated on the first and last 5,000 trials (conditions "first-trials" and "lasttrials") suggests that participants' performance improvement over the course of the experiment did not rely on a modification of the listening strategy. A possible alternative mechanism would be a diminution of internal noise: listeners are more likely to provide the same answer when presented with the same stimulus twice at the end of the experiment than at the beginning. Unfortunately, the estimation of internal noise requires a two-pass experiment [58,59], which was impossible to implement here given the large number of trials.

Finally, the non-significant comparison between conditions highSNR and lowSNR suggests that the listening strategy did not depend crucially on the level of noise during the experiment. Rather, it may rely on the same acoustic cues, regardless of whether the background noise was important. Across a series of studies, Allen and colleagues carefully studied the confusion

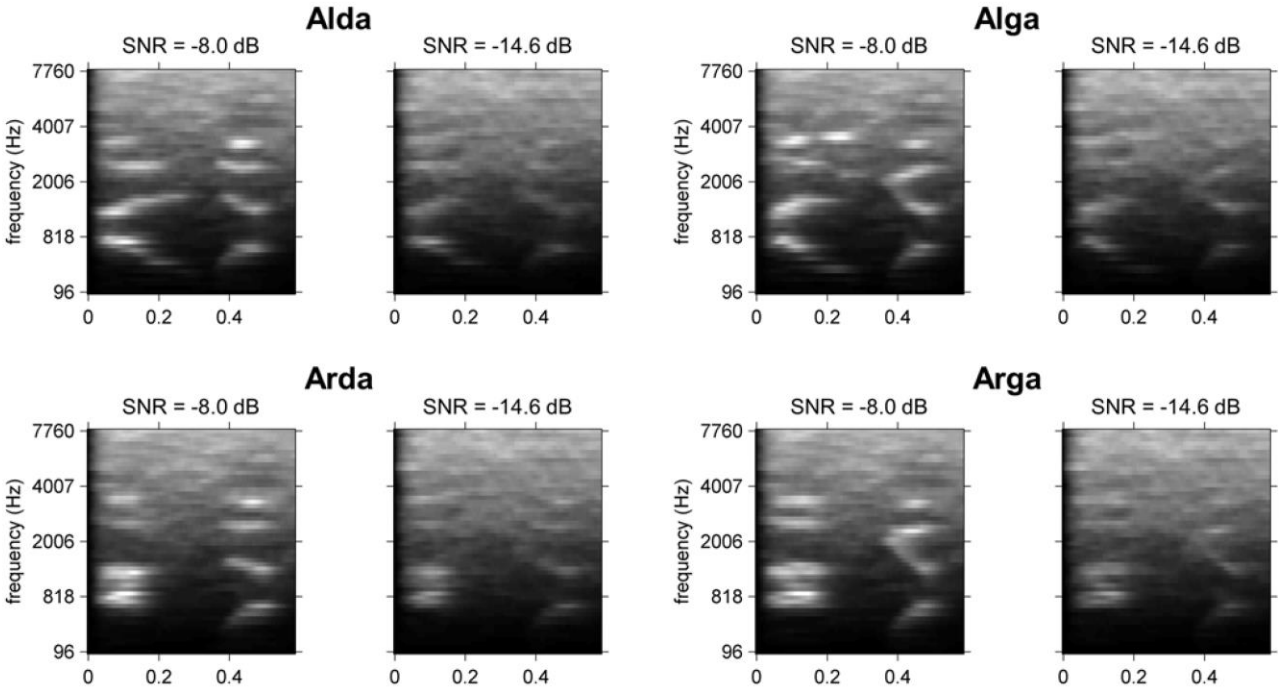


Fig 6. Cochleograms of the 4 signals in noise, with SNR = -8.0 dB or -14.6 dB. The parameters used for the cochleograms are described in paragraph Materials and Methods.

doi:10.1371/journal.pone.0118009.g006

patterns in a phoneme-recognition task as a function of SNR while linking discontinuities in the probabilities of a given answer and the robustness of the critical acoustic cues [17–19]. It may appear surprising that similar noise-dependent cues were not observed in our study. However, in our case, the range of SNR values was considerably smaller: overall, 90% of the trials were between -8.0 and -14.6 dB, whereas in the experiment conducted by Allen and colleagues, the SNR value varied from 12 to -22 dB. One may assume that no critical acoustic cues are masked in our lowSNR condition compared with the highSNR condition, as confirmed by the representations of the four signals in noise at -8.0 and -14.6 dB (Fig. 6).

In the next sections, we discuss the assumptions underlying the ACI method and possible improvements.

Cochleogram representation

The use of a GLM does not require the noise samples in $\underline{X}_i$ to be normally distributed, thus allowing us to select from most auditory models to represent the sounds. In their 2013 article, Varnet and colleagues chose to derive their ACIs from the spectrogram of the noise. However, the spectrogram is not the most suitable representation for studying speech perception because it does not consider the specificities of processing in the outer and middle ear, such as the spacing and bandwidth of the auditory filters. Thus, we decided to use a more biologically inspired representation of speech, the cochleogram, thereby yielding a "higher-level" representation of the functional acoustic cues. Because of the quasi-logarithmic frequency axis, mimicking the resolution of the auditory system, the acoustic cues in different frequency bands should be similar in size. This similarity is important for applying the smoothness prior, which acts here as a low-pass filter. Indeed, spatial smoothing would make it impossible to detect cues of large and small sizes simultaneously, as will be discussed further below.

Nonetheless, other representations could be used. A promising approach would be to combine the ACI technique with the multiresolution model developed by Chi and colleagues [25]. This combination would yield a 3-dimensional matrix of weights reflecting the importance of different regions of the time-frequency space for the phonemic categorization.

Are smoothness priors well adapted?

The introduction of a smoothness prior in the GLM estimation provides a means of selecting the amount of filtering to be applied to the ACI by minimizing the error predicting new data rather than applying an arbitrary degree of smoothing. This powerful tool is highly useful when estimating a matrix of parameters with unknown smoothness from a series of noisy measurements. However, the determination of the optimal spatial smoothing in a principled manner is not immune to other filtering problems, such as those described in [60]. One limitation is that if patterns of multiple scales are present, then the filtering can make the detection of both patterns simultaneously nearly impossible. In other words, our smoothness optimization introduces one assumption in the estimation process: that all relevant acoustic cues must be of similar sizes. However, the bandwidth of the auditory filters varies with their center frequencies. The cochleogram representation considers, at least partially, this differential sensitivity along the basilar membrane. Nevertheless, acoustic cues covering several auditory filters may have different sizes.

Indeed, when dividing our frequency axis into three bands with equal numbers of parameters (low frequencies: 90–1,100 Hz; middle frequencies: 1,100–3,100 Hz; high frequencies: 3,100–8,000 Hz) and estimating three separate ACIs for each participant, we obtained different degrees of smoothing for the three frequency bands. The same acoustic cues were obtained in the middle frequency band, and no significant weight was found in the high-frequency band. Unexpectedly, a clear acoustic cue appeared in the low-frequency band, with a much lower degree of smoothing different than in the middle frequency band (Fig. 7). This small-sized low-frequency cue was not predicted by the previous studies on this task, as they focused solely on the F2-F3 transition. Thus, our band-limited ACI indicates that this simple categorization task involves the processing of several spectral and temporal cues. One possible interpretation may be found in [13]. This synthetic speech study suggests that the identity of the consonant may be affected by the synchronicity between the F1 onset and the locus of the transition. Thus, a temporal translation of the first formant might change the consonant percept, thereby explaining the presence of a temporal cue on the onset of the first formant in our ACI. This low-frequency cue was not detected during the "global" ACI estimation because the middle-frequency cues, which are of different size, more accurately predict the participants' responses. Therefore, the CVD curve attains its minimum for the lambda value corresponding to the smoothness of the main cues, a value that is too high to render a good resolution of the secondary cues.

The presence of multiple resolutions clearly shows a limitation of the smoothness prior: cues of multiple sizes cannot be found simultaneously in a single estimate. One solution in our case would be to implement the constraint on not the smoothness but the number of cues to be detected. This adjustment could be enabled by the "sparse prior on a smooth basis" described in the work by Mineault et al. [39]. Using the same GLM, this penalization would seek to improve the accuracy of the prediction of the participants' answers by placing a restricted number of Gaussian-shaped patterns of weights of various scales on the ACI. Moreover, in their visual experiment, Mineault and colleagues demonstrated that the sparse prior offers a more accurate prediction than the smoothness prior for a given number of trials in terms of CV deviance.

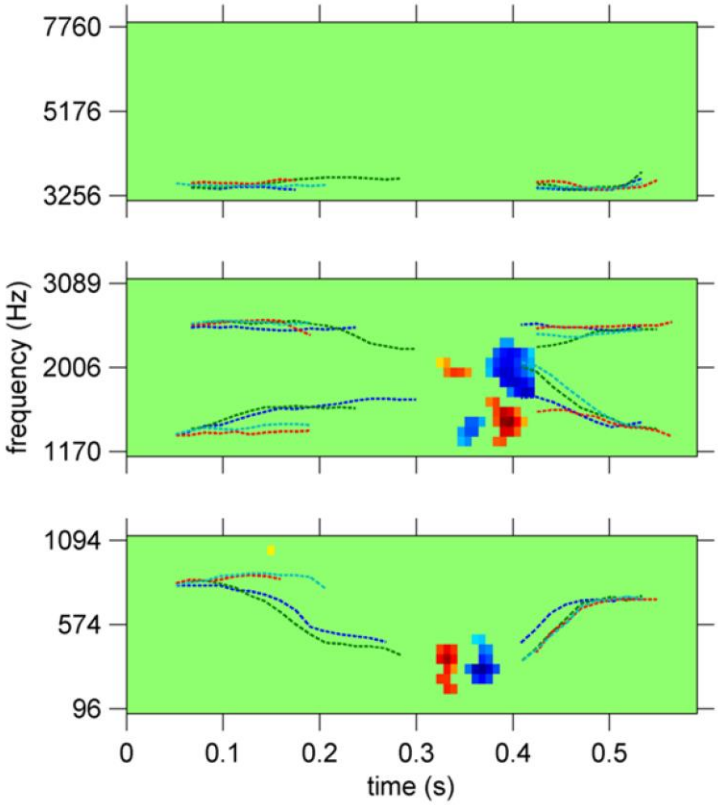


Fig 7. Band-limited ACIs, in low-frequency (90–1,100 Hz, $\tilde{\lambda}$ = 36, middle frequency (1,100–3,100 Hz, $\tilde{\lambda}$ = 144) and high frequency (3,100–8,000 Hz, $\tilde{\lambda}$ = 144) bands. In each band, weights are divided by their maximum absolute value and formant trajectories are superimposed (same legend as in Fig. 6). Non-significant weights are set to 0 (corrected t-test, FDR<0.01).

doi:10.1371/journal.pone.0118009.g007

Number of trials required

One crucial question for the application of the ACI method relates to the length of the experiment. We examined separately how the number of trials influences the hyperparameter selection and the quality of the estimated templates.

Number of trials required for lambda optimization. Fig. 8A. depicts the optimal lambda values obtained with different numbers of trials (in red). By definition, $\tilde{\lambda}$, the point at which the CVD becomes smaller than the minimum plus one standard deviation, is smaller than the lambda value corresponding to the minimum of the CVD curve (in blue). Both points appear to be biased toward higher values when the number of trials is insufficient to provide a reliable ACI (approximately with less than 5,000 trials). Indeed, in this case, the CVD curve does not attain a minimum but plateaus after an abrupt decrease. The participants' response bias toward 'ga,' a bias that has been shown to be stronger in the first sessions of this experiment, may also affect the overestimation of $\tilde{\lambda}$. Nevertheless, the lambda selection appears to be relatively robust, even with as few as 1,000 trials. Comparatively, a selection based on the minimum of the CVD curve would perform less well in terms of both the bias and variability across participants.

Number of trials required for reliable ACI estimation. As a second step, we investigated the number of trials necessary to ensure a reliable estimation of the underlying template. The

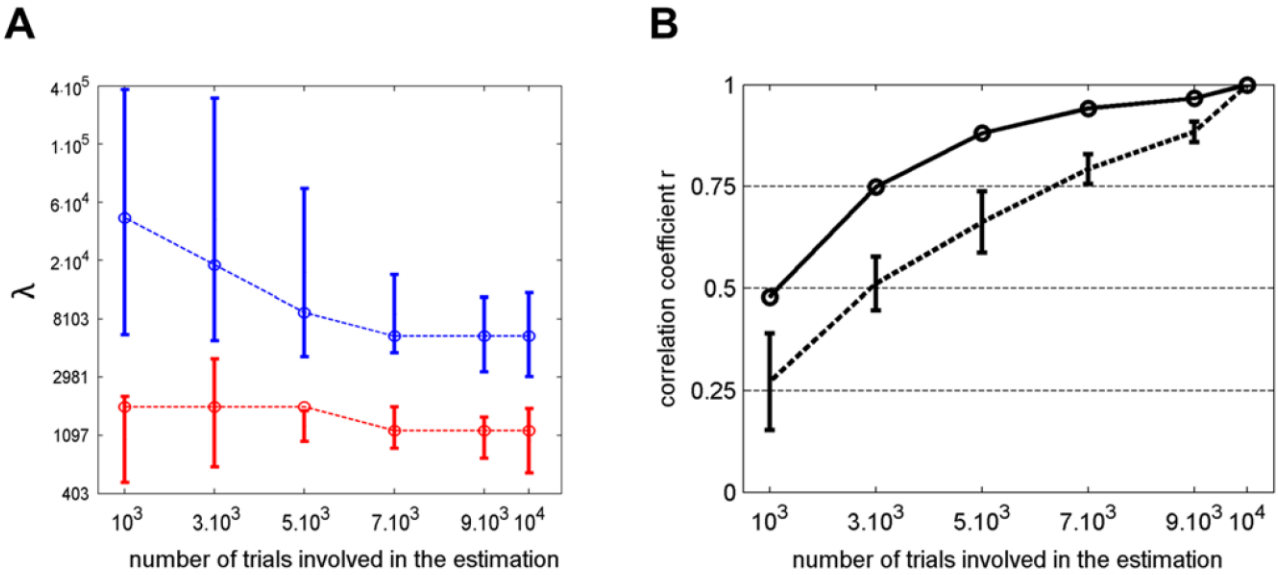


Fig 8. Effect of the number of trials involved in the estimation. A. Evolution of $\tilde{\lambda}$ for a different number of trials. Red circles indicate the result of the conjoint optimization, and red error bars indicate the standard deviation of the individual lambdas. Blue circles and blue error bars indicate the position of the minimum value of the CVD function for groups and individuals, respectively. B. Correlation between ACIs estimated on n trials, and final ACI (λ = 1,191). Dotted line: mean and standard deviation of the correlation for individual ACIs; thick line: correlation for the mean ACI for the participants collectively.

doi:10.1371/journal.pone.0118009.g008

accuracy of one ACI was evaluated by examining its correlation with the final ACI calculated on 10,000 trials. All ACIs were estimated with λ = 1,191 ($\tilde{\lambda}$ for 10,000 trials). The results of the individual and mean ACIs are presented in Fig. 8B. Whereas the accuracy of individual estimates decreases steadily with a decreasing number of trials involved in the estimation, the mean ACI for all 16 participants in total remains high (r>0.75) until approximately 3,000 trials (the estimation noise being reduced by the averaging).

Overall, the data collected from 16 participants enable the number of trials required from each participant to be reduced to approximately 3,000 by selecting the lambda value conjointly and considering the mean image for all participants. More importantly, this multi-participants study offers the opportunity to apply statistical tests at the group level rather than at the individual level.

Future directions. Here we have described a new methodology to investigate the way in which the human speech perception system achieves fast and efficient categorization of phonemes in noise. An appealing application would be to combine the ACI approach with electrophysiological measurements, such as EEG recordings or intracranial recordings. This would offer a direct way to identify the neural correlates of acoustic cue detection during speech perception. Furthermore the similarities with statistical methods employed in time-frequency analyses of electrophysiological data [22,48] would make it possible to draw parallel analyses of neural and behavioral responses. For the time being however, the duration of the experiment would constitute a major impediment.

Two plausible solutions to overcome this problem should be considered in future studies. At present we can obtain a good level of precision for individual images using 5000 trials, as mentioned above. A solution to further reduce this number of trials would be to introduce some additional a priori knowledge about the acoustic cues to be sought. For example, if we assume that the cues could be well represented by a limited number of Gaussian bumps we can use a GLM with sparse priors on a smooth basis, which is far more powerful, as done by

Mineault et al. [39]. Alternatively, future studies investigating the neural signatures of speech categorization using the ACI approach could explore the recording of Speech Auditory Brainstem Response (ABR) [61,62]. This type of experiment typically requires a few thousand presentations of speech stimuli. In this context, one could derive the ACI directly from the ABR instead of the behavioral response of the participant.

Conclusion

We demonstrated how the GLM with smoothness prior approach, combined with a cluster-based test, provides a reliable approach for investigating the acoustic cues involved in a specific phoneme categorization task. Through the example of a da/ga categorization in the contexts of Ar- and Al-, we confirmed that listeners relied on the F2 and F3 onsets. We also demonstrated that the perceived timing of F1influences the categorization. Finally, the method was proven precise enough to track fine modifications in the weighting of the different cues depending on the specific utterance presented. Three constraints of the ACI technique and possible solutions were discussed: the dependency on the sound representation, the choice of the prior, and the number of trials required. Despite these limitations, such a psychoacoustic method, which involves no prior knowledge of the spectrotemporal locations of the acoustic cues being sought, offers a valuable insight into the mechanisms of speech perception. Additionally, the ACI technique can be combined with statistical tests at a group level, thus making it a powerful tool to investigate hypotheses on human speech recognition.

Supporting Information

S1 Table. Results of preliminary screening tests for all participants.
(XLSX)

Author Contributions

Conceived and designed the experiments: LV MH. Performed the experiments: LV. Analyzed the data: LV MH. Contributed reagents/materials/analysis tools: LV MH FM KK. Wrote the paper: LV MH FM KK WS.

References

1. Norris D, McQueen JM, Cutler A (2000) Merging information in speech recognition: feedback is never necessary. Behav Brain Sci 23: 299–325; discussion 325–370. PMID: 11301575.
2. Segui J, Frauenfelder U, Mehler J (1981) Phoneme monitoring, syllable monitoring and lexical access. Br J Psychol 72: 471–477. doi: 10.1111/j.2044-8295.1981.tb01776.x
3. Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychol Rev 74: 431–461. PMID: 4170865
4. Moore BC., Tyler LK, Marslen-Wilson W (2008) Introduction. The perception of speech: from sound to meaning. Philos Trans R Soc B Biol Sci 363: 917–921. doi: 10.1098/rstb.2007.2195 PMID: 17827100.
5. Fant G (1970) Acoustic Theory of Speech Production. Walter de Gruyter. 344 p.
6. Johnson K (2011) Acoustic and Auditory Phonetics. John Wiley & Sons. 235 p.
7. Mann VA (1980) Influence of preceding liquid on stop-consonant perception. Percept Psychophys 28: 407–412. PMID: 7208250
8. Fowler CA (2006) Compensation for coarticulation reflects gesture perception, not spectral contrast. Percept Psychophys 68: 161–177. PMID: 16773890
9. Lotto AJ, Kluender KR (1998) General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. Percept Psychophys 60: 602–619. PMID: 9628993
10. Sonderegger M, Yu A (2010) A rational account of perceptual compensation for coarticulation. Proc 32nd Annu Meet Cogn Sci Soc. Available: http://palm.mindmodeling.org/cogsci2010/papers/0063/paper0063.pdf. Accessed 30 October 2013.

11. Viswanathan N, Fowler CA, Magnuson JS (2009) A critical examination of the spectral contrast account of compensation for coarticulation. Psychon Bull Rev 16: 74–79. doi: 10.3758/PBR.16.1.74 PMID: 19145013
12. Juneja A (2012) A comparison of automatic and human speech recognition in null grammar. J Acoust Soc Am 131: EL256–EL261. doi: 10.1121/1.3684744
13. Delattre PC, Liberman AM, Cooper FS (1955) Acoustic Loci and Transitional Cues for Consonants. J Acoust Soc Am 27: 769–773. doi: 10.1121/1.1908024
14. Liberman AM, Delattre PC, Cooper FS, Gerstman LJ (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol Monogr Gen Appl 68: 1–13. doi: 10.1037/h0093673
15. Apoux F, Healy EW (2009) On the number of auditory filter outputs needed to understand speech: further evidence for auditory channel independence. Hear Res 255: 99–108. doi: 10.1016/j.heares.2009.06.005 PMID: 19539016
16. Xu L, Thompson CS, Pfingst BE (2005) Relative contributions of spectral and temporal cues for phoneme recognition. J Acoust Soc Am 117: 3255–3267. doi: 10.1121/1.1886405 PMID: 15957791
17. Régnier MS, Allen JB (2008) A method to identify noise-robust perceptual features: application for consonant /t/. J Acoust Soc Am 123: 2801–2814. doi: 10.1121/1.2897915 PMID: 18529196
18. Li F, Menon A, Allen JB (2010) A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. J Acoust Soc Am 127: 2599–2610. doi: 10.1121/1.3295689 PMID: 20370041
19. Li F, Trevino A, Menon A, Allen JB (2012) A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise. J Acoust Soc Am 132: 2663–2675. doi: 10.1121/1.4747008 PMID: 23039459
20. Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic Feature Encoding in Human Superior Temporal Gyrus. Science: 1245994. doi: 10.1126/science.1245994
21. Obleser J, Eisner F (2009) Pre-lexical abstraction of speech in the auditory cortex. Trends Cogn Sci 13: 14–19. doi: 10.1016/j.tics.2008.09.005 PMID: 19070534
22. Calabrese A, Schumacher JW, Schneider DM, Paninski L, Woolley SMN (2011) A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. PloS One 6: e16104. doi: 10.1371/journal.pone.0016104 PMID: 21264310
23. Woolley SMN, Gill PR, Theunissen FE (2006) Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. J Neurosci Off J Soc Neurosci 26: 2499–2512. doi: 10.1523/JNEUROSCI.3731-05.2006
24. Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485: 233–236. doi: 10.1038/nature11020 PMID: 22522927
25. Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am 118: 887–906. PMID: 16158645
26. Varnet L, Knoblauch K, Meunier F, Hoen M (2013) Using auditory classification images for the identification of fine acoustic cues used in speech perception. Front Hum Neurosci 7: 865. doi: 10.3389/fnhum.2013.00865 PMID: 24379774
27. Al Ahumada J, Lovell J (1971) Stimulus Features in Signal Detection. J Acoust Soc Am 49: 1751–1756. doi: 10.1121/1.1912577
28. Ahumada AJ Jr (1996) Perceptual classification images from vernier acuity masked by noise. ECVP'96 Abstracts. doi: 10.1155/2013/627230 PMID: 24494432
29. Gold JM, Sekuler AB, Bennett PJ (2004) Characterizing perceptual learning with external noise. Cogn Sci 28: 167–207.
30. Kurki I, Eckstein MP (2014) Template changes with perceptual learning are driven by feature informativeness. J Vis 14: 6. doi: 10.1167/14.11.6 PMID: 25515764
31. Gold JM, Murray RF, Bennett PJ, Sekuler AB (2000) Deriving behavioural receptive fields for visually completed contours. Curr Biol CB 10: 663–666. PMID: 10837252
32. Thomas JP, Knoblauch K (2005) Frequency and phase contributions to the detection of temporal luminance modulation. J Opt Soc Am A Opt Image Sci Vis 22: 2257–2261. PMID: 16277294
33. Bouet R, Knoblauch K (2004) Perceptual classification of chromatic modulation. Vis Neurosci 21: 283–289. PMID: 15518201
34. Liu J, Li J, Feng L, Li L, Tian J, et al. (2014) Seeing Jesus in toast: Neural and behavioral correlates of face pareidolia. Cortex 53: 60–77. doi: 10.1016/j.cortex.2014.01.013 PMID: 24583223
35. Levitt H (1971) Transformed up-down methods in psychoacoustics. J Acoust Soc Am 49: Suppl 2:467 +. PMID: 5541744

36. Slaney M, Lyon RF (1988) Lyon's cochlear model. Apple Computer, Advanced Technology Group. 72 p.
37. Wu MC- K, David SV, Gallant JL (2006) Complete functional characterization of sensory neurons by system identification. Annu Rev Neurosci 29: 477–505. doi: 10.1146/annurev.neuro.29.051605.113024 PMID: 16776594
38. Knoblauch K, Maloney LT (2008) Estimating classification images with generalized linear and additive models. J Vis 8: 10.1–19. doi: 10.1167/8.16.10 PMID: 19146352
39. Mineault PJ, Barthelmé S, Pack CC (2009) Improved classification images with sparse priors in a smooth basis. J Vis 9: 17.1–24. doi: 10.1167/9.10.17 PMID: 20055550
40. Machens CK, Wehr MS, Zador AM (2004) Linearity of cortical receptive fields measured with natural sounds. J Neurosci Off J Soc Neurosci 24: 1089–1100. doi: 10.1523/JNEUROSCI.4445-03.2004 PMID: 22773137
41. Wood SN (2006) Generalized additive models: an introduction with R. Boca Raton, FL: Chapman & Hall/CRC. PMID: 23242683
42. Willmore B, Smyth D (2003) Methods for first-order kernel estimation: simple-cell receptive fields from responses to natural scenes. Netw Bristol Engl 14: 553–577.
43. Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33: 1–22. PMID: 20808728
44. Chauvin A, Worsley KJ, Schyns PG, Arguin M, Gosselin F (2005) Accurate statistical tests for smooth classification images. J Vis 5: 659–667. doi: 10:1167/5.9.1 PMID: 16356076
45. Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods 164: 177–190. doi: 10.1016/j.jneumeth.2007.03.024 PMID: 17517438
46. Nichols TE, Holmes AP (2003) Nonparametric Permutation Tests for Functional Neuroimaging. Human Brain Function. R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny.
47. Ménoret M, Varnet L, Fargier R, Cheylus A, Curie A, et al. (2014) Neural correlates of non-verbal social interactions: a dual-EEG study. Neuropsychologia 55: 85–97. doi: 10.1016/j.neuropsychologia.2013.10.001 PMID: 24157538
48. Oostenveld R, Fries P, Maris E, Schoffelen J- M (2011) FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci 2011: 156869. doi: 10.1155/2011/156869 PMID: 21253357
49. Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15: 870–878. doi: 10.1006/nimg.2001.1037 PMID: 11906227
50. Kingdom FAA, Prins N (2010) Psychophysics: a practical introduction. London: Academic.
51. Knoblauch K, Maloney LT (2012) Modeling Psychophysical Data in R. Springer Science & Business Media. 376 p.
52. Abbey CK, Eckstein MP (2006) Classification images for detection, contrast discrimination, and identification tasks with a common ideal observer. J Vis 6: 335–355. doi: 10.1167/6.4.4 PMID: 16889473
53. Holt LL (2006) Speech categorization in context: joint effects of nonspeech and speech precursors. J Acoust Soc Am 119: 4016–4026. PMID: 16838544
54. Viswanathan N, Magnuson JS, Fowler CA (2010) Compensation for coarticulation: disentangling auditory and gestural theories of perception of coarticulatory effects in speech. J Exp Psychol Hum Percept Perform 36: 1005–1015. doi: 10.1037/a0018391 PMID: 20695714
55. Mann VA, Repp BH (1981) Influence of preceding fricative on stop consonant perception. J Acoust Soc Am 69: 548–558. doi: 10.1121/1.385483 PMID: 7462477
56. Repp BH, Mann VA (1981) Perceptual assessment of fricative-stop coarticulation. J Acoust Soc Am 69: 1154–1163. PMID: 7229203
57. Holt LL, Lotto AJ (2002) Behavioral examinations of the level of auditory processing of speech context effects. Hear Res 167: 156–169. PMID: 12117538
58. Ahumada AJ Jr (2002) Classification image weights and internal noise level estimation. J Vis 2: 121–131. doi: 10:1167/2.1.8 PMID: 12678600
59. Joosten ERM, Neri P (2012) Human pitch detectors are tuned on a fine scale, but are perceptually accessed on a coarse scale. Biol Cybern 106: 465–482. doi: 10.1007/s00422-012-0510-x PMID: 22854977
60. Stelzer J, Lohmann G, Mueller K, Buschmann T, Turner R (2014) Deficient approaches to human neuroimaging. Front Hum Neurosci 8: 462. doi: 10.3389/fnhum.2014.00462 PMID: 25071503

61. Johnson KL, Nicol TG, Kraus N (2005) Brain stem response to speech: a biological marker of auditory processing. Ear Hear 26: 424–434. PMID: 16230893

62. Banai K, Nicol T, Zecker SG, Kraus N (2005) Brainstem timing: implications for cortical processing and literacy. J Neurosci Off J Soc Neurosci 25: 9850–9857. doi: 10.1523/JNEUROSCI.2373-05.2005 PMID: 22773137

10.3. Résumé de l'étude 2

L'étude 2 a tout d'abord été l'occasion de vérifier nos hypothèses concernant le rôle des F2 et F3 dans la catégorisation /da/-/ga/, au moyen de la méthode des ACIs. Comme suggéré par de précédentes expériences en parole synthétique, les stimuli possédant des fréquences d'attaque proches pour ces formants sont plus volontiers interprétés comme /ga/ ; inversement des attaques de formants éloignées engendrent le percept /da/. En conséquence, la présence de bruit dans cette région temps-fréquence influe sur la catégorisation, ce qui se traduit par des pondérations importantes de l'ACI. Une analyse limitée aux basses fréquences a également mis en évidence l'implication de l'attaque de F1 dans cette tâche.

L'observation d'indices non attendus illustre indirectement la grande précision qui est désormais possible dans l'estimation de l'ACI. Celle-ci repose sur deux modifications de l'algorithme : d'une part, une représentation biologiquement inspirée des indices acoustiques, basée sur le cochléogramme du bruit, et non plus sur le spectrogramme ; d'autre part, l'utilisation d'un hyperparamètre unique dans le GLM pénalisé, ce qui permet une sélection plus efficace du lissage optimal.

Enfin, cette étude démontre également que certains outils statistiques développés pour la neuro-imagerie (*FDR correction* et test non paramétrique basé sur les clusters) peuvent être utilisés en vue d'une analyse conjointe des ACIs d'un groupe de participants. À ce stade, nous disposons donc d'une procédure complète et d'un ensemble de scripts MATLAB (voir en Annexe 4) pour l'identification des indices acoustiques utilisés dans une tâche de catégorisation dans le bruit, leur validation statistique, ainsi que la comparaison des pondérations obtenues dans différentes conditions.

11. Étude 3 : Comparaison des stratégies d'écoute de participants musiciens et non musiciens dans une tâche de parole dans le bruit.

11.1. Présentation de l'étude 3

Les deux premières études ayant établi l'efficacité et la précision de la méthode des ACIs, nous avons alors cherché à l'appliquer dans un second temps dans le contexte d'un questionnement scientifique actuel. Nous nous sommes ainsi intéressés aux différences entre les stratégies d'écoute de plusieurs groupes d'auditeurs. La présente étude vise à comparer les ACIs de participants ayant reçu une formation instrumentale prolongée à celles de participants contrôles ne possédant pas ce bagage musical.

Un vif débat s'est engagé récemment dans la communauté scientifique afin de déterminer par quels mécanismes l'expertise particulière acquise par les musiciens dans des tâches auditives se transfère aux tâches parolières. À ce jour, la question de savoir si – et dans quelles conditions – les musiciens sont effectivement plus performants pour la compréhension de parole dans le bruit reste ouverte. Dans ce contexte, la méthode des ACIs semble offrir une nouvelle voie pour déterminer la stratégie d'écoute des musiciens pour la reconnaissance de stimuli de parole bruitée, en comparaison avec celle de non-musiciens. En particulier, les musiciens utilisent-ils des indices acoustiques différents ou bien, simplement, une pondération différente des mêmes indices acoustiques ?

Nous avons choisi ici de reprendre le paradigme expérimental utilisé dans l'étude 2 (catégorisation /da/-/ga/ en contexte /al/ ou /aʁ/), puisque les indices acoustiques impliqués nous sont déjà connus. Deux groupes de 20 participants furent soumis à cette tâche, l'un composé de musiciens experts, l'autre de non-musiciens, puis les ACIs furent calculées individuellement pour chaque auditeur. Des comparaisons de groupe furent ensuite réalisées pour identifier les différences significatives au niveau des pondérations de l'ACI. Cette analyse fut alors complétée par une étude de la généralisabilité du modèle, ajusté sur les données d'un participant, aux réponses d'un autre participant.

Ce manuscrit été publié dans la revue en accès ouvert Scientific Reports (Varnet et al., 2015b). Une version préliminaire de ces résultats a également fait l'objet d'une communication orale lors de la conférence Society for the Neurobiology of Language 2014 à Amsterdam (Wang, et al., 2014).

11.2. Article 3 : How musical expertise shapes speech perception: evidence from auditory classification images.

SCIENTIFIC REPORTS

OPEN

How musical expertise shapes speech perception: evidence from auditory classification images

Received: 24 June 2015
Accepted: 18 August 2015
Published: 24 September 2015

Léo Varnet[1,2,3], Tianyun Wang[1,2,3], Chloe Peter[1,3], Fanny Meunier[2,3] & Michel Hoen[1,3,4]

It is now well established that extensive musical training percolates to higher levels of cognition, such as speech processing. However, the lack of a precise technique to investigate the specific listening strategy involved in speech comprehension has made it difficult to determine how musicians' higher performance in non-speech tasks contributes to their enhanced speech comprehension. The recently developed Auditory Classification Image approach reveals the precise time-frequency regions used by participants when performing phonemic categorizations in noise. Here we used this technique on 19 non-musicians and 19 professional musicians. We found that both groups used very similar listening strategies, but the musicians relied more heavily on the two main acoustic cues, at the first formant onset and at the onsets of the second and third formants onsets. Additionally, they responded more consistently to stimuli. These observations provide a direct visualization of auditory plasticity resulting from extensive musical training and shed light on the level of functional transfer between auditory processing and speech perception.

Acoustically, music and speech are two broad classes of sounds that share a number of characteristics. Both consist of quasi-periodic segments of rich harmonic complexity (notes and syllables) separated by silence, bursts of noise and/or transients and are organized with precise timing. These perceptual elements are mostly defined in terms of their spectral envelope, duration, and fundamental frequency[1]. The spectral envelope is one main component of a musical sound's timbre (which allows identification of an instrument being played), whereas in speech its maxima, also called formants, are critical acoustic cues used to discriminate one vowel from another. In speech, as in music, pitch and rhythm are the two constituents of melody or prosody. Furthermore, both notes and vowels are perceived categorically, leading to a discretization of pitch space or formant space, respectively[2]. Finally, in real-world listening, sounds are rarely produced in isolation; however, the perception system demonstrates a remarkable ability to extract one auditory object from a background (e.g., the voice of your conversation partner in the presence of other speakers or one instrument in the orchestra)[3]. All of these commonalties suggest that similar basic auditory processes are involved during the analysis of both speech[4] and music[5].

Musical practice provides intensive training of these auditory capacities[6]. Musicians typically spend hours on tasks requiring the precise discrimination of sounds and melodies, leading to true auditory expertise. The most striking aspect of this sensory learning is the structural brain changes induced by instrumental training; musicians show increased grey matter volume in cortical areas of critical importance for music performance (auditory cortex, motor cortex, and visuo-spatial cortex)[7–9]. Such cortical reorganization can be connected to changes in cognitive abilities. Indeed, musicians have been demonstrated to outperform non-musicians in a number of non-speech auditory tasks. It is now well established that compared with non-musicians, musicians have larger auditory working memory[10], finer frequency[10–13]

[1]Lyon Neuroscience Research Center, CNRS UMR 5292, INSERM U1028, Auditory Language Processing (ALP) research group, Lyon, France. [2]Laboratoire sur le Langage le Cerveau et la Cognition, CNRS UMR 5304, Auditory Language Processing (ALP) research group, Lyon, France. [3]Université de Lyon, Université Lyon 1, Lyon, France. [4]Oticon Medical – 2720 Chemin Saint Bernard, 06220 Vallauris, France. Correspondence and requests for materials should be addressed to L.V. (email: leo.varnet@isc.cnrs.fr)

and duration discrimination[14], enhanced auditory attention[13], facilitated pitch processing[15,16], better detection of tones masked by noise[13] and greater perceptual acuity of rapid spectro-temporal changes[17]. This behavioral evidence for the benefit of musical experience on basic auditory skills has been supplemented with a series of electrophysiological studies. By directly measuring electromagnetic brain activity in response to various musical stimuli, researchers have been able to demonstrate increased sensitivity of musicians, compared with non-musicians, to mistuning[18–20] and spectral complexity[21,22].

The affinity between speech and music, described in the first paragraph, raises the question whether the improvement of basic auditory processes observed in musicians is domain-specific or can generalize beyond music processing to speech comprehension. There is general agreement that acuity enhancements observed in musical tasks percolate to higher-level skills. Musicians' expertise is associated with a wide number of cognitive benefits for processing speech sounds, including word segmentation[23], pitch processing in prosody[15,16], and integration of metric structure[24]. Furthermore, a large literature has analyzed speech auditory brainstem responses (sABRs)[25] in non-musicians and musicians, establishing that the latter show more robust encoding of speech sounds in the brainstem, even in situations of reverberation[26] or noisy background[27,28]. Musicians also exhibit an enhanced brainstem representation of pitch contours[29–31] and harmonics[27,32,33]. This superior encoding is significantly correlated with the amount of musical training[29,31,32] and occurs from an early age[34].

Taken together, these findings suggest that brain plasticity induced by musical expertise is not selective to musical stimuli but rather provides an advantage across domains. Notably, the effects of musical expertise transfer to speech processing and result in reinforced auditory skills and speech representations, leading to facilitated speech-in-noise perception. However, how musicians process their superior representations of speech to lead to better comprehension in noisy conditions remains unknown. In particular musical training could lead to a selective subcortical enhancement of relevant properties of speech sounds and eventually to a more focused auditory processing. Alternatively, musicians could exploit more efficient strategies involving different acoustic information than non-musicians.

In the present article, we asked whether musicians and non-musicians use the same acoustic cues during phoneme categorization in noise. This question was addressed here using a new auditory psychophysical method developed by some of the authors, which was named "Auditory Classification Image".

The Auditory Classification Image (ACI) is a powerful tool and originates from the work of Ahumada on tone-in-noise detection[35,36]. Informally, ACIs can be thought of as a visual description of the participant's listening strategy in an auditory categorization task. The experimental paradigm consists of introducing random fluctuations to the stimulus and then measuring the influence of these fluctuations on the participant's behavior. This approach reveals how noise masking of different "parts" of the sound biases the listener toward a specific response. Several statistical methods have been proposed for the derivation of Classification Images from visual categorization experiments, including linear regression[35,36], reverse correlation[37,38], and generalized linear models (GLMs)[39–41]. The most recent developments in the field involve penalized GLMs (also called generalized additive models)[39,42,43]. This last technique offers sufficient power to translate Classification Images back to the auditory domain. This step has been performed in our group through the example of /aba/-/ada/ categorization, yielding a map of the stimulus regions in which noise consistently affects participants' responses[44]. ACIs are a behavioral counterpart of spectro-temporal receptive fields (STRFs), a widely used model to capture the relationship between the acoustic characteristics of a stimulus and the firing of a specific auditory neuron[45,46]. At the population level, it has been demonstrated that the ACIs of several participants can be combined and analyzed using specific statistical tests from functional neuroimaging. This approach has been employed to unmask the auditory primitives used by normal-hearing participants to differentiate the syllables /da/ and /ga/[47]. In addition to the expected acoustic cues present in the F2 and F3 onsets, which have been previously determined through other methods, the authors were able to reveal a neglected source of information about the identity of the stimulus: the F1 onset.

The objectives pursued in the present study are twofold: 1) Determining whether musicians and non-musicians use different sets of acoustic cues when performing a /da/-/ga/ categorization in noise. 2) Estimating the specificity of each participant's listening strategy, in both musicians and non-musicians. These issues translate into two testable hypotheses: (H1) Particularities of musicians' listening strategies should be reflected in their ACIs. Therefore, there should be significant differences between musicians' ACIs and ACIs estimated for a group of non-musician participants. (H2) A measure of specificity can be found by evaluating how the ACI from one participant successfully predicts the responses of other participants. Individual characteristics in the ACIs would result in low generalizability across participants.

Methods

Participants. Forty adults without any known audiological or neurological pathology participated in this study. They provided written informed consent and were paid 100€ for participation. They were divided into two groups: 20 musical experts, who reported more than 7 years of musical practice and started musical training before the age of 13, and 20 control participants (normal readers who reported no musical practice). Seventeen control participants were individuals already included in a previous study[47]. A pre-test confirmed that all participants had audiometric pure-tone thresholds $\leq$20 dB over the 125 Hz–8000 Hz range for both ears.

Variable	Musicians	Non-musicians	t-test
Age (year)	23 (±2.89 S.D.)	22.68 (±4.39 S.D.)	p = 0.78
Gender (M/F)	9/5	6/13	
Handedness	64.21 (±53.99 S.D.)	73.95 (±57.72 S.D.)	p = 0.61
ANT			
Alerting effect	29.95 (±22.68 S.D.)	32.37 (±21.97 S.D.)	p = 0.78
Orienting effect	46.05 (±14.98 S.D.)	40.21 (±23.25 S.D.)	p = 0.36
Conflict effect	131.74 (±49.13 S.D.)	131.16 (±39.95 S.D.)	p = 0.97
Main experiment			
Score (%)	79.29 (±0.35 S.D.)	78.83 (±0.42 S.D.)	p = 0.0008*
SNR (dB)	−13.37 (±1.22 S.D.)	−11.91 (±1.04 S.D.)	p = 0.0004*
Reaction Time (s)	1.37 (±0.19 S.D.)	1.28 (±0.13 S.D.)	p = 0.11
Sensitivity d'	1.68 (±0.05 S.D.)	1.64 (±0.04 S.D.)	p = 0.0036*
Decision criterion	0.638 (±0.128 S.D.)	0.639 (±0.125 S.D.)	p = 0.97

Table 1. Summary of the characteristics of the two groups.

Two participants achieved unexpectedly low performances in the experiment, suggesting misunderstanding of task instructions, and their data were not included in further analyses. The two resulting groups ($N = 2 \times 19$) were matched with regard to age, gender, and handedness (independent t-tests, all $p > .05$). Attentional capacities were evaluated using the Attention Network Test (ANT)[48]. Groups did not differ significantly ($p > .05$) in Alerting, Orienting and Conflict effect scores. Table 1 shows the mean characteristics of the participants in each group.

Musician participants were additionally administered questionnaires on their musical practice and were tested for absolute pitch using a simple pitch-naming task without feedback. The procedure was automatized with notes played on a synthetic piano, randomly drawn from the full-range of the keyboard (from C1 to B6). Eight participants obtained a score of at least 17/20 on this task, and will be further considered as absolute pitch possessors. All results are reported in Table 2.

The study was approved by the Comité d'évaluation éthique de l'Inserm/Institutional Review Board (IORG0003254, FWA00005831, IRB00003888). All methods were carried out in accordance with the approved guidelines.

Stimuli. The targets were 4 productions of VCCV non-words (/alda/, /alga/, /aʁda/, and /aʁga/) used in a previous behavioral study[47]. They were recorded by a male speaker in a soundproof booth (48 kHz; 16 bits). Inter-syllable gap durations were reduced such that the time onsets of the 2nd syllable in all target sounds were the same (t = 328 ms), and they were then made equivalent with respect to duration (680 ms). The target sounds can be downloaded as .wav files at https://zenodo.org/record/12300/, and their cochleograms are shown in Fig. 1. For each participant, a set of 10,000 white-noise stimuli of the same length as the targets were generated and stored prior to the experiment. They can be found at the addresses https://zenodo.org/record/19104 and https://zenodo.org/record/23064.

On trial i, the stimulus presented consisted of one target signal, $\underline{t}_{k_i}$, that was embedded in noise, $\underline{n}_i$ (both root-mean-square normalized), with a given SNR_i, as shown in equation (1).

$$\underline{s}_i = (\alpha_i \cdot \underline{t}_{k_i} + \underline{n}_i)/A_i \tag{1}$$

where $\alpha_i = 10^{\frac{SNR_i}{20}}$ and $A_i = \sqrt{\mathrm{var}(\alpha_i \cdot \underline{t}_{k_i} + \underline{n}_i)}$, a factor allowing the power normalization of the stimulus. During presentation, all stimuli were preceded with a short Gaussian fade-in of Gaussian noise for the listener's comfort.

Experimental procedure. The stimuli were presented diotically over Sennheiser's HD 448 headphones at each listener's most comfortable sound pressure level. The participants were instructed to listen carefully to the speech sound and to press one of the two response keys as rapidly as possible according to whether they heard the last syllable as /da/ or /ga/. The response to trial i was denoted $r_i = 0$ for 'da' and 1 for 'ga'). Repetition of the stimulus was allowed as many times as required, but this option was rarely used. In case the speech could not be identified, the subject was instructed to guess one of the two syllables.

Stimulus presentation was randomized across the two targets and was divided into 20 sessions of 500 trials each (10,000 trials total). To prevent excessive fatigue, the experiment was scheduled over 4 days.

Musician	Age onset (year)	Years of practice	Instrument	Absolute pitch test (/20)
#1	13	7	Double bass; guitar	1
#2	6	20	Trombone	17
#3	6	13	Piano	20
#4	4	17	Violoncello	19
#5	6	15	Violin	8
#6	5	14	Accordion	17
#7	5	22	Violin; piano	18
#8	7	14	Violin	10
#9	5	18	Double bass	20
#10	6	16	Violin	6
#11	5	22	Piano	20
#12	7	13	Piano	13
#13	5	14	Flute; bassoon	2
#14	3.5	18	Opera singing	5
#15	5	21	violin, viola	15
#17	7	10	guitar	19
#18	7	13	guitar	2
#19	7	17	Double bass	14
#20	5	17	viola da gamba; bowed viol	8

Table 2. Details of all participants' musical experience.

Typically, one day of the experiment lasted 2 h and consisted of a short practice block (~5 min), 3 sessions (~45 min) followed by a long break (~15 min), one cognitive test (~10 min), and then 3 more sessions of the main experiment (~45 min). Sessions were separated by 3 minute breaks. During the practice block, participants received automatic feedback about the accuracy of their answers, and the SNR was set to −11 dB. During the main experiment, however, no feedback was provided, and the SNR was adapted from trial to trial using a 3-down, 1-up staircase procedure[49]. The SNR was increased by one step after each incorrect response and decreased by one step after three consecutive correct responses from the last change in stimulus intensity. At the beginning of each session, the step size was set to 2 dB to accelerate convergence and then decreased by 10% each step until a step size of 0.2 dB was attained. The initial SNR level was −11 dB, and each session began with the final SNR of the previous session. This procedure, which was designed to target a threshold corresponding to 79% correct answers, allowed us to control for task difficulty over the course of the experiment even in the case of a momentary loss of attention. Data from all participants can be downloaded at the following addresses: https://zenodo.org/record/21808 and https://zenodo.org/record/19134.

Derivation of Auditory Classification Images. ACIs were measured using the penalized GLM method[44,47]. Noisy stimuli ($\underline{s}_i$) were pre-processed before entering the model, using Lyon's cochlea model (quality factor = 8; overlapping = 40%)[50]. These cochleogram representations, binned at 15.6 ms in time and in 54 logarithmically spaced spectral bins between 96 and 7760 Hz, were used as predictors of the participant's response. The estimation process was the same as in a previous study[47], except for one notable improvement described below.

In the previous model, the prediction of the participant's response was based on the cochleogram of the noise plus a two-level factor corresponding to the presented target. This dissociation between the noise and the target, inspired by the literature on visual Classification Images[39,43], presupposes that these two elements are linearly combined in predictor space. However, this is not the case here because the cochleogram is not a linear operator. A more proper way to formulate the model is to directly use the cochleogram of the stimulus as the predictor (hereafter the cochleogram of stimulus $\underline{s}_i$ will be denoted $\underline{S}_i$). In line with the linear observer model derived from signal detection theory[42,51], a decision variable is formed by taking the dot product of $\underline{S}_i$ and an internal template $\underline{\beta}$, the ACI, and adding the single parameter c, which reflects the general bias of the participant in favoring 'da' or 'ga':

$$P(r_i = 1) = \phi(\underline{S}_i * \underline{\beta} - c) \quad (2)$$

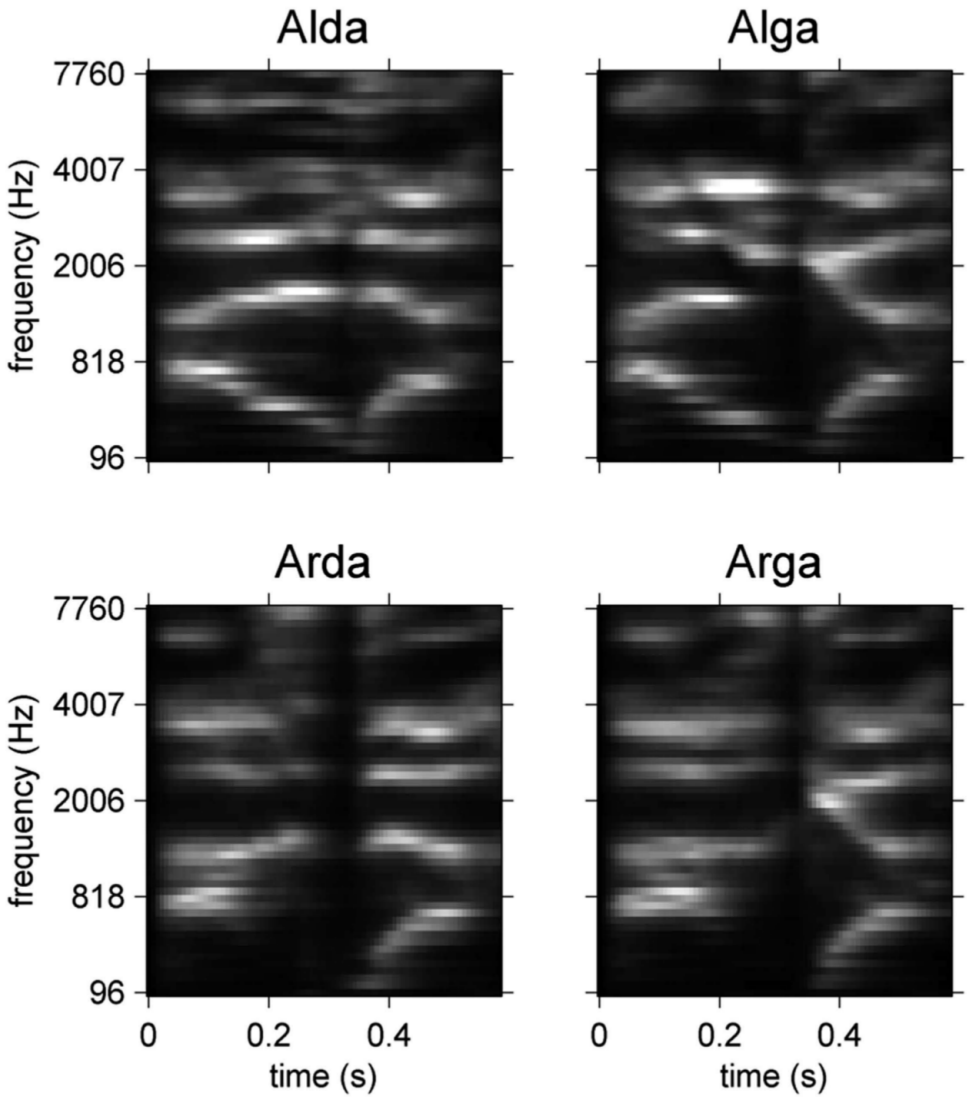


Figure 1. Cochleagrams of the four stimuli involved in the experiment. Parameters for spectral and temporal resolution are the same as those used for the derivation of ACIs (see details in the text).

ϕ is a link function (here, an inverse logit function) that transforms the decision variable $\underline{S}_i * \underline{\beta} - c$ into an answer probability, $P(r_i = 1) = 1 - P(r_i = 0)$. The model parameters $\underline{\beta}$ and c are estimated via the penalized likelihood (maximum a posteriori) of the GLM. The penalty chosen here is a smoothness prior and provides data-based, low-pass filtering of the ACI. The amount of penalty to be applied was determined by minimizing the mean cross-validated deviance of the participants' models. The same level of penalty was obtained in both groups (musicians and non-musicians), thereby confirming a posteriori our choice of prior, and this value was applied to all estimations in this study.

This new method for the calculation of ACIs poses a problem, however, as it is sensitive to imbalances between correctly and incorrectly categorized trials, potentially leading to distorted representations of the underlying categorization template. Yet, the staircase algorithm necessitates that all participants obtain a 79% correct score. A simple solution adopted here was to balance the number of errors and correct categorizations before estimating the ACI. This goal was achieved by discarding a number of randomly chosen correct trials.

A comparison of the two methods using the same group of data (not shown) indicates that the ACIs obtained using this new method are less noisy and that the secondary cues are more distinct, although the estimation was based on ~4,200 trials versus 10,000 trials for the previous method.

Statistical analysis. Once ACIs were estimated for all participants, they were individually z-scored and averaged to compute group ACIs, in accordance with previous work[52]. Because they are only approximations of the actual template used by the participant, encompassing a certain amount of estimation

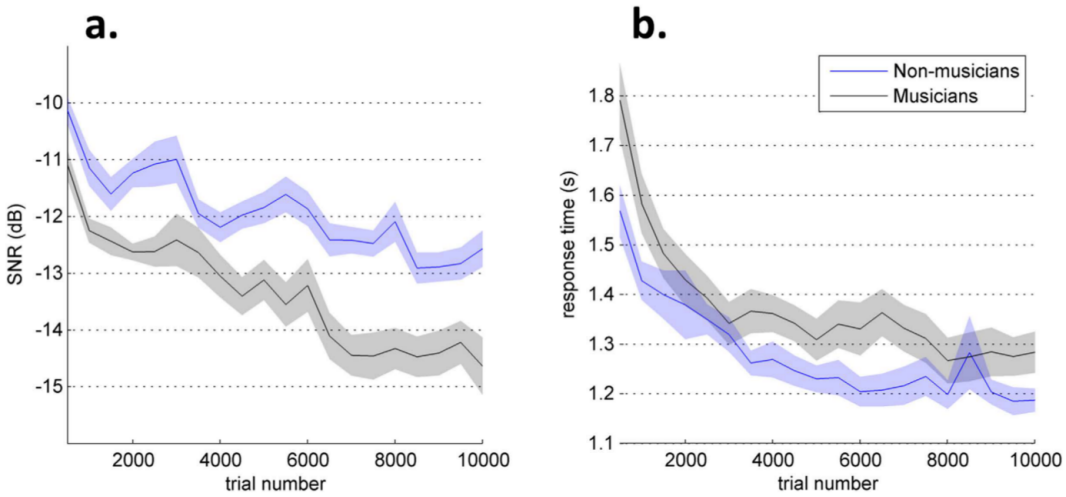


Figure 2. Evolution of SNR (**a**) and response time (**b**) over the course of the experiment, averaged across participants (the shaded region shows the s.e.m.). Each data point corresponds to a mean over 500 trials.

noise, it is imperative to apply some sort of statistical test to determine whether the resulting mean ACI is meaningful or can simply be due to a random process. Here, we used two different statistical tests from neuroimaging to decide which spectro-temporal regions truly reflected the listening strategy of a group of participants (t-test against zero with FDR correction[53]) or were significantly different between two groups of participants (t-test with cluster-based correction[54]).

Results

Overall, participants in both groups obtained very similar patterns of correct response rates across all 4 signals (all $p > .4$). Their mean correct response rates were all nearly equal to 79% throughout the experiment thanks to the staircase algorithm (musicians: $79.28\% \pm 0.34\%$ S.D.; non-musicians: $78.83\% \pm 0.40\%$ S.D.). These percentages, although very close, were significantly different ($t(36) = 3.6$; $p < .001$). We estimated the sensitivity and decision criteria, as defined in signal detection theory[51], for the two groups. Only sensitivity differed significantly (musicians: 1.68 ± 0.05 S.D.; non-musicians: 1.64 ± 0.04 S.D.; $t(36) = 3.11$; $p = 0.0036$). As a group, musician participants enrolled in this study also obtained better performances than the non-musician group in terms of SNR. Specifically, the mean SNR over all trials was significantly lower for musicians than for non-musicians (musicians: $-13.37\,\text{dB} \pm 1.19\,\text{dB}$ S.D.; non-musicians: $-11.91\,\text{dB} \pm 1.01\,\text{dB}$ S.D.; $t(36) = -3.97$; $p < .001$) resulting in a slightly superior recognition rate. The mean response time (RT) did not differ significantly between the groups (musicians: $1.37\,\text{s} \pm 0.18\,\text{s}$ S.D.; non-musicians: $1.27\,\text{s} \pm 0.12\,\text{s}$ S.D.; $t(36) = 1.77$; $p = 0.086$). These results are summarized in Table 1.

However, the participants' performance level is not a stable state. Rather, SNR and RT decrease slowly over the course of the experiment (Fig. 2). A 2-way repeated-measures ANOVA showed significant effects of session number ($F(19,36) = 26.93$, $p < .0001$) and group ($F(1,36) = 15.77$, $p = 0.0003$) on SNR, with a significant interaction effect ($F(19,684) = 1.83$, $p = 0.017$). Thus, musicians were not only better performers in general but they also demonstrate faster learning. RTs were only affected by session number ($F(19,36) = 32.88$, $p < .0001$), and by a significant interaction between group and session ($F(19,684) = 1.7$, $p = 0.0319$).

Because musicians performed the task in significantly higher levels of background noise than non-musicians, SNR acts as a confounding variable in the comparisons between the two groups. Therefore, we additionally compared two subgroups of equivalent SNRs. For this purpose, we excluded 4 participants with mean SNRs that deviated more than 1.65 SD from the mean SNR of the non-musician group (e.g., outside the range of $-11.91\,\text{dB} \pm 1.67\,\text{dB}$). Four additional non-musician participants were randomly discarded to result in two subgroups ($N = 13$) of equivalent SNR. Response times were also comparable (unpaired t-test, $t(24) = -0.64$, $p = 0.53$). Note, however, that the higher-performing musicians were obviously not included in the subgroup comparison.

Individual ACIs were derived, z-scored, and group-averaged. The group-ACIs obtained are shown in Fig. 3a,b, with non-significant parameters set to zero (FDR $> .01$). High positive (red) and negative (blue) clusters of weights are time-frequency regions where noise biased the response of the participants towards 'da' or 'ga', respectively. ACIs for both groups demonstrated a very similar distribution of positive and negative weights.

For greater clarity, we defined 11 sets of weights by gathering CIs from all 38 participants regardless of their group, performing a t-test against zero for each weight, and considering only regions that exceeded the arbitrary threshold of $p < 10^{-10}$ and formed a cluster of at least 7 adjacent significant pixels. Based

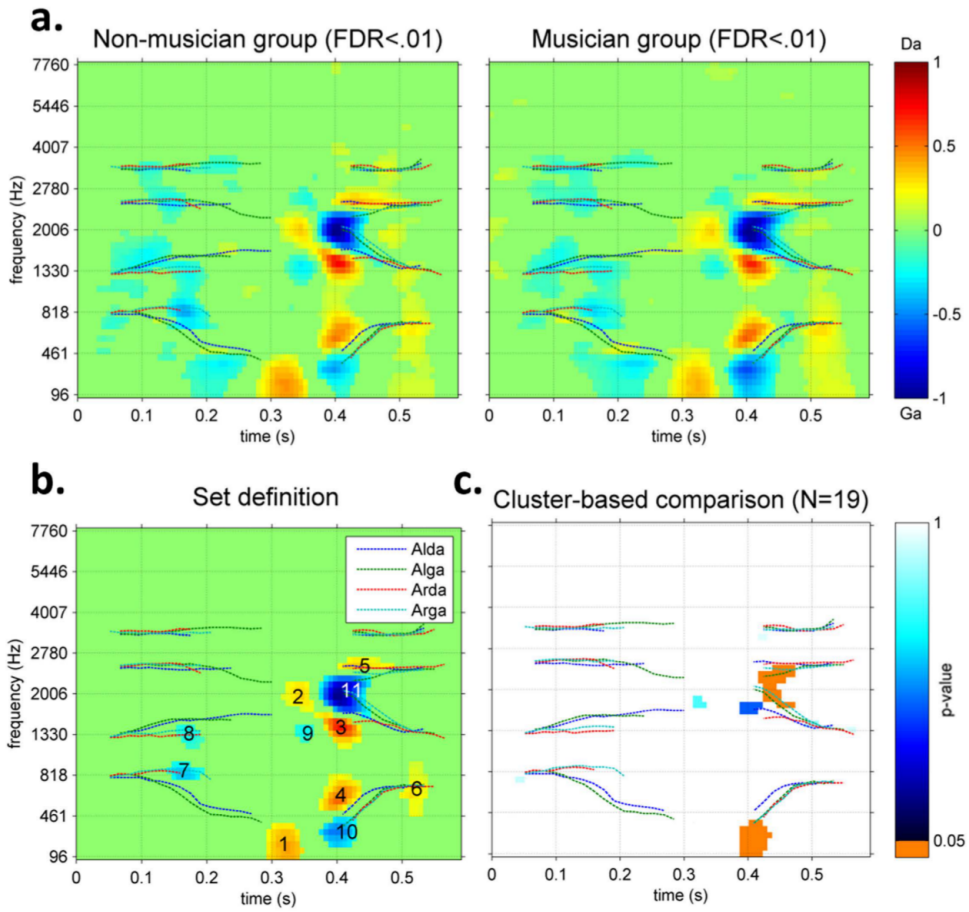


Figure 3. (**a**) ACIs for the two groups of participants (N = 19). Non significant weights (FDR > 0.01) are set to zero (**b**). Mean ACI over all 38 participants. Only weights sets (min. 7 adjacent weights with $p < 10^{-10}$) are shown (**c**). Cluster-based nonparametric test between ACIs for the Non-musician and Musician groups (N = 19).

on this profile, we characterized each set of weights by its size, time-frequency position (defined as the location of its centroid), spatial and temporal extent, sign (whether it was composed of positive or negative weights), and match with the acoustic features of the signals. The contours of all sets are plotted in Fig. 3c, and a summary of their characteristics is provided in Table 3.

As in STRFs, the ACI combines excitatory and inhibitory regions at close spectro-temporal positions to encode the presence of acoustic cues. Figure 3a,b. show that the main acoustic cue used by participants performing the task was located at the onsets of the F2 and F3 formant transitions in the second syllable (approximately 413 ms and 1960 Hz). The main acoustic cue consists of an inhibitory center in favor of the response 'ga' (set #11), with an excitatory surround in favor of the response 'da' (sets #2, #3 and #5). That is, the probability of the 'ga' answer increased in response to an increment in energy in set #11 and in response to a decrement in energy before, above and below this area. A similar configuration was observed at the onset of the F1 formant transition (sets #1, #4 and #10). Additionally, a number of weaker but significant weights were located in the first syllable and at the end of the second syllable (sets #6, #7 and #8).

A cluster-based nonparametric test indicated that musician listeners placed greater weight on the onsets of F2 and F3 (the latest part of set #11, $p = 0.029$) and on the onset of F1 (set #10, $p = 0.027$) than non-musician listeners (Fig. 3c). Arguably, differences in cue weighting between the two groups could simply be due to the global difference between the SNR at which they carried out the task (−13.37 dB on average for musicians versus −11.91 dB for non-musicians). To rule out this possibility, we also ran the same cluster-based test on two subgroups (N = 13) with equivalent SNRs. The same significant clusters were found (set #11: $p = 0.048$; set #10: $p = 0.01$), indicating that SNR differences were not the only cause of weighting changes between the musicians and non-musicians. We also made sure that these conclusions did not depend on the threshold used to delimit the clusters in the nonparametric test by replicating the above results with various thresholds values ($\alpha_{threshold} = .01$, .025 and .05).

One advantage of the cluster-based analysis is that it makes no assumptions about the location of the effect. However, to obtain a more complete picture of the strategies of the two groups, we also

Set	size (pxls)	Centroid		Extent		Correspondence with formants	Bias towards	Set weights				
		t (ms)	f (Hz)	t (ms)	f (Hz)			NM (mean)	NM (SD)	M (mean)	M (SD)	t-test (M vs. NM)
#1	41	317,9	215,8	51,04	256,8	offset F1, 1st syllable	'da'	0,0173	0,0097	0,0158	0,0065	p = 0,58
#2	25	340,8	1913	43,75	429,5	onset F2/F3, 2nd syllable	'da'	0,0070	0,0030	0,0075	0,0026	p = 0,60
#3	34	406	1383	65,63	425,4	onset F2, 2nd syllable	'da'	0,0166	0,0047	0,0160	0,0048	p = 0,68
#4	42	405	658,4	58,33	300,6	onset F1, 2nd syllable	'da'	0,0158	0,0060	0,0171	0,0044	p = 0,45
#5	19	443,5	2548	80,21	137,7	onset F3, 2nd syllable	'da'	0,0066	0,0018	0,0052	0,0022	p = 0,051
#6	34	521,6	697,1	36,46	489,7	offset F1, 2nd syllable	'da'	0,0075	0,0050	0,0090	0,0025	p = 0,27
#7	18	166,1	876,4	51,04	133,5	offset F1, 1st syllable	'ga'	−0,0067	0,0036	−0,0060	0,0030	p = 0,50
#8	17	173,9	1297	36,46	244,9	offset F1, 2nd syllable	'ga'	−0,0040	0,0017	−0,0042	0,0020	p = 0,63
#9	10	350	1347	21,88	166,4	onset F2, 2nd syllable	'ga'	−0,0029	0,0010	−0,0023	0,0017	p = 0,17
#10	32	404,9	301,3	51,04	209,2	onset F1, 2nd syllable	'ga'	−0,0113	0,0066	−0,0180	0,0043	**$p = 7.10^{-4}$***
#11	60	413,9	1960	72,92	645,6	onset F2/F3, 2nd syllable	'ga'	−0,0389	0,0076	−0,0461	0,0061	**$p = 2.10^{-3}$***

Table 3. Summary of the characteristics of all sets of weights, sorted by bias and latency.

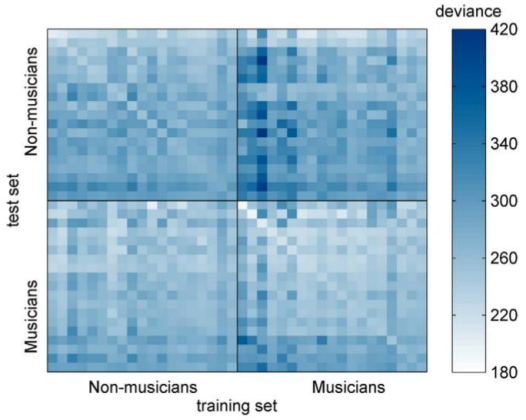


Figure 4. Cross-prediction deviances for all participants. Auto-predictions are represented along the diagonal of the matrix. Participants in each group are sorted by auto-prediction values.

performed a more classic ROI analysis on the mean weights in the previously defined sets. Table 3 shows the p-values for these t-tests. The hierarchy of the weights was the same in musicians and non-musicians, with relatively low inter-subject variability. Only sets #10 and #11 were weighted differently by the two groups, confirming the above results.

Similarly, we also investigated the specificity of each participant's listening strategy for both groups. A straightforward way of evaluating this characteristic is to test how a model trained on one participant's data can generalize to predict the responses of other participants. All goodness-of-fit measures are presented in terms of deviance (prediction error rates are also displayed for better understanding). As above, the proportions of correctly and incorrectly categorized trials were equated to make sure that the performance level of the participant did not impact the prediction of the model.

Ten-fold cross-validated predictions were calculated for each listener and termed "auto predictions". This method provides a measure of how a model trained on a subset of the listener's data can successfully predict other responses from the same participant. The data were randomly divided into 10 equal subsets, 9 of which constituted the training-set for fitting the GLM. Then, the predictive power of the resulting model was evaluated on the remaining subset (test-set). This operation was repeated 9 times until all subsets were used as test-sets once. Finally, the 10 deviance values were averaged to obtain the auto-prediction deviance for the participant. Auto-predictions were significantly better (unpaired t-test, $t(36) = 2.7$, $p = 0.01$) in the Musician group (233.41 ± 23.5 S.D.) than in the Non-Musician group (252.22 ± 19.1 S.D.).

Our aim was to compare the auto-prediction for each listener with his/her "cross-predictions". The latter were calculated in precisely the same way as the auto-predictions except that the test-set used was taken from a different subject than the training-set. The matrix of cross-prediction values is shown in Fig. 4. As expected, a given model was better able to predict unseen data from the participant on which it had been trained than data from another participant (paired t-test, $t(37) = -5.9805$, $p < 10^{-6}$).

On average, the auto-prediction performance reached 69.1% correct predictions, whereas the mean cross-prediction performance was approximately 63.9% correct (both performances were superior to the chance level of 50%). As described in the Methods section, the amount of the penalty applied to all maximum a posteriori estimations calculated in this article was determined by minimizing the mean auto-prediction deviance across all participants. To confirm the accuracy of our approach, we also verified that the same value was obtained by minimizing the mean cross-prediction deviance instead.

Additionally, cross-prediction deviances within the Musician group were lower than those within the Non-Musician group (unpaired t-test, $t(36) = -4.28$, $p < 10^{-3}$). It could not be inferred from the prediction data that GLMs fit on the musicians' responses were better predictors of the whole group of 38 participants than GLMs fit on the non-musicians' responses (unpaired t-test, $t(36) = -1.46$, $p = 0.15$). However, when averaging across training-sets instead of across test-sets, the musicians' responses were more accurately predicted than the non-musicians' responses (unpaired t-test, $t(36) = -3.44$, $p = 0.0015$).

A measure of the specificity of a listener's strategy can be obtained by taking the difference between his/her auto-prediction deviance and the cross-prediction deviance of his/her data produced by GLMs fit to the other participants. This value is higher when the subject's responses are better predicted by his/her own ACI than by others' ACIs. The specificity measures did not differ significantly between the two groups (Non-Musicians: -31.76 ± 7.04 S.D.; Musicians: -30.71 ± 14.88 S.D., unpaired t-test, $t(36) = -0.28$, $p = 0.78$).

Discussion

In this study, we aimed to determine how the enhanced cognitive abilities of expert musicians in non-speech acoustic tasks, compared with those of non-musicians, are linked to their better performance in speech-in-noise comprehension. For this purpose, we used the Auditory Classification Image approach in a simple phoneme categorization task in noise. Nineteen musicians and 19 normal-hearing participants with no musical practice were asked to discriminate the final syllable in 4 non-word stimuli (/alda/, /alga/, /aʁda/, and /aʁga/). The noise level was adjusted throughout the experiment so that the percentage of correct answers was 79%. ACIs were derived for each participant to reveal the precise time-frequency regions where noise influenced his phonemic categorization.

As expected, musicians enrolled in this study demonstrated better phoneme perception in background noise than non-musicians. Although the percentages of correct answers were close to 79% in both groups, the musicians performed the task at an average SNR that was 2 dB less favorable. Furthermore, their estimated sensitivities were slightly, but significantly, superior to those of non-musicians. These results confirm that musical expertise enhances resistance to noise, as has been demonstrated in previous studies[10]. This validates our experimental paradigm and enables us to explore the ACIs to identify correlates of this behavioral effect in the acoustic cues used. A tentative explanation for the slight, marginally significant, difference in response times between musicians and non-musicians could be that musicians are more meticulous in the task, being slower in their responses but obtaining better performances. Additionally the two groups show different evolution of performances with time, as indicated by the significant interaction effects between trial number and SNR or response time, suggesting faster learning for musicians.

All individual ACIs exhibited similar weight patterns, consistent with the results of a previous study that used a slightly different model (see Methods section) and a smaller number of participants[47]. The main sets of weights are depicted in Fig. 3c, and a summary of their characteristics is provided in Table 3. At least two acoustic cues appeared to be involved in the /da/-/ga/ categorization: 1) the configuration of the F2/F3 transitions was different for the two syllables (close onsets for /ga/, distant onsets for /da/). This property was encoded by weights in sets #3, #5 and #11. 2) The position of F1, although less informative for proper identification of the consonant, was precisely reflected in the ACI by sets #4 and #10. Both cues were preceded by opposite weights in the same frequency bands (sets #1, #2 and #9), indicating that energy detection in one channel at time t also depends on energy in the same channel at previous instances in time. Although the general listening strategies used by the participants in this study appear to have been very similar, we cannot exclude that there may have been fine group differences.

The main goal of this study was to determine whether and how the extensive training of auditory abilities in music experts modifies the auditory cues used for the categorization of /da/ and /ga/ in noise. If this was the case, there should have been a significant difference between the ACIs of the Musician and Non-Musician groups. Indeed, a cluster-based, nonparametric test indicated that musicians placed greater weights on the end of the central negative set #11 ($p = 0.029$) (Fig. 3c). Therefore, the contour of the auditory cue more closely traced the second and third formant transitions in the second syllable. Another difference was found for the second acoustic cue, more precisely in set #10. These two acoustic cues were more heavily weighted by musician participants, suggesting better selectivity for the task-relevant characteristics of the stimuli, which was enabled by the hypersensitivity of their over-trained auditory skills.

This picture is complicated by a potentially confounding variable: differences between the ACIs of the two groups can be explained by the differences in performance between the two groups, reflected by the lower SNRs for musicians compared to non-musicians. This possibility has been ruled out, however, because the two subgroups that were matched in SNR (and of equivalent reaction times) showed the same pattern of differences in their ACIs. It is noteworthy that although the higher-performing musicians

were not included in this last comparison, the result remains the same. This finding suggests that the plasticity of the auditory process induced by musical practice might impact speech processing even when it provides no or very little advantage for comprehension. To further investigate this effect we performed a ROI analysis on the ACIs of the low- and high- performing musicians. Significant differences were obtained only on sets #6 ($t(17) = -2.2597$, $p = 0.037$) and #7 ($t(17) = 2.7183$, $p = 0.015$) (the complete list can be found as Supplementary Table S1 online). This set of results would imply that all musician experts selectively focus on the relevant information contained in the speech stimuli, but some of them may be unable to process it effectively by combining it with secondary cues. However, the sample size being very small ($N = 6$ for the high-performer subgroup), further study is needed to confirm this tentative explanation.

Absolute pitch, a change in sensory representations of tones associated to early musical training for some people[55], may pose a quite serious confound for the interpretability of the results. Indeed the possession of this ability might pave the way for alternative strategies in our task, or enforce the existing strategy, possibly explaining the differences between musicians' and non-musicians' ACIs. This concern needs to be addressed here by a complementary analysis. Fortunately, it was possible to evaluate the effect of absolute pitch by dividing the musicians group into 2 subgroups, according to whether or not they scored more than 17/20 on the pitch-naming task. The differences between the ACIs of musicians with ($N = 8$) or without ($N = 11$) absolute pitch was determined by means of a ROI analysis. None of the 11 t-tests indicated a significant differential weighting of the cues (all $p > 0.05$). The detailed results are reported in Supplementary Table S1 online. This result supports the claim that the change in auditory processing observed in musicians, compared to non-musicians, is driven by group differences in musical experience rather than by a subgroup of absolute pitch possessors.

It may seem contradictory that intensive musical training, which has been shown to confer sharper tuning of cochlear filter responses[56] and finer spectral and temporal discrimination[10–14], resulted in larger cues in the ACI in the current study. However, in the context of speech-in-noise perception, this adaptation provides an advantage because it is a way to obtain more reliable estimates of formant positions by gathering information from a wider area. To summarize, although they used the same listening strategy as non-musicians, musicians selectively focused on a small time-frequency region critical for correct /da/-/ga/ categorization. In light of these results, the musicians' performance could be at least partially explained by the enhanced selectivity for the most behaviorally relevant aspects of the sound. One tentative explanation is that the more robust encoding of speech in the musicians' brainstem and their enhanced auditory abilities pave the way for more flexible cue weighting, such as the one described here. Even if they use the same acoustic cues, their extensive training of auditory functions makes them better able to focus on those cues and to not be influenced by non-behaviorally relevant acoustic information.

Another aspect that we wanted to investigate in this study was the specificity of each participant's listening strategy. This was performed by comparing the auto-predictions (prediction of a subject's response using his own ACI) and the cross-predictions (prediction of a subject's responses using the ACIs of the other subjects) (Fig. 4). We first confirmed the validity of this approach by verifying that the ACIs were generalizable from one participant to another: on average, the cross-prediction rate was significantly above chance (64%), although it was inferior to the mean auto-prediction rate (69%). While a study using Bubble Images, a method closely related to ACIs, demonstrated that intelligibility maps can be used to predict the intelligibility of other phonemes and/or other talkers[57], the current result shows that a comparable generalization is possible across listeners. The comparison of cross-prediction deviances between the two groups offers another insight into the differences in speech processing induced by musical training. On average, musicians' data were more accurately cross-predicted than non-musicians' responses ($p = 0.0015$), indicating that they were more consistent in their responses. This result could be due to a reduction of internal noise in the Musician group (fewer unpredictable mistakes), due perhaps to their increased sustained auditory attention capabilities[13]. However, musicians and non-musicians did not seem to differ in terms of the specificity of their ACIs, based on the gap between auto-prediction and mean cross-prediction for each participant in both groups ($p = 0.78$).

Of course, the acoustic cues identified depend on the categorization task studied (/da/-/ga/ categorization), and their precise time-frequency locations depend on the particular utterances presented. Nonetheless, our results demonstrate a phenomenon that is highly unlikely to be limited to this particular choice of phonemes and utterances. Furthermore, due to the high degree of nonlinearity in the speech comprehension process, it is possible that the exact weight ratio in the ACIs do not exactly reflect the real relative importance of acoustic cues[58]. However, this does not invalidate the difference in weighting between the two groups. Because the ACIs are estimated using the same set of targets, we completely avoid the issue of stimulus-dependent estimation.

A recent series of studies have questioned the initial finding of Parbery-Clark and colleagues[10] that musicians have a general enhancement of speech-in-noise comprehension compared to non-musicians. They varied the type of task used, the targets, and the complexity of the maskers, showing that generally, the group difference was very small and in most cases non-significant[59–61]. They presented a plausible explanation for this absence of positive results: the musicianship advantage would only be revealed in the most difficult listening situations[59]. However, this interpretation is challenged by our observations. Indeed, a significant difference was obtained in our study using a forced-choice task with word targets, which is less demanding than the task used by Boebinger *et al.* (open-choice task with sentence

targets). Furthermore, our adaptive staircase algorithm targeted a higher percentage of correct responses (79% compared with 50% in the previous study). Two alternative explanations can be proposed. First, the stronger effect of musicianship in our study could be due to a more strict selection of participants compared to the previous studies: non-musicians had no practice of an instrument, even for less than 2 years[10,59–61]. Our subjects were also required to have pure-tone thresholds better than 20 dB at the audiometric test, whereas normal hearing was not tested in Boebinger *et al.*[59]. These two factors can be a major source of inter-individual variability: on one hand, even as little as 6 months of musical training is sufficient to influence linguistic abilities[62]; on the other hand, musicians are more likely to experience hearing problems. Another explanation could be related to the non-naturalness of the forced-choice task we used. Here, we hypothesized that improved frequency discrimination enhances formant trajectory detection, thus resulting in better /da/-/ga/ categorization in noise. Nonetheless, it is possible that this ability does not strongly impact speech-in-noise perception in more natural listening situations, such as those employed in the studies above[59–61].

In this paper, we showed that the ACI is a suitable tool for providing direct visualization of auditory plasticity resulting from intensive musical training. We also showed its effect on speech perception. This approach fills a gap in the current debate on the beneficial effects of musical training on speech perception. In line with what has been shown previously in sABR studies[27,29–33] and in psychoacoustic studies[10–17], we were able to demonstrate an enhancement of specific acoustic characteristics in the encoding/decoding of speech sounds. One important aspect of our method, however, is that it is based on speech comprehension data instead of on electrophysiological data or behavioral data for non-speech sounds. Hence, the acoustic cues identified are those that are behaviorally relevant for the phoneme categorization task (whereas sABR representations, for example, show all of the information extracted from the signal, even the information that is not used in later stages of the comprehension process). In the context of this study, the ACI approach allowed us to determine which cues were actually - not only potentially - better extracted and combined by musicians to yield better speech-in-noise perception. As such, this method provides a bridge between two sets of evidence: the existence of a musicianship advantage in speech-in-noise perception on one hand and improvements of basic auditory abilities on the other hand. This article has set the stage for future studies using the ACI approach to investigate group differences in auditory plasticity. In contrast with the enhancement of speech perception, another interesting application would be to explore its impairments, such as in dyslexia.

References

1. Kraus, N., Skoe, E., Parbery-Clark, A. & Ashley, R. Experience-induced Malleability in Neural Encoding of Pitch, Timbre, and Timing. *Ann. N. Y. Acad. Sci.* **1169,** 543–557 (2009).
2. Wolfe, J. Speech and music, acoustics and coding, and what music might be 'for'. In *Proceedings of the 7th International Conference on Music Perception and Cognition,* 10–13 (2002).
3. DeLiang, W. & Brown, G. J. Fundamentals of computational auditory scene analysis. In *Computational Auditory Scene analysis* 1–44 (Wang, DeLiang & Brown, Guy J., 2006).
4. Moore, B. C., Tyler, L. K. & Marslen-Wilson, W. Introduction. The perception of speech: from sound to meaning. *Philos. Trans. R. Soc. B Biol. Sci.* **363,** 917–921 (2008).
5. Zatorre, R. J. & Salimpoor, V. N. From perception to pleasure: Music and its neural substrates. *Proc. Natl. Acad. Sci. USA* **110,** 10430–10437 (2013).
6. Patel, A. D. Why would Musical Training Benefit the Neural Encoding of Speech? The OPERA Hypothesis. *Front. Psychol* **2,** 142 (2011).
7. Gaser, C. & Schlaug, G. Brain Structures Differ between Musicians and Non-Musicians. *J. Neurosci.* **23,** 9240–9245 (2003).
8. Gaser, C. & Schlaug, G. Gray Matter Differences between Musicians and Nonmusicians. *Ann. N. Y. Acad. Sci.* **999,** 514–517 (2003).
9. Schneider, P. *et al.* Morphology of Heschl's gyrus reflects enhanced activation in the auditory cortex of musicians. *Nat. Neurosci.* **5,** 688–694 (2002).
10. Parbery-Clark, A., Skoe, E., Lam, C. & Kraus, N. Musician enhancement for speech-in-noise. *Ear Hear.* **30,** 653–661 (2009).
11. Kishon-Rabin, L., Amir, O., Vexler, Y. & Zaltz, Y. Pitch discrimination: are professional musicians better than non-musicians? *J. Basic Clin. Physiol. Pharmacol.* **12,** 125–143 (2001).
12. Micheyl, C., Delhommeau, K., Perrot, X. & Oxenham, A. J. Influence of musical and psychoacoustical training on pitch discrimination. *Hear. Res.* **219,** 36–47 (2006).
13. Strait, D. L., Kraus, N., Parbery-Clark, A. & Ashley, R. Musical experience shapes top-down auditory mechanisms: evidence from masking and auditory attention performance. *Hear. Res.* **261,** 22–29 (2010).
14. Rammsayer, T. & Altenmüller, E. Temporal Information Processing in Musicians and Nonmusicians. *Music Percept. Interdiscip. J* **24,** 37–48 (2006).
15. Magne, C., Schön, D. & Besson, M. Musician Children Detect Pitch Violations in Both Music and Language Better than Nonmusician Children: Behavioral and Electrophysiological Approaches. *J. Cogn. Neurosci* **18,** 199–211 (2006).
16. Schön, D., Magne, C. & Besson, M. The music of speech: music training facilitates pitch processing in both music and language. *Psychophysiology* **41,** 341–349 (2004).
17. Gaab, N. *et al.* Neural correlates of rapid spectrotemporal processing in musicians and nonmusicians. *Ann. N. Y. Acad. Sci.* **1060,** 82–88 (2005).
18. Brattico, E. *et al.* Neural discrimination of nonprototypical chords in music experts and laymen: an MEG study. *J. Cogn. Neurosci* **21,** 2230–2244 (2009).
19. Koelsch, S., Schröger, E. & Tervaniemi, M. Superior pre-attentive auditory processing in musicians. *Neuroreport* **10,** 1309–1313 (1999).
20. Zendel, B. R. & Alain, C. Concurrent Sound Segregation Is Enhanced in Musicians. *J. Cogn. Neurosci* **21,** 1488–1498 (2009).
21. Pantev, C. *et al.* Increased auditory cortical representation in musicians. *Nature* **392,** 811–814 (1998).
22. Shahin, A., Roberts, L. E., Pantev, C., Trainor, L. J. & Ross, B. Modulation of P2 auditory-evoked responses by the spectral complexity of musical sounds. *Neuroreport* **16,** 1781–1785 (2005).

23. François, C., Jaillet, F., Takerkart, S. & Schön, D. Faster Sound Stream Segmentation in Musicians than in Nonmusicians. *PLoS ONE* **9,** e101340 (2014).
24. Marie, C., Magne, C. & Besson, M. Musicians and the metric structure of words. *J. Cogn. Neurosci* **23,** 294–305 (2011).
25. Kraus, N. & Chandrasekaran, B. Music training for the development of auditory skills. *Nat. Rev. Neurosci.* **11,** 599–605 (2010).
26. Bidelman, G. M. & Krishnan, A. Effects of reverberation on brainstem representation of speech in musicians and non-musicians. *Brain Res.* **1355,** 112–125 (2010).
27. Parbery-Clark, A., Skoe, E. & Kraus, N. Musical Experience Limits the Degradative Effects of Background Noise on the Neural Processing of Sound. *J. Neurosci*. **29,** 14100–14107 (2009).
28. Strait, D. L., Parbery-Clark, A., Hittner, E. & Kraus, N. Musical training during early childhood enhances the neural encoding of speech in noise. *Brain Lang.* **123,** 191–201 (2012).
29. Musacchia, G., Sams, M., Skoe, E. & Kraus, N. Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. *Proc. Natl. Acad. Sci* **104,** 15894–15898 (2007).
30. Song, J. H., Skoe, E., Wong, P. C. M. & Kraus, N. Plasticity in the adult human auditory brainstem following short-term linguistic training. *J. Cogn. Neurosci.* **20,** 1892–1902 (2008).
31. Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T. & Kraus, N. Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nat. Neurosci.* **10,** 420–422 (2007).
32. Lee, K. M., Skoe, E., Kraus, N. & Ashley, R. Selective Subcortical Enhancement of Musical Intervals in Musicians. *J. Neurosci.* **29,** 5832–5840 (2009).
33. Parbery-Clark, A., Tierney, A., Strait, D. L. & Kraus, N. Musicians have fine-tuned neural distinction of speech syllables. *Neuroscience* **219,** 111–119 (2012).
34. Strait, D. L., O'Connell, S., Parbery-Clark, A. & Kraus, N. Musicians' enhanced neural differentiation of speech sounds arises early in life: developmental evidence from ages 3 to 30. *Cereb. Cortex N. Y. N 1991* **24,** 2512–2521 (2014).
35. Ahumada, A. J., Jr & Lovell, J. Stimulus Features in Signal Detection. *J. Acoust. Soc. Am.* **49,** 1751–1756 (1971).
36. Ahumada, A. J., Jr, Marken, R. & Sandusky, A. Time and frequency analyses of auditory signal detection. *J. Acoust. Soc. Am.* **57,** 385–390 (1975).
37. Ahumada, A. J., Jr. Classification image weights and internal noise level estimation. *J. Vis.* **2,** 121–131 (2002).
38. Abbey, C. K. & Eckstein, M. P. Theory for Estimating Human-Observer Templates in Two-Alternative Forced-Choice Experiments. In *Information Processing in Medical Imaging* (eds. Insana, M. F. & Leahy, R. M.) 24–35 (Springer Berlin Heidelberg, 2001).
39. Knoblauch, K. & Maloney, L. T. Estimating classification images with generalized linear and additive models. *J. Vis.* **8,** 10.1–19 (2008).
40. Kurki, I., Saarinen, J. & Hyvärinen, A. Investigating shape perception by classification images. *J. Vis.* **14,** 24 (2014).
41. Kurki, I. & Eckstein, M. P. Template changes with perceptual learning are driven by feature informativeness. *J. Vis.* **14,** 6 (2014).
42. Knoblauch, K. & Maloney, L. T. *Modeling Psychophysical Data in R*. (Springer Science & Business Media, 2012).
43. Mineault, P. J., Barthelmé, S. & Pack, C. C. Improved classification images with sparse priors in a smooth basis. *J. Vis.* **9,** 17.1–24 (2009).
44. Varnet, L., Knoblauch, K., Meunier, F. & Hoen, M. Using auditory classification images for the identification of fine acoustic cues used in speech perception. *Front. Hum. Neurosci.* **7,** 865 (2013).
45. David, S. V., Mesgarani, N. & Shamma, S. A. Estimating sparse spectro-temporal receptive fields with natural stimuli. *Netw. Bristol Engl.* **18,** 191–212 (2007).
46. Theunissen, F. E. & Elie, J. E. Neural processing of natural sounds. *Nat. Rev. Neurosci.* **15,** 355–366 (2014).
47. Varnet, L., Knoblauch, K., Serniclaes, W., Meunier, F. & Hoen, M. A Psychophysical Imaging Method Evidencing Auditory Cue Extraction during Speech Perception: A Group Analysis of Auditory Classification Images. *PLoS ONE* **10,** e0118009 (2015).
48. Fan, J., McCandliss, B. D., Sommer, T., Raz, A. & Posner, M. I. Testing the efficiency and independence of attentional networks. *J. Cogn. Neurosci.* **14,** 340–347 (2002).
49. Levitt, H. Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.* **49,** Suppl 2:467+(1971).
50. Slaney, M. & Lyon, R. F. *Lyon's cochlear model* (Apple Computer, Advanced Technology Group, 1988).
51. Abdi, H. Signal Detection Theory in *Encyclopedia of Measurement and Statistics* (Neil Salkind, 2007).
52. Neri, P. & Levi, D. M. Evidence for joint encoding of motion and disparity in human visual perception. *J. Neurophysiol.* **100,** 3117–3133 (2008).
53. Genovese, C. R., Lazar, N. A. & Nichols, T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* **15,** 870–878 (2002).
54. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164,** 177–190 (2007).
55. Zatorre, R. J. Absolute pitch: a model for understanding the influence of genes and development on neural and cognitive function. *Nat. Neurosci.* **6,** 692–695 (2003).
56. Bidelman, G. M., Schug, J. M., Jennings, S. G. & Bhagat, S. P. Psychophysical auditory filter estimates reveal sharper cochlear tuning in musicians. *J. Acoust. Soc. Am.* **136,** EL33–EL39 (2014).
57. Mandel, M. I., Yoho, S. E. & Healy, E. W. Generalizing time-frequency importance functions across noises, talkers, and phonemes. In *Proceedings of Interspeech* (2014).
58. Christianson, G. B., Sahani, M. & Linden, J. F. The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. *J. Neurosci. Off. J. Soc. Neurosci.* **28,** 446–455 (2008).
59. Boebinger, D. *et al.* Musicians and non-musicians are equally adept at perceiving masked speech. *J. Acoust. Soc. Am.* **137,** 378–387 (2015).
60. Fuller, C. D., Galvin, J. J., Maat, B., Free, R. H. & Başkent, D. The musician effect: does it persist under degraded pitch conditions of cochlear implant simulations? *Front. Neurosci*. **8,** 179 (2014).
61. Ruggles, D. R., Freyman, R. L. & Oxenham, A. J. Influence of Musical Training on Understanding Voiced and Whispered Speech in Noise. *PLoS ONE* **9,** e86980 (2014).
62. Moreno, S. *et al.* Musical Training Influences Linguistic Abilities in 8-Year-Old Children: More Evidence for Brain Plasticity. *Cereb. Cortex* **19,** 712–723 (2009).

Acknowledgements

LV is funded by a PhD grant from the Ecole Doctorale Neurosciences et Cognition (http://nsco.universite-lyon.fr/), Lyon-1 University, France. This research was partially supported by a European Research Council (http://erc.europa.eu/) grant for the SpiN project (No. 209234) attributed to FM and by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083).

Author Contributions

L.V., T.V., M.H. and F.M. conceived and design the experiment. L.V. developed the matlab toolbox for derivation of the images. T.V. and C.P. carried out the experiment and L.V. performed data analysis. L.V., M.H. and F.M. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Varnet, L. *et al.* How musical expertise shapes speech perception: evidence from auditory classification images. *Sci. Rep.* **5**, 14489; doi: 10.1038/srep14489 (2015).

11.3. Résumé de l'étude 3

La présente étude a tout d'abord fourni l'occasion d'étendre les résultats décrits dans l'article 2, pour des participants non-musiciens, au cas de participants musiciens experts. Pour ces derniers également, la catégorisation de /da/ et /ga/ dans le bruit repose sur l'extraction des attaques des premier, second et troisième formants. Les performances supérieures des musiciens dans la tâche ne découlent donc pas de l'utilisation d'une stratégie d'écoute radicalement différente. En revanche la pondération des indices acoustiques n'est pas identique entre les deux groupes, les musiciens apparaissant davantage à même de « focaliser » leur écoute sur les informations cruciales contenues dans le stimulus, en faisant abstraction des régions temps-fréquence sans importance pour la tâche considérée. Ces résultats apportent une indication importante concernant le transfert des capacités musicales à des situations de compréhension de la parole dans le bruit : l'encodage plus fin de la parole chez les auditeurs musiciens leur permettrait une meilleure extraction des indices acoustiques – observable sur l'ACI – et, par suite, une identification plus robuste des phonèmes.

Une étape a également été franchie dans le calcul des ACIs qui sont désormais dérivées à partir des cochléogrammes des stimuli, et non plus seulement des cochléogrammes des bruits. Ce procédé est non seulement plus correct mathématiquement mais, également, plus efficace, puisque la sélection de l'hyperparamètre est à présent extrêmement robuste et que l'étape de découpage en bandes de fréquences n'est plus nécessaire pour faire apparaître les indices basses-fréquences.

12. Étude 4 : Exploration des stratégies compensatoires mises en place par les auditeurs dyslexiques pour la catégorisation de phonèmes dans le bruit

12.1. Présentation de l'étude 4

À la suite de l'analyse des résultats des musiciens experts, nous avons entrepris dans une quatrième étude d'appliquer le même protocole expérimental à des auditeurs atteints de dyslexie développementale. En effet, comme nous l'avons souligné dans la partie théorique (paragraphe 5.3), ce trouble de l'apprentissage de la lecture implique dans la grande majorité des cas un déficit phonologique. Si les poids de l'ACI reflètent indirectement les représentations phonologiques ou, du moins, les indices acoustiques qui activent ces représentations, nous pouvons espérer obtenir ainsi un moyen de tester directement les hypothèses concernant le déficit de perception catégorielle dans la dyslexie et les stratégies compensatoires éventuellement mises en place.

Vingt participants diagnostiqués comme dyslexiques furent recrutés pour cette expérience et accomplirent la même tâche de catégorisation /da/-/ga/ que dans les études 2 et 3. Ils passèrent également une batterie de tests cognitifs, attentionnels et phonologiques visant à confirmer la dyslexie. Les ACIs individuelles calculées sur la base de leurs réponses furent confrontées à celles de participants contrôles du même âge (identiques à ceux de l'étude 3), tout d'abord au moyen d'une comparaison par groupe puis par une analyse individuelle basée sur la spécificité. Nous nous attendions ainsi à observer soit une différence générale portant sur les indices acoustiques utilisés par les participants dyslexiques pour réaliser la tâche demandée, soit une plus grande hétérogénéité des stratégies d'écoute dans le groupe dyslexique.

Cet article est actuellement soumis pour publication dans la revue en accès ouvert PLoS ONE. Par ailleurs, une partie des résultats a déjà été présentée lors de la conférence Society for the Neurobiology of Language 2014 à Amsterdam (Varnet et al., 2014a).

12.2. Article 4 : Direct viewing of dyslexics' compensatory strategies in speech in noise using auditory classification images.

Title: Direct viewing of dyslexics' compensatory strategies in speech in noise using auditory classification images.

Authors: Léo Varnet [1,2,3*], Fanny Meunier [2,3], Gwendoline Trollé [1,3], Michel Hoen [1,3]

Affiliations:
[1] Lyon Neuroscience Research Center, CNRS UMR 5292, INSERM U1028, Auditory Language Processing (ALP) research group, Lyon, France.

[2] Laboratoire sur le Langage le Cerveau et la Cognition, CNRS UMR 5304, Auditory Language Processing (ALP) research group, Lyon, France.

[3] Université de Lyon, Université Lyon 1, Lyon, France.

***Corresponding Author**
E-mail: leo.varnet@isc.cnrs.fr

Abstract:

A vast majority of dyslexic children exhibit a phonological deficit, particularly noticeable in phonemic identification or discrimination tasks. The gap in performance between dyslexic and normotypical listeners appears to decrease into adulthood, suggesting that some individuals with dyslexia develop compensatory strategies. Some dyslexic adults however remain impaired in more challenging listening situations such as in the presence of background noise. This paper addresses the question of the compensatory strategies employed, using the recently developed Auditory Classification Image (ACI) methodology. The results of 18 dyslexics taking part in a phoneme categorization task in noise were compared with those of 18 normotypical age-matched controls. By fitting a penalized Generalized Linear Model on the data of each participant, we obtained his/her ACI, a map of the time-frequency regions he/she relied on to perform the task. Even though dyslexics performed significantly less well than controls, we were unable to detect a robust difference between the mean ACIs of the two groups. This is partly due to the considerable heterogeneity in listening strategies among a subgroup of 7 low-performing dyslexics, as confirmed by a complementary analysis. When excluding these participants to restrict our comparison to the 11 dyslexics performing as well as their average-reading peers, we found a significant difference in the F3 and F4 onsets of the first syllable, suggesting that this population developed a compensatory strategy based upon 2 additional allophonic cues.

MANUSCRIPT:

1. Introduction

Developmental dyslexia is an extensively researched and documented learning disability, and one of the most common causes of reading difficulties, affecting about 5%-10% of school-age children and persisting in adulthood. It is characterized by reading performances well below the normal range of the age group and IQ level, and not explained by sensory deficits or inadequate scholarship only. This concise definition contrasts with the heterogeneity of associated cognitive impairments, which are observed in reading tasks [1], but also speech comprehension tasks [2], auditory processing of rapid sounds [3], visual tasks [4], and even postural tasks [5].

Although causes of developmental dyslexia remain opaque, it is acknowledged that a vast majority (75% to 100%) of dyslexic individuals show a phonological deficit, noticeable in tasks involving phonological awareness (e.g. spoonerism), verbal short-term memory (e.g. non-word repetition) or lexical retrieval (e.g. rapid automatic naming tasks) [6–8]. However, the exact nature of this impairment is still debated. Some authors have proposed that it may result from an impairment in the access to phonological representations [9–11], an abnormal auditory sampling [12,13], or from underspecified [14] or overspecified [15,16] phonological representations, leading in all cases to a blurring of boundaries between phonological categories. One major difficulty in testing these hypotheses arises from the fact that dyslexics do not form a homogeneous population, showing very different patterns of errors. Thus dyslexia is often divided into subtypes, possibly originating from deficits at various stages of the comprehension system [1,17]. The pattern of speech deficits in developmental dyslexia has been usually investigated by comparing a group of dyslexics to a group of controls matched in chronological age or in reading level. However, given the variety of subtypes, group differences may mask a wide variability in behavioral responses of dyslexic participants. Therefore, some authors have also employed individual deviance analyses to asses abnormal performances at the individual level [7,8,18].

With experience, dyslexics often develop compensatory strategies for speech recognition, giving the misleading impression that dyslexia disappears with time. Indeed, they usually demonstrate no or weak deficits for perception in quiet, probably because they capitalize on the very redundant nature of speech. However, difficulties are still reported under more challenging conditions, such as speech-in-noise perception. By the age of 8, dyslexic children perform as well as matched normal readers in a speech-in-quiet task, but they still experience difficulties understanding noisy speech [19,20], which can in turn become particularly disabling to achieve normal school progress in the context of a noisy classroom. In the same line, it is now well acknowledged that dyslexics [21,22] and children with a familial history of dyslexia [18] are generally more affected by the presence of strong background noise, compared to matched control groups without learning disabilities. By contrast, tone-in-noise detection deficits are weaker and found only in specific subgroups [8,23] suggesting that speech-in-noise difficulties stem from a phonological problem rather than an auditory problem. It may have roots in poorer encoding of speech sounds at the

subcortical level, as measured by speech auditory brainstem responses [24–27]. The impaired robustness of speech representations in noise has even been proposed as a core deficit of dyslexia, as behavioral measures predict reading performances better than other cognitive, auditory or attentional abilities [20], and neurophysiological measures in preschoolers predict future reading development [27].

Some authors have nevertheless highlighted that the speech-in-noise impairment is not always observed [28], and highly dependent to the type of background and the listening configuration used [7,21], dyslexics even showing better release from masking than normal reading controls in some cases [20,21]. This again suggests that alternative processing strategies may be employed to compensate for the speech recognition in noise deficit. For instance, concurrent babbles are particularly deleterious to dyslexic's comprehension, when information about sounds stream localization is not available, but not in more natural conditions, when speech and noise originate from two separate sources [21]. In the latter case, however, dyslexics show a right hemisphere over-activation, suggesting a reallocation of neural resources [22]. In the same line, two studies demonstrated that dyslexics [29] or children at risk for dyslexia [30] with no apparent behavioral deficits in a phoneme categorization task nevertheless show heightened sensitivity to non-linguistic information in their neural responses. Therefore, even when they behaviorally compensate for their phonological difficulties, dyslexics still show neurophysiological evidence of a less efficient processing of phonemes.

The present experiment aimed at exploring the listening strategies employed by dyslexic individuals for phoneme categorization. Here we chose to test the distinction between places of articulation in stop consonants, which has been shown to be particularly impacted in dyslexics [20,31]. As we mentioned earlier, the addition of a sufficient amount of background noise is necessary to reveal their difficulties, which are otherwise compensated. In this context, the Auditory Classification Image (ACI) method seems perfectly adapted as it relies on a phoneme categorization task in noise to derive individual maps of participant's listening strategies.

The ACI method [32,33] has been developed as an auditory version of the Classification Image method, which aims at revealing the primitives used in various visual detection or categorization [34–37] and more broadly the strategies used to make decision in forced-choice tasks [38–42]. Another way of seeing the ACI is as a behavioral spectrotemporal receptive field, derived from the responses of a participant instead of those of an auditory neuron [43–45]. This technique relies on a forced-choice phoneme categorization task in noise. The idea here is to capitalize on the masking noise to predict the responses of the listener on a trial-by-trial basis: by training a statistical model on the categorization data we can uncover how a specific noise configuration misleads the participant towards on particular phoneme. The result is a spectrotemporal perceptual map showing the time-frequency regions where the presence of energy influences the phonemic decision. Therefore this method allows us to visualize which parts of the speech stimuli serve as auditory cues for phoneme categorization.

The ACI method has already been successfully used on normal-hearing participants performing a /da/-/ga/ discrimination in context /al/ or /aʁ/ [33]. We have thus identified 3

acoustic cues involved in this task: the F1, F2 and F3 onsets. This estimate was precise enough to allow comparison between different groups of participants as a second step. Researchers in our team calculated in the same way the images for a group of musician experts to examine differences between the weighting strategies of the two groups [46]. As expected, professional musicians showed a better resistance to noise and their classification images appeared to be more focused on critical cues, revealing their finer auditory abilities. However, contrary to what may be initially assumed, the analysis of cross-predictions between musicians and non-musicians does not show an increased specificity of individual listening strategies in the second group.

The present experiment followed the same procedure, with a group of dyslexics compared to a group of average reader participants. The purpose was to determine if there was any significant group difference in the weighting of the acoustic cues that may reveal the phonological deficit in dyslexics or the compensatory strategies employed. We reasoned here that if their representations of phonemes were under- or overspecified in a consistent way, one should observe large dissimilarities in their ACI, in average, compared to those of normal-hearing controls.

2. Materials and Methods

The study was approved by the Comité d'évaluation éthique de l'Inserm / Institutional Review Board (IORG0003254, FWA00005831, IRB00003888). Participants provided written informed consent before participating in the experiment.

Participants

Twenty French-native volunteers with dyslexia took part in this study. Participants were informed about the experimental procedure used and they received a financial compensation for their participation (100€). They all had prior diagnosis of dyslexia from a psychologist. Participants had the option of ending the experiment at any time, but none of them did. From these original recordings, 2 participants had to be rejected from the dataset due to excessively low performances. The analyses reported in the following are thus based on 18 recordings (11 females, age 22.8 years ± 6.5 years S.D.).

Eighteen typical readers were selected from a previous study (Varnet et al., 2015) to match the dyslexic group in age, nonverbal IQ and handedness (12 females, 22.8 years ± 4.5 years S.D.). All participants included in the present study had normal hearing (audiometric pure-tone threshold ≤ 20 dB on the 125 Hz – 8000 Hz range for both ears), and reported no history of neurological disorders.

Cognitive and phonological tests

Attention capacities were evaluated using the Attention Network Test (ANT) [47]. Nonverbal IQ was assessed by the Raven's Standard Progressive Matrices. All participants obtained normal scores above the 50^{th} percentile of their age category (corresponding to a

score of at least 42/60). Literacy and phonological skills were assessed by the ECLA-16+ [48]. This battery includes the French-language standardized L'Alouette Reading Test [49], phonological awareness tests (phoneme deletion and spoonerism), reading and spelling tests, and working memory tests (digit span, backward/forward). All results are reported Table 1. The details of the results can be found at https://zenodo.org/record/29239.

		Dyslexic group	Control group	t-test
N		18	18	
Gender (f/m)		11/9	12/8	
Age		22.83 (± 6.52 S.D.)	22.83 (± 4.48 S.D.)	p=1
Handedness (Edinburgh test)		73.33 (± 30.48 S.D.)	62.78 (± 55.17 S.D.)	p=0.469
Raven's Standard Progressive Matrices (score /60)		49.67 (± 4,40 S.D.)	50.83 (± 4,12 S.D.)	p=0.436
Reading age (in months)		124.06 (± 26,54 S.D.)	186.78 (± 24,7 S.D.)	p=3.043
Reading tests	regular words (score /20)	19.11 (± 0,74 S.D.)	19.72 (± 0,45 S.D.)	$p=6.10^{-3}$**
	regular words (time in s)	21.39 (± 8,43 S.D.)	11.94 (± 1,99 S.D.)	$p=7.10^{-5}$***
	irregular words (score /20)	18.06 (± 1,93 S.D.)	19.11 (± 0,81 S.D.)	p=0.045*
	irregular words (time in s)	20.22 (± 7,64 S.D.)	12.17 (± 2,11 S.D.)	$p=1.10^{-4}$***
	pseudowords (score /20)	17.17 (± 3,27 S.D.)	18.78 (± 2,27 S.D.)	p=0.105
	pseudowords (time in s)	35.94 (± 14,98 S.D.)	18.5 (± 3,24 S.D.)	$p=4.10^{-5}$***
Spelling tests	sentences: orthography (score /10)	6.56 (± 1,86 S.D.)	9.28 (± 0,65 S.D.)	$p=2.10^{-6}$***
	sentences: grammar (s/10)	6.17 (± 2,39 S.D.)	8.67 (± 0,67 S.D.)	$p=2.10^{-4}$***
	regular words (score /10)	7.17 (± 1,34 S.D.)	9 (± 0,75 S.D.)	$p=2.10^{-5}$***
	regular words (time in s)	48.39 (± 6,30 S.D.)	38.72 (± 5,64 S.D.)	$p=4.10^{-5}$***
	irregular words (score /10)	4.39 (± 1,92 S.D.)	7.833 (± 1,30 S.D.)	$p=5.10^{-7}$***
	irregular words (time in s)	53.33 (± 10,96 S.D.)	42.06 (± 8,98 S.D.)	$p=2.10^{-3}$**
	pseudowords (score /10)	7.06 (± 1,54 S.D.)	8.78 (± 1,84 S.D.)	$p=5.10^{-3}$**
	pseudowords (time in s)	59.44 (± 14,80 S.D.)	45.22 (± 4,88 S.D.)	$p=6.10^{-4}$***
Phonological awareness tests	Phoneme deletion (score /10)	6.5 (± 2,22 S.D.)	9.22 (± 1,90 S.D.)	$p=5.10^{-4}$***
	Phoneme deletion (time in s)	45.78 (± 11,14 S.D.)	27.78 (± 6,20 S.D.)	$p=1.10^{-6}$***
	Spoonerism (score /20)	15.5 (± 4,04 S.D.)	19.22 (± 0,97 S.D.)	$p=7.10^{-4}$***
	Spoonerism (time in s)	141.72 (± 55,71 S.D.)	69.22 (± 19,68 S.D.)	$p=1.10^{-5}$***
Memory span tests	Pseudowords repetition (score /20)	18.72 (± 1,24 S.D.)	19.5 (± 0,69 S.D.)	p=0.030*
	Forward digit	6.17 (± 1.01 S.D.)	7.22 (± 0.98 S.D.)	$p=3.10^{-3}$**
	Backward digit	4.56 (± 1.21 S.D.)	6.11 (± 1.20 S.D.)	$p=6.10^{-4}$***
ANT	alerting effect	32.89 (± 26.28 S.D.)	33.06 (± 21.76 S.D.)	p=0.984
	orienting effect	53.00 (± 15.38 S.D.)	38.83 (± 22.46 S.D.)	p=0.039*
	conflict effect	155.39 (± 42.23 S.D.)	130.44 (± 36.82 S.D.)	p=0.075

Table 1. Summary of the characteristics of the dyslexic and normal-reading groups

Stimuli

Targets were borrowed from a previous study [33]. They consisted in 4 /aCCa/ nonwords with a continuant consonant (/l/ or /ʁ/) followed by a stop consonant (/d/ or /g/). All sounds were recorded by a male speaker and digitized at a sampling rate of 4.8 kHz (8-bit). Targets were equated both in total length (680 ms) and duration of the 1st syllable (328 ms) by cutting of the end of the syllables when necessary. The resulting audio .wav files can be downloaded at https://zenodo.org/record/12300.

For each participant, a set of 10,000 white noise instances of same duration as the targets were generated and stored before the beginning of the experiment. These files can be found at the addresses https://zenodo.org/record/23064 and https://zenodo.org/record/19102.

Experimental procedure

Participants sat in an acoustically isolated chamber in front of video monitor where they read instructions for the experiment. They wore Sennheiser's HD 448 headphones. On a given trial, they were presented with one of the four possible targets superimposed with additive white noise. The SNR was adapted from one trial to the next based on the performance level using a 3-down 1-up staircase procedure to target the 79% correct point [50]. All stimuli were power-normalized and presented at each participant's most comfortable sound level.

The task of the participant was to identify the last syllable of the stimulus as /da/ or /ga/, independently of the preceding consonantal context, and to respond as quickly as possible by a button press. The response to trial i is denoted r_i (0 for 'da' and 1 for 'ga'). Participants were allowed to play the stimulus as many times as needed, however they nearly always respond after the first listening. The experiment was divided into 20 sessions of 500 trials each, separated with breaks and completed over 4 days. The total length of the experiment (10,000 stimuli plus cognitive and phonological tests) was approximately 4 hours. Data from all participants are available at https://zenodo.org/record/21808 and https://zenodo.org/record/19129.

Auditory Classification Images

We previously described a method for deriving individual Auditory Classification Images (ACIs) for a closed-set categorization task in noise [33,46]. This method is inspired from similar works in the visual domain [34–36]. The ACI calculation has three steps. First a cochleogram is generated for each sound stimulus (54 frequency steps spaced quasi-logarithmically and 81 time steps). For each trial i, time-frequency bins of the cochleogram are treated as a vector of predictors and denoted $\underline{S_i}$. Second, data are randomly divided into 10 sets of 1000 trials that will be assigned to the test set or training set during cross-validation. ACIs are derived using a regularized logit regression between the physical properties of the stimulus $\underline{S_i}$ and the corresponding response of the participant r_i [41]. The resulting vector of parameters $\underline{\beta}$ can be thus seen as a weighting function of the cochleogram. Here we implemented a smoothing constraint penalizing abrupt variations in the ACI, with a level of

smoothing λ determined by a 10-fold cross-validation [51]. For each value of λ, 10 ACIs are estimated, by each time putting aside one set of 1000 trials and using the remaining 9000 trials as training set. Once the ACIs are obtained, their generalizability is measured in terms of cross-validated deviance (CVD) and cross-validation rate (CVR) by predicting data that were not used in the estimation (test set). During this process, proportions of correctly and incorrectly categorized trials are equated in each training set and each test set to make sure that the performance level of the participant does not impact the estimation or evaluation of the models. Third, the level of smoothing λ yielding the lowest mean CVD over all participants is selected and ACIs are re-computed on the complete datasets (10,000 trials) using this value of lambda.

Statistical analyses

Participant's listening strategies were compared on a group basis. Two aspects were investigated: whether the weighting of time-frequency information was similar between dyslexics and control participants, and whether the individual listening strategies were more specific in one group. Correspondingly, two types of statistical analyses were performed.

Firstly, we used a cluster-based non parametric test to know if there was a significant difference between the ACIs of the two groups. This statistical test is appropriate when dealing with highly dimensional data where the location of the potential effect is unknown [52]. The general procedure is as follows: Clusters of adjacent pixels weighted significantly differently between two conditions are identified by a running t-test. Then a permutation test (5000 randomizations) is performed to determine which of them were unlikely to have occurred by chance. Applied to ACIs, this allows us to detect fine differences in the template of weights between two groups or two conditions [33]. This picture was completed by a classic ROI analysis. As in [46], ROIs contours were defined as clusters of at least 7 adjacent time-frequency bins identified as significant in a running t-test ($p<10^{-10}$). Then the mean weights in each set were separately compared between the two groups with another t-test.

Secondly, the generalizability of each participant's ACI was evaluated by measuring how well it can predict the responses of other participants. For this purpose, each model was characterized not only by its 10-fold CVD and CVR derived during the estimation process ("auto-prediction deviance" and "auto-prediction rate") but also 2N-1 between-subject CVDs ("cross-prediction deviances"). Here, the predictive power is assessed in the same way as before, except that the test set is now taken from another participant's data. A measure of the specificity of a listener's strategy can be obtained by taking the difference between his/her auto-prediction deviance and the mean cross-prediction deviance of his/her data produced by ACIs of the other participants. This value is high when the responses are better predicted by his/her own ACI than those of the other.

3. Results

Cognitive and phonological tests

As a group, dyslexic participants enrolled in this study performed significantly lower than control participants ($p<.05$) on all tests but Raven's ($p=0.436$) and pseudo-words reading ($p=0.105$) tests (see Table 1).

To confirm the reading impairment on an individual basis, we performed an "individual deviance analysis" as described in [8] (see also [7]). This two-step procedure considers a performance as deviant when it exceeds 1.65 S.D. from the mean of a given distribution (the fifth percentile of the normal distribution). For each measured characteristics, the mean and standard deviation were calculated on the control group. Any abnormal performances deviating of more than 1.65 S.D. from the mean were then removed from the group and the mean and standard deviation were recomputed. Finally, deviant dyslexic participants were identified on the basis of these new values, with the same 1.65 S.D. threshold. According to this criterion, all dyslexic participants deviated from the normal range of performance in at least 12 of the 26 characteristics, confirming the diagnosis.

Performances in the main experiment

As expected, dyslexic participants were poorer than normal hearing participants on the main phoneme categorization-in-noise experiment. Although both groups obtained similar correct response rates thanks to the adaptive SNR algorithm (dyslexics: 78.8 % ± 0.4 % S.D.; controls: 78.8 % ± 0.4 % S.D.; $t(34)=0.42$; $p=0.67$), and similar sensitivity (d') as defined in signal detection theory (dyslexics: 1.64 ± 0.04 S.D.; controls: 1.65 ± 0.06 S.D.; $t(34)=0.49$; $p=0.62$), dyslexics performed the task at a +1.31 dB SNR higher than normotypical controls on average (dyslexics: - 10.59 dB ± 1.87 dB S.D.; controls: -11.90 dB ± 1.07 dB SNR; $t(34)=2.58$; $p=0.014$). Furthermore, dyslexic participants responded slower than controls (dyslexics: 1.38 s ± 1.87 s S.D.; controls: 1.28 s ± 0.12 s S.D.; $t(34)=2.38$; $p=0.023$).

To assess the extent of any learning effect we carried out two separate 2-way repeated-measures analyses of variance (ANOVAs) on the SNR and response time data. The results show that participants' performances improved over the course of the experiment (Figure 1): we obtained significant effects of session number ($F(19,34)=16.63$; $p<0.001$) and group ($F(1,34)=6.68$; $p<0.05$) on SNR, and significant effects of session number ($F(19,34)=33.42$; $p<0.001$) and group ($F(1,34)=5.65$; $p<0.05$) on RT. In both cases interaction effects were not significant, suggesting that the magnitude of learning effect is equivalent in the two groups.

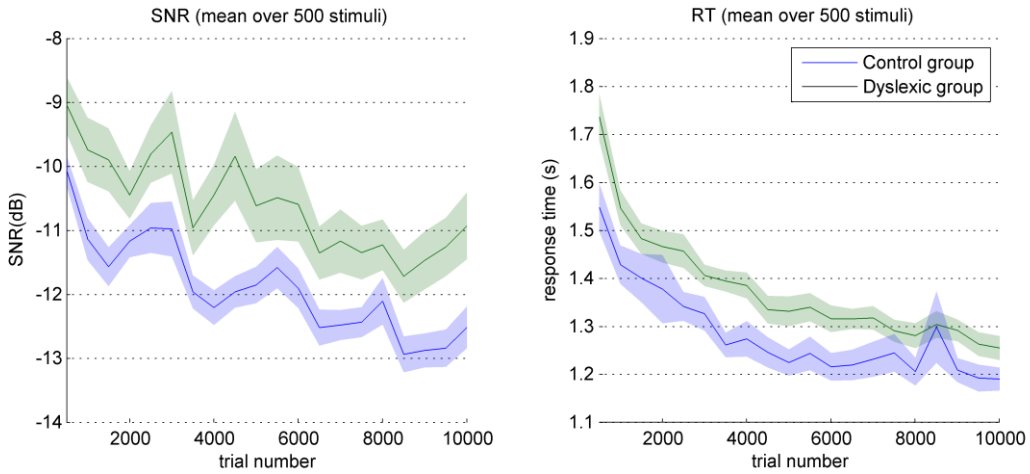


Figure 1. Evolution of performances over the course of the experiment. Mean SNR (left) and mean RT (right) by sessions of 500 trials, for the two groups. Shaded regions denote s.e.m. over participants.

ACIs

ACIs were calculated for each dyslexic participant and compared with those of their average reader peers. Individual images are shown in Figure 2. For each participant, the quality of the ACI was assessed by its cross-validation rate (auto-prediction rate), ranging from 55.5% to 76.9%.

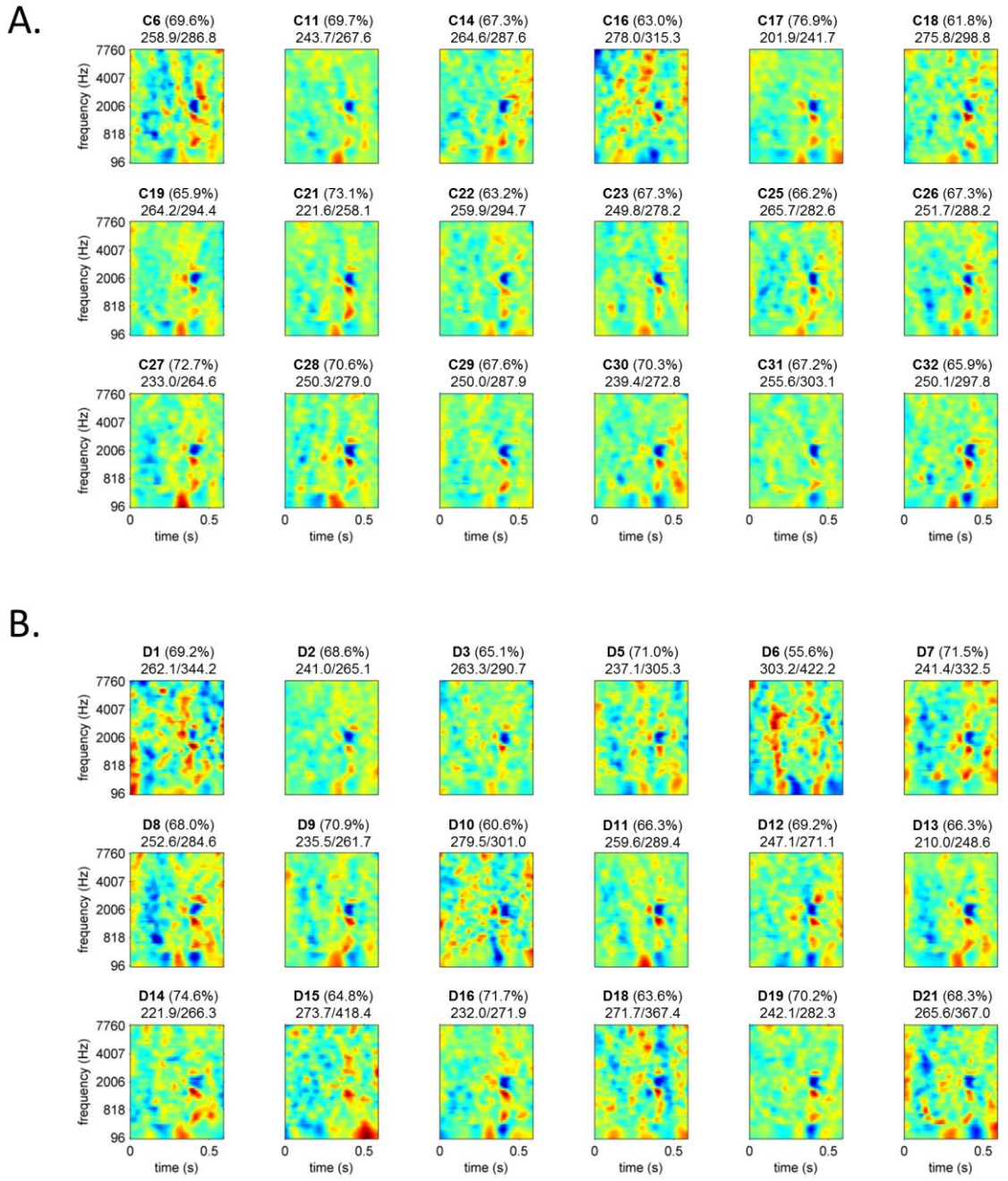


Figure 2. Individual ACIs for all control (A.) and dyslexic (B.) participants. For each ACI the auto-prediction rate is given (in brackets), followed by the auto- and cross- prediction deviances.

Generally, they all shared a similar pattern of weight despite some inter-individual variability. This common pattern becomes clearer when considering the mean ACI over all participants (Figure 3A). As already noticed in a previous study [46], the most consistently weighted areas ($p<10^{-10}$) are the onsets of the F1 F2 and F3 in the second syllable (Figure 3A).

Our first aim being the comparison of the two groups, we averaged separately the dyslexics' and controls' ACIs to obtain two group-ACIs (Figure 3B). The striking similarity between them is corroborated by a cluster-based non-parametric test eliciting no significant differences ($p>0.05$ for all clusters, Figure 3B).

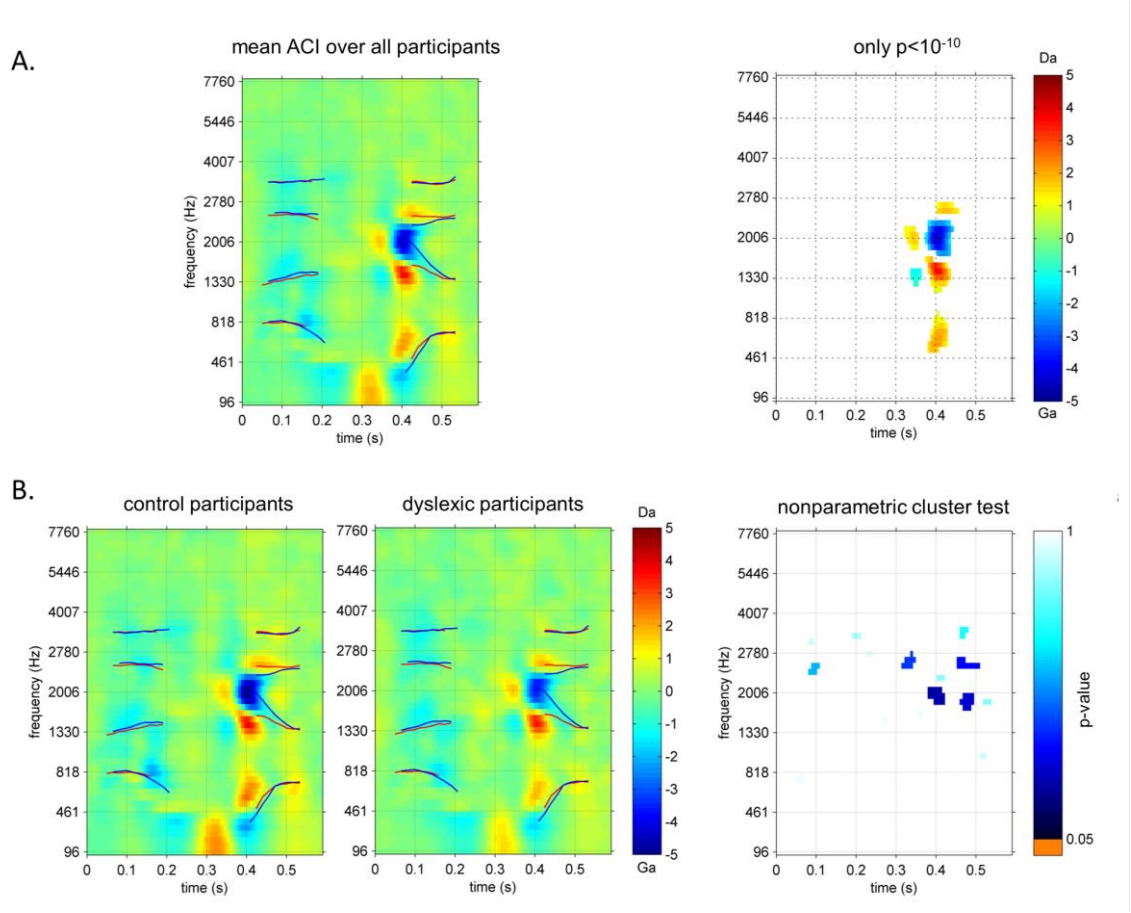


Figure 3. Diagram of the group-analysis of ACIs used in this study. A. mean ACI over all participants (left) and same ACI with all non-significant pixels ($p>10^{-10}$) plotted in white (right panel) defining the regions used in the ROI analysis. B. mean ACIs for the control (left panel) and dyslexic (center panel) groups and output of the cluster-based non-parametric test (right panel). In the ACIs, lines correspond to mean formant trajectories for /alda/ and /aʁda/ (red) and for /alga/ and /aʁga/ (blue).

In order to confirm this result, a less-conservative ROI analysis was performed in the significant regions identified on the pooled images of the two groups, in the previous step. In each of the 6 ROIs, the weights were averaged for each participant, and then compared using a non-paired t-test. The central negative region appeared to be significantly less weighted by dyslexic than control participants ($p=0.009$), while all other differences were non-significant. The characteristics of the ROIs are summarized in Table 2.

Set	size (pxl)	Centroid		Extent		Correspondence with formants	Bias towards	Set weights				
		t (ms)	f (Hz)	t (ms)	f (Hz)			C group (mean)	C group (SD)	D group (mean)	D group (SD)	t-test (D vs. C)
#1	16	340	1967	29	329	onset F2/F3, 2nd syllable	'da'	0.0049	0.0021	0.0051	0.0023	0.79
#2	34	406	1380	58	425	onset F2, 2nd syllable	'da'	0.0167	0.0049	0.0161	0.0044	0.73
#3	33	406	665	44	366	onset F1, 2nd syllable	'da'	0.0137	0.0050	0.0103	0.0054	0.06
#4	14	427	2557	51	138	onset F3, 2nd syllable	'da'	0.0051	0.0014	0.0047	0.0019	0.44
#5	10	350	1347	22	166	onset F2, 2nd syllable	'ga'	-0.0028	0.0010	-0.0024	0.0015	0.27
#6	48	409	1977	66	549	onset F2/F3, 2nd syllable	'ga'	-0.0358	0.0061	-0.0263	0.0132	**0.009***

Table 2. Summary of the characteristics of all sets of weights, sorted by bias and latency.

These mitigated results encouraged us to look more closely at potential individual strategies which may have obscured the picture. To this end, we measured how well each participant's data are predicted by his/her own ACI and by those of the others. This resulted in two measures, respectively the auto- and cross-prediction deviances. Auto-prediction deviances did not differ significantly between the two groups (dyslexics: 252.2 ± 22.5 S.D.; controls: 250.8 ± 18.6 S.D.; t(34)=0.21; p=0.84), but cross-prediction deviances are significantly higher for dyslexic participants (dyslexics: 310.5 ± 53.1 S.D.; controls: 283.3 ± 17.7 S.D.; t(34)=2.06; p=0.049). It is clear from the plot of the cross-prediction deviance against the auto-prediction deviance (Figure 4) that the first is always higher than the second. This stems from the fact that there is a part of each participant's data that can be accurately predicted by his own ACI only, which captures some particularities of the strategy.

Our main interest was therefore on the difference between these two measures, viewed here as an indicator of the specificity of the listener's strategy (Figure 4). On average, the result was dissimilar for the two groups (dyslexics: 58.3 ± 38.0 S.D.; controls: 32.5 ± 8.2 S.D.; t(34)=2.81; p=0.008). This effect was mainly due to a little subgroup of 7 dyslexic participants with high values (participants D1, D5, D6, D7, D15, D18 and D21), the other being in the normal range. The specificity measure is strongly correlated with the mean SNR at which participants performed the task ($r(34)=-0.63$, $p=4.10^{-5}$), even when restricting the calculus to dyslexics only to avoid any group effect ($r(16)=-0.72$, $p=8.10^{-4}$).

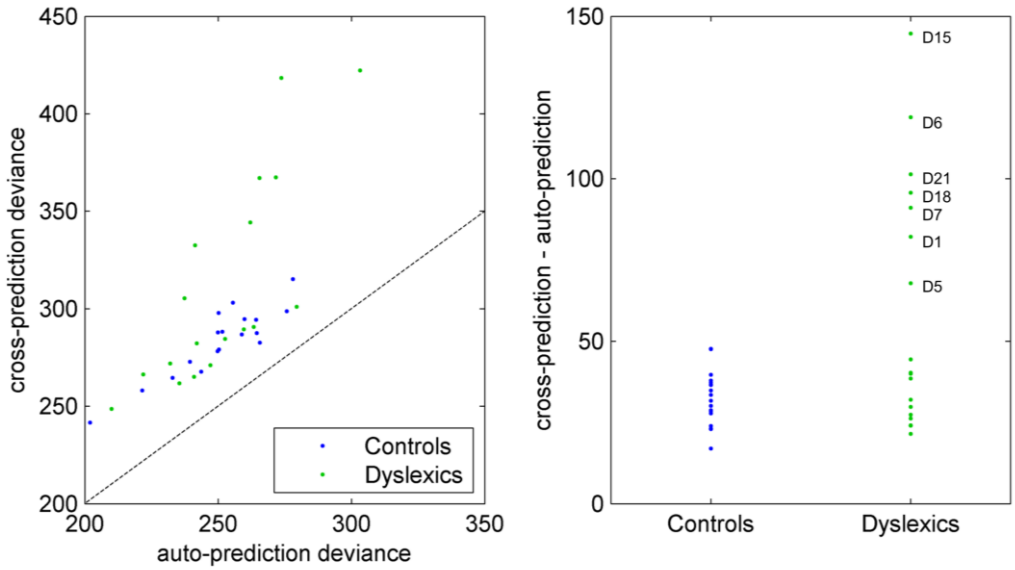


Figure 4. Auto- and cross-predictions for all participants. Left panel: cross-prediction deviance as a function of auto-prediction deviance. The dotted line indicates the cross-prediction = auto-prediction line. Right panel: difference between cross- and auto-prediction deviances for the two groups.

In the group comparison Figure 3, SNR acts as a confounding variable because dyslexics performed the task in significantly lower levels of background noise. Therefore, we additionally compared two subgroups of equivalent SNRs. The 7 outliers previously identified, corresponding to the lower-performing dyslexics, were excluded from the dyslexic group. Additionally, 7 control participants were randomly discarded to result in two subgroups (N=11) of equivalent SNR. A cluster-based comparison of the ACIs of the two groups was performed, revealing a significant cluster on the onset of F3 in the first syllable ($p=0.02$) and a cluster approaching significance on the onset of F4 ($p=0.061$) (Figure 5).

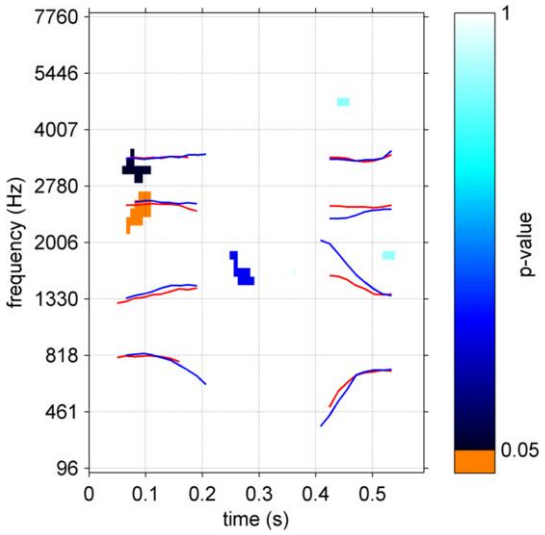


Figure 5. Comparison between the dyslexic and control subgroups matched in SNR (N=11). Output of the cluster-based non-parametric test. Lines correspond to mean formant trajectories for /alda/ and /aʁda/ (red) and for /alga/ and /aʁga/ (blue).

4. Discussion

Unlike previous studies directly contrasting the performances of dyslexic and normotypical participants in different verbal or non-verbal tasks, this work focused on the estimation and comparison of listening strategies in the two groups. For this purpose we used the recently developed ACI technique, a psychophysical tool that has proven to be efficient for identifying the acoustic cues used in a phoneme categorization task in noise. This offered us an insight into the roots of the phonological impairment in dyslexia and enabled us to visualize the compensatory strategies that might be developed to overcome these difficulties.

Here the task was a simple /da/-/ga/ discrimination with two phonetic contexts, /al/ or /aʁ/. The noise level was adapted online on an individual basis to target the 79% correct point. Each participant completed 10,000 trials and special care was taken to ensure that they were not subject to fatigue of weariness. Data from one group of N=18 participants with a prior history of developmental dyslexia, and from a control group of N=18 average readers matched in age, nonverbal IQ and handedness, were analyzed.

As expected, dyslexic participants enrolled in this study obtained performances well below the normal range in the phonological tasks, as a group (see Table 1) but also on an individual basis as indicated by a deviance analysis. This, combined with their normal scores in the cognitive tests and their significant deficit in attention, confirmed the diagnosis of dyslexia. Coherently, the results of the main experiment clearly show that the two groups are dissimilar in their processing of the stimuli: compared to normo-typical participants, dyslexic listeners performed the task with a gap of 1.31 dB SNR and a delay of 100 ms for the same correct response rate (Figure 1). Our first aim in this study was to explore the group-ACIs to identify correlates of this behavioral deficit in the acoustic cues used.

ACIs estimated for all participants are relatively coherent, resulting in a well-defined average pattern of weights in the overall ACI (Figure 3). Two critical time-frequency regions, composed of highly positive (red) or negative (blue) weights, stand out clearly against the non-significant white background (threshold arbitrarily set to $p>10^{-10}$). They are both located around 400 ms, i.e. approximately at the beginning of the 2nd syllable, on which the phonetic decision is made. In these two regions, the presence of noise consistently interferes with the categorization of the stimulus. When we match their time-frequency positions with the acoustical content of the stimuli (formant trajectories symbolized as lines in Figure 3), it appears that participants use two distinct acoustic cues for performing the task: namely the combined onsets of the 2nd and 3rd formants, and the onset of the 1st formant. This observation is coherent with the literature [53,54], and reproduces the results of two previous ACI studies with the same Alda/Alga/Arda/Arga task performed by normotypic [33] or musician [46] listeners.

However this apparent similarity between the various groups of listeners can possibly mask more subtle differences in the listening strategies employed. A tempting explanation for the poor performances of dyslexic participants is that their phonological representations are somewhat noisy, or not precisely attuned to phonetic categories. This impairment might result here in a less efficient weighting of the acoustic cues in their ACIs than in the normotypical ones. Accordingly, we compared the ACIs for the dyslexic and control groups, using a cluster-based nonparametric test. Unfortunately, we were not able to detect a significant difference in ACIs of the two groups of listeners. Even a less conservative analysis showed that only one among 6 tested ROIs, corresponding to the central negative cue, was weighted differently in the two groups.

This null result could be due to the pattern being highly variable at the individual level. Indeed when looking more closely at the non-averaged ACIs Figure 3, we notice an important heterogeneity, especially in the dyslexic group, which may reflect specificities of each participant's listening strategy, and/or a certain amount of noise due to the estimation process itself. In order to disentangle between these two possible causes, we calculated the individual auto- and cross-prediction deviances, a measure of the amount of error when predicting one participant's data by his/her own ACI or by those of the others, respectively. The difference between the two predictions directly relates to the amount of the participant's data accurately predicted only by his/her own ACI and mispredicted by the others. Therefore it is an indicator of how the estimated ACI reflects the specificities of the listener's strategy.

Our rationale was the following: if the observed variability between the dyslexics' ACIs was primarily due to the presence of individual strategies, as hypothesized earlier, the absolute difference between auto- and cross-prediction deviances should be higher in the dyslexic group than in the control group. On the contrary, if the ACIs in the dyslexic group are just less accurately estimated (due for example to the higher SNR in their stimuli), this would mainly impact the auto-prediction. In this case the absolute difference between auto- and cross-prediction deviances should be lower in the dyslexic group than in the control group.

The measured cross-prediction deviance was always higher than the auto-prediction deviance (i.e. data points are above the dotted line on Figure 4A) indicating that the

participants' responses are better predicted by their own ACI, which captures some idiosyncratic aspects of the processing. Moreover the difference between auto- and cross-prediction deviances is larger in the dyslexic group than in the control group. This result is in line with our hypothesis that participants with dyslexia show individual, less efficient, weighting strategies in the task. However, according to Figure 4, this observation does not hold for the group as a whole but it is only due to a subgroup of 7 dyslexics with very high "specificity". Indeed the distribution of weights inside their ACIs suggests a specific preference in this population for other auditory cues than those preferentially exploited by control participants. The most striking effect is observed for participant D15, whose ACI does not show the usual negative cluster of weights around 2500 Hz, nor the low-frequency cue. Consequently, his/her responses are poorly predicted by the other ACIs (cross-prediction deviance = 418.4). However the auto-prediction deviance within the normal range (= 273.7) assesses that this ACI can generalize to unseen data from participant D15, and therefore that it accurately models the underlying process. The same holds for participants D1, D5, D6, D7, D18 and D21, corresponding to the most scattered distributions of perceptual weights.

The finding that some dyslexics in our sample used individual listening strategies is coherent with the hypothesis that dyslexia is related to an impairment of phonological representation. However, the specificity measure is strongly correlated with the mean SNR at which participants performed the task. Furthermore we did not find any robust correlation with the tests in the cognitive and phonological battery, even after grouping the tests in 5 composite scores [7] reflecting phonological awareness ($r(16)=0.01$, $p=0.95$), literacy ($r(16)=0.02$, $p=0.92$), short term memory ($r(16)=-0.44$, $p=0.06$), nonverbal IQ ($r(16)=0.35$, $p=0.15$), and attention ($r(16)=-0.06$, $p=0.82$). Therefore, it is not clear at this point whether the strategy change in the dyslexic subgroup has caused their lower performances in the categorization task or, conversely, if the dissimilarities observed in their ACIs are a consequence of the gap in SNR. Indeed we have previously demonstrated that the level of noise influences the processing of stimuli [32] (it should be noted however that in this study we only evidenced a change in cue weighting, not in the strategy as it is the case here). In order to clarify this issue, it would be very useful to compare the ACIs of the dyslexic group with those of a second control group, matched in reading-level rather than in chronological age.

Finally, we excluded the subgroup of 7 lower-performing participants to explore in further details the group differences between the ACIs of dyslexics and average readers completing the task with equivalent SNR. Our hypothesis was that the 11 remaining high-performing dyslexics successfully developed compensatory strategies for their phonologic impairment, yielding normal recognition in the phoneme in noise categorization task. A new cluster-based comparison (N=11) revealed that they indeed have a slightly different strategy than controls, relying more on the onsets of F3 and F4 in the first syllable (Figure 5). We assume that these regions correspond to a fine acoustic cue, specific to the utterances we used as targets. To reach the same level of performances than their average reader peers, dyslexics seems to be using a slightly different listening strategy, involving the "classic" F1, F2 and F3 onsets from the syllable to be categorized, but also two anticipatory cues.

The interpretation of these cues is not clear, however. They may correspond to subtle differences in the amplitude or timing of the F3 and F4 formant onsets. According to the allophonic perception theory [15,16,30], dyslexic individuals demonstrate an excessive sensitivity to non-linguistic information in the acoustic signal. Therefore, they may be able to extract allophonic cues to build their compensatory strategy, relying on the redundancies in our speech targets. On the ACI these two cues appear as small but significant clusters of negative weights, suggesting that these cues may not be used for all trials. One possibility is that these secondary cues affect the decision only when the primary cues are ambiguous [55].

This finding that high-performing dyslexics used allophonic cues can be linked to a series of recent neuroimaging studies revealing that, even when dyslexics show normal behavioral responses in a speech-in-noise task [22], or in a phoneme categorization task [29,30], their deficit still manifest in the form of enhanced neurophysiological activity. This has been proposed by the authors as a demonstration of the less efficient strategies dyslexics employ to compensate for their impaired phonological processing. In the present study we showed that these strategies can be revealed by the purely behavioral ACI methodology. For our subgroup of 11 dyslexics performing as well as average readers in a phoneme categorization task in noise, the extraction of allophonic cues requires additional cognitive resources, potentially leading to increased neural activations and to stronger mental fatigue.

Two limitations of this study must be highlighted here. First, the last finding relies on a comparison between two groups of N=11 participants, and therefore calls for a replication with a larger sample. Second, one general constraint of the ACI technique is the large amount of trials required for the estimation (10,000 categorizations per participant in the present study), and the limited number of targets. As a consequence, listeners might use non-speech strategies for performing the task, which casts doubts on the authenticity of the acoustic cues revealed in this way. That does not seem to be the case for the control group, as their results are consistent with the literature. However one can ask whether the alternative strategy employed by dyslexic participants is not due to their difficulties in such long and repetitive tasks.

5. References

1. Castles A, Coltheart M. Varieties of developmental dyslexia. Cognition. 1993;47: 149 – 180.

2. Manis FR, Mcbride-Chang C, Seidenberg MS, Keating P, Doi LM, Munson B, et al. Are speech perception deficits associated with developmental dyslexia? J Exp Child Psychol. 1997;66: 211–235.

3. Tallal P. Auditory temporal perception, phonics, and reading disabilities in children. Brain Lang. 1980;9: 182–198.

4. Stein J. Dyslexia: the Role of Vision and Visual Attention. Curr Dev Disord Rep. 2014;1: 267–280.

5. Fawcett AJ, Nicolson RI, Dean P. Impaired performance of children with dyslexia on a range of cerebellar tasks. Ann Dyslexia. 1996;46: 259–283.

6. Boets B, De Smedt B, Cleuren L, Vandewalle E, Wouters J, Ghesquière P. Towards a further characterization of phonological and literacy problems in Dutch-speaking children with dyslexia. Br J Dev Psychol. 2010;28: 5 – 31.

7. Law JM, Vandermosten M, Ghesquiere P, Wouters J. The relationship of phonological ability, speech perception, and auditory perception in adults with dyslexia. Front Hum Neurosci. 2014;8: 482.

8. Ramus F, Rosen S, Dakin SC, Day BL, Castellote JM, White S, et al. Theories of developmental dyslexia: insights from a multiple case study of dyslexic adults. Brain J Neurol. 2003;126: 841–865.

9. Boets B, Beeck HPO de, Vandermosten M, Scott SK, Gillebert CR, Mantini D, et al. Intact But Less Accessible Phonetic Representations in Adults with Dyslexia. Science. 2013;342: 1251–1254.

10. Ramus F. Neuroimaging sheds new light on the phonological deficit in dyslexia. Trends Cogn Sci. 2014;18: 274–275.

11. Ramus F, Szenkovits G. What phonological deficit? Q J Exp Psychol 2006. 2008;61: 129–141.

12. Lehongre K, Morillon B, Giraud A-L, Ramus F. Impaired auditory sampling in dyslexia: further evidence from combined fMRI and EEG. Front Hum Neurosci. 2013;7.

13. Lehongre K, Ramus F, Villiermet N, Schwartz D, Giraud A-L. Altered low-γ sampling in auditory cortex accounts for the three main facets of dyslexia. Neuron. 2011;72: 1080–1090.

14. Elbro C. Early linguistic abilities and reading development: A review and a hypothesis. Read Writ. 1996;8: 453–485.

15. Bogliotti C, Serniclaes W, Messaoud-Galusi S, Sprenger-Charolles L. Discrimination of speech sounds by children with dyslexia: Comparisons with chronological age and reading level controls. J Exp Child Psychol. 2008;101: 137–155.

16. Serniclaes W, Van Heghe S, Mousty P, Carré R, Sprenger-Charolles L. Allophonic mode of speech perception in dyslexia. J Exp Child Psychol. 2004;87: 336–361.

17. Ziegler JC, Castel C, Pech-Georgel C, George F, Alario F-X, Perry C. Developmental dyslexia and the dual route model of reading: simulating individual differences and subtypes. Cognition. 2008;107: 151–178.

18. Boets B, Ghesquière P, van Wieringen A, Wouters J. Speech perception in preschoolers at family risk for dyslexia: relations with low-level auditory processing and phonological ability. Brain Lang. 2007;101: 19–30.

19. Brady S, Shankweiler D, Mann V. Speech perception and memory coding in relation to reading ability. J Exp Child Psychol. 1983;35: 345–367.

20. Ziegler JC, Pech-Georgel C, George F, Lorenzi C. Speech-perception-in-noise deficits in dyslexia. Dev Sci. 2009;12: 732–745.

21. Dole M, Hoen M, Meunier F. Speech-in-noise perception deficit in adults with dyslexia: Effects of background type and listening configuration. Neuropsychologia. 2012;50: 1543–1552.

22. Dole M, Meunier F, Hoen M. Functional correlates of the speech-in-noise perception impairment in dyslexia: An MRI study. Neuropsychologia. 2014;60: 103–114.

23. Boets B, Wouters J, van Wieringen A, Ghesquière P. Auditory temporal information processing in preschool children at family risk for dyslexia: relations with phonological abilities and developing literacy skills. Brain Lang. 2006;97: 64–79.

24. Banai K, Nicol T, Zecker SG, Kraus N. Brainstem timing: implications for cortical processing and literacy. J Neurosci Off J Soc Neurosci. 2005;25: 9850–9857.

25. Chandrasekaran B, Kraus N. The scalp-recorded brainstem response to speech: neural origins and plasticity. Psychophysiology. 2010;47: 236–246.

26. Hornickel J, Skoe E, Nicol T, Zecker S, Kraus N. Subcortical differentiation of stop consonants relates to reading and speech-in-noise perception. Proc Natl Acad Sci U S A. 2009;106: 13022–13027.

27. White-Schwoch T, Woodruff Carr K, Thompson EC, Anderson S, Nicol T, Bradlow AR, et al. Auditory Processing in Noise: A Preschool Biomarker for Literacy. PLoS Biol. 2015;13: e1002196.

28. Messaoud-Galusi S, Hazan V, Rosen S. Investigating speech perception in children with dyslexia: is there evidence of a consistent deficit in individuals? J Speech Lang Hear Res JSLHR. 2011;54: 1682–1701.

29. Noordenbos MW, Segers E, Serniclaes W, Verhoeven L. Neural evidence of the allophonic mode of speech perception in adults with dyslexia. Clin Neurophysiol Off J Int Fed Clin Neurophysiol. 2013;124: 1151–1162.

30. Noordenbos MW, Segers E, Serniclaes W, Mitterer H, Verhoeven L. Allophonic mode of speech perception in Dutch children at risk for dyslexia: A longitudinal study. Res Dev Disabil. 2012;33: 1469–1483.

31. Tallal P, Stark RE. Speech acoustic-cue discrimination abilities of normally developing and language-impaired children. J Acoust Soc Am. 1981;69: 568–574.

32. Varnet L, Knoblauch K, Meunier F, Hoen M. Using auditory classification images for the identification of fine acoustic cues used in speech perception. Front Hum Neurosci. 2013;7: 865.

33. Varnet L, Knoblauch K, Serniclaes W, Meunier F, Hoen M. A Psychophysical Imaging Method Evidencing Auditory Cue Extraction during Speech Perception: A Group Analysis of Auditory Classification Images. PLoS ONE. 2015;10: e0118009.

34. Knoblauch K, Maloney LT. Estimating classification images with generalized linear and additive models. J Vis. 2008;8: 10.1–19. doi:10.1167/8.16.10

35. Knoblauch K, Maloney LT. Modeling Psychophysical Data in R. Springer Science & Business Media; 2012.

36. Mineault PJ, Barthelmé S, Pack CC. Improved classification images with sparse priors in a smooth basis. J Vis. 2009;9: 17.1–24.

37. Murray RF. Classification images: A review. J Vis. 2011;11. doi:10.1167/11.5.2

38. Gold JM, Murray RF, Bennett PJ, Sekuler AB. Deriving behavioural receptive fields for visually completed contours. Curr Biol CB. 2000;10: 663–666.

39. Nagai M, Bennett PJ, Rutherford MD, Gaspar CM, Kumada T, Sekuler AB. Comparing face processing strategies between typically-developed observers and observers with autism using sub-sampled-pixels presentation in response classification technique. Vision Res. 2013;79: 27–35

40. Pritchett LM, Murray RF. Classification images reveal decision variables and strategies in forced choice tasks. Proc Natl Acad Sci. 2015;112: 7321–7326.

41. Schönfelder VH, Wichmann FA. Sparse regularized regression identifies behaviorally-relevant stimulus features from psychophysical data. J Acoust Soc Am. 2012;131: 3953–3969.

42. Sekuler AB, Gaspar CM, Gold JM, Bennett PJ. Inversion leads to quantitative, not qualitative, changes in face processing. Curr Biol CB. 2004;14: 391–396.

43. Calabrese A, Schumacher JW, Schneider DM, Paninski L, Woolley SMN. A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. PloS One. 2011;6: e16104.

44. Woolley SMN, Gill PR, Fremouw T, Theunissen FE. Functional Groups in the Avian Auditory System. J Neurosci. 2009;29: 2780–2793.

45. Woolley SMN, Gill PR, Theunissen FE. Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. J Neurosci Off J Soc Neurosci. 2006;26: 2499–2512.

46. Varnet L, Wang, T, Peter C, Meunier F, Hoen M. How musical expertise shapes speech perception: evidence from auditory classification images. Sci Rep. in press;

47. Fan J, McCandliss BD, Sommer T, Raz A, Posner MI. Testing the efficiency and independence of attentional networks. J Cogn Neurosci. 2002;14: 340–347.

48. Gola-Asmussen C, Lequette C, Pouget G, Rouyer C, Zorman M. ECLA-16+ Evaluation des Compétences de Lecture chez l'Adulte de plus de 16 ans [Internet]. Grenoble: Laboratoire Cogni-Sciences, IUFM de Grenoble; 2010.

49. Lefavrais P. Le test de L'Alouette. Edition du Centre de Psychologie Appliquée. Paris, France; 1967.

50. Levitt H. Transformed up-down methods in psychoacoustics. J Acoust Soc Am. 1971;49: Suppl 2:467+.

51. Wu MC-K, David SV, Gallant JL. Complete functional characterization of sensory neurons by system identification. Annu Rev Neurosci. 2006;29: 477–505.

52. Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods. 2007;164: 177–190.

53. Mann VA. Influence of preceding liquid on stop-consonant perception. Percept Psychophys. 1980;28: 407–412.

54. Viswanathan N, Magnuson JS, Fowler CA. Compensation for coarticulation: disentangling auditory and gestural theories of perception of coarticulatory effects in speech. J Exp Psychol Hum Percept Perform. 2010;36: 1005–1015.

55. Serniclaes W, Arrouas Y. Perception des traits phonétiques dans le bruit. Verbum. 1995; 131–144.

12.3. Résumé de l'étude 4

L'objectif de cette dernière étude était d'aborder la problématique du déficit phonologique dans un contexte de dyslexie développementale ainsi que les stratégies mises en place pour le compenser. Notre hypothèse était que ce trouble, observé dans des tâches de catégorisation et d'identification de phonèmes, devait pouvoir être visualisé et quantifié au moyen des ACIs. L'approche suivie ici était identique à celle de l'étude 3 mais appliquée à un groupe de 20 participants diagnostiqués comme dyslexiques et à un groupe de 20 participants contrôles.

Au final, le groupe dyslexique obtint des performances globalement inférieures à celles du groupe contrôle dans la tâche de catégorisation /da/-/ga/ dans le bruit, ayant besoin d'un SNR significativement plus élevé pour parvenir à un même taux de réponses correctes. En revanche, contrairement à nos attentes, la comparaison des ACIs entre les deux groupes n'a pas permis d'identifier une différence robuste qui expliquerait directement cet écart de performances. Une inspection visuelle des ACIs individuelles semblait indiquer une grande variabilité au sein du groupe dyslexique, qui expliquait en partie l'absence de résultat dans la comparaison de groupe. Comme dans l'article 3, nous avons donc complété cette étude par une analyse en prédictions croisées des GLM ajustés à chaque participant. Nous avons pu ainsi mettre en évidence que l'hétérogénéité des ACIs observées n'est pas due à la présence d'un bruit d'estimation mais effectivement à une plus grande diversité des stratégies d'écoute, en particulier pour un sous-groupe de sept participants dyslexiques réalisant la tâche à un SNR élevé. Il n'est cependant pas possible, à ce stade, de déterminer si l'emploi de stratégies individuelles est une cause ou une conséquence des faibles performances dans la tâche considérée. En excluant ces participants pour restreindre la comparaison de groupe aux seuls participants dyslexiques obtenant des SNRs équivalents à ceux du groupe contrôle, nous avons alors pu identifier des différences significatives au sein de l'ACI. Les participants dyslexiques atteignant le niveau de performances standard utilisent donc une stratégie compensatoire consistant à extraire deux indices acoustiques supplémentaires : l'attaque des formants F3 et F4 dans la première syllabe.

13. Résumé des principaux résultats obtenus

L'objectif premier de cette thèse est l'élaboration et la mise à la disposition des chercheurs en psycholinguistique d'un nouvel outil, appelé Images de Classification Auditives (ACI), qui leur permette de « révéler » et d'étudier les indices acoustiques impliqués dans la compréhension de la parole dégradée. Au travers des 4 articles ci-dessus, nous avons eu l'occasion de démontrer, par le calcul et par l'exemple, le potentiel et la fiabilité de cette méthode. Le cadre théorique a été emprunté à une technique développée pour le domaine visuel, les Image de Classification (CI). Ainsi que nous l'espérions, sa traduction dans le domaine auditif a permis de dériver des profils de poids correspondant à la position des régions temps-fréquence cruciales pour une tâche de catégorisation phonémique dans le bruit. À cet effet, nous avons implémenté un paradigme expérimental et un algorithme d'estimation (basé sur un GLM pénalisé). On trouvera dans l'Annexe 4 de cette thèse une *toolbox* MATLAB regroupant les principaux scripts et fonctions utilisés ici. Grâce aux données expérimentales de 70 participants, pour un total d'environ 210 heures de passation, nous avons pu calculer les ACIs correspondant aux catégorisations /ba/-/da/ et /da/-/ga/ (respectivement en contexte /a/ et /al/-/aʁ/). L'ensemble des stimuli employés et des données récoltées est librement disponible sur Internet (voir Annexes 1, 2 et 3).

Ces quatre études ont permis de vérifier un certain nombre d'hypothèses à propos de l'interface acoustico-phonétique, formulées à la suite d'expériences utilisant le paradigme du continuum de parole synthétique. Plutôt que d'isoler une variable supposée cruciale puis de tester son influence sur une mesure de la perception, ce qui induit un biais d'analyse, l'approche des ACIs permet de laisser le choix des variables pertinentes à l'auditeur. Nous avons ainsi pu confirmer que les attaques des formants F2 et F3 constituent effectivement un indice naturel pour la compréhension des occlusives voisées dans le bruit. Mais, de manière inattendue, le calcul des ACIs a également établi que d'autres informations se trouvaient aussi impliquées dans cette catégorisation : des indices anticipatoires (fin du formant F2 dans la syllabe précédente, étude 1) ou simultanés (attaque de F1, études 2, 3 et 4). Ainsi, nous avons pu mettre en relief la pluralité des indices contribuant à la décision phonétique. La redondance de l'information assure ainsi une meilleure robustesse de la compréhension de la parole dégradée, en autorisant une certaine plasticité des stratégies d'écoute que nous avons notamment pu observer dans l'étude 1 (évolution de l'ACI en fonction du SNR). Par ailleurs, la différence entre les ACIs restreintes à l'une ou l'autre des cibles (études 1 et 2) confirme que les traitements sous-jacents sont non linéaires.

Une fois établie la stratégie d'écoute d'un groupe de participants sans trouble phonologique ni entraînement particulier, nous avons cherché dans un second temps à la comparer avec les résultats de participants musiciens experts, d'une part (étude 3), puis dyslexiques, d'autre part (étude 4). Pour les premiers, le renforcement des

capacités auditives se traduit par de meilleures performances dans la tâche de catégorisation phonémique dans le bruit et, au niveau de l'ACI, par une plus grande focalisation de l'attention auditive sur les deux principaux indices acoustiques. Les seconds se divisent en deux sous-groupes : certains participants dyslexiques montrent des performances significativement plus faibles lors de la catégorisation dans le bruit et une forte hétérogénéité dans leurs stratégies d'écoute ; d'autres atteignent le même niveau de performance que le groupe contrôle, en utilisant une stratégie compensatoire basée sur l'extraction de deux indices complémentaires situés dans la première syllabe.

Entre les études 1 et 4 nous avons apporté plusieurs améliorations à notre algorithme de calcul : utilisation d'une représentation biologiquement inspirée, basée sur le cochléogramme et non plus sur le spectrogramme ; optimisation d'un hyperparamètre unique ; ACIs dérivées du stimulus (cible + bruit) plutôt que du bruit seul. Ces modifications successives ont permis d'augmenter l'efficacité de l'algorithme (diminution du nombre d'essais nécessaires au calcul, meilleure sélection du degré de lissage) et la résolution des ACIs (réduction du bruit d'estimation, apparition de détails plus fins). À titre d'illustration, nous représentons Figure 25 deux ACIs calculées à partir des mêmes données en utilisant la première version de l'algorithme et la dernière version. L'optimisation du lissage de l'algorithme le plus récent se montre sans conteste plus efficace. De plus, les régions critiques se détachent de façon plus marquée sur le fond, laissant même apparaître un possible indice supplémentaire sur l'attaque du F1.

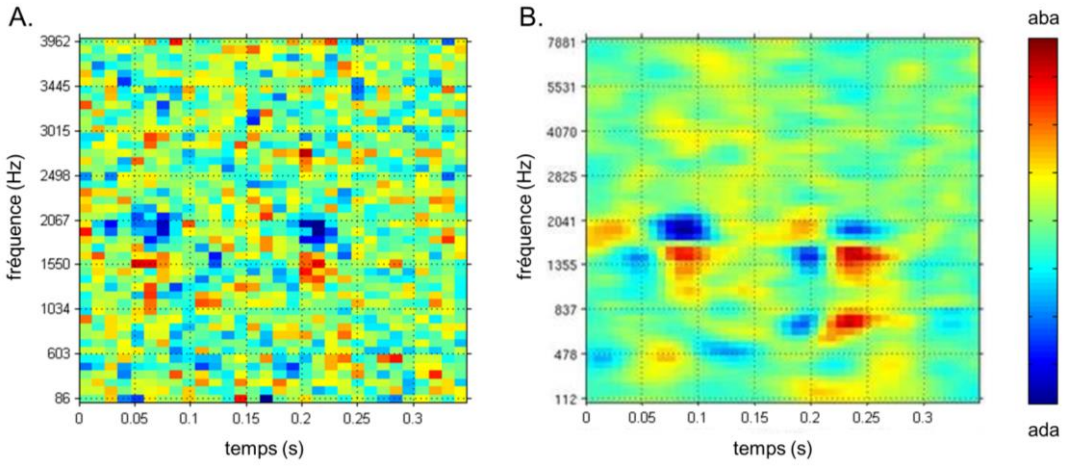


Figure 25 – ACIs correspondant à la catégorisation /aba/-/ada/, pour le participant LV de l'étude 1. A. ACI dérivée en utilisant l'algorithme de l'étude 1 (calcul basé sur le spectrogramme des bruits, 2 hyperparamètres). B. ACI dérivée en utilisant l'algorithme de l'étude 4 (calcul basé sur le cochléogramme des stimuli, 1 seul hyperparamètre).

Le second volet du développement méthodologique visait à la mise en œuvre de tests statistiques pour l'analyse des ACIs, au niveau individuel et au niveau du groupe. Dans les deux cas se pose le problème des comparaisons multiples, aggravé par

l'existence de dépendances entre les pixels temps-fréquence adjacents. Plusieurs tests statistiques ont été mis à contribution au fil de ces quatre études : comparaison avec un ensemble d'ACIs obtenues par randomisation et correspondant à l'hypothèse nulle ; emploi d'une *FDR correction* ; mise en place d'un test non paramétrique basé sur les clusters. Ces divers tests nous ont permis de tirer des conclusions concernant le profil de poids perceptuels (ou les différences de profil entre deux conditions) à l'échelle de la population considérée. Enfin, dans les études 3 et 4, l'examen par validation croisée des capacités de prédiction des modèles nous a permis de distinguer deux sources de variabilité au sein des ACIs : le bruit d'estimation et les spécificités individuelles des stratégies d'écoute.

14. Discussion générale

Pour clore ce manuscrit, nous reprendrons l'ensemble de nos résultats pour les discuter suivant deux axes principaux. Nous reviendrons tout d'abord sur certains aspects de la technique (notamment ses limites théoriques et contraintes expérimentales) qui ne sont pas toujours abordés directement dans les articles précédents mais qui peuvent être éclairés par nos observations et par la littérature s'intéressant à la modalité visuelle. Cet examen nous permettra de mieux cerner le champ d'application propre des ACIs, en comparaison avec les autres méthodes psycholinguistiques destinées à identifier les indices acoustiques impliqués dans une tâche de catégorisation phonémique (partie 14.1). Dans un second temps, nous replacerons les résultats obtenus dans le contexte de la recherche sur l'interface acoustico-phonétique, en en proposant une interprétation plus « spéculative » et, enfin, nous proposerons une réflexion quant au rôle que pourrait tenir la méthode des ACIs dans l'étude du décodage de la parole par le cerveau (partie 14.1).

14.1. Discussion méthodologique

Comme toute expérimentation dans le domaine de la cognition (Tiberghien & Jeannerod, 1995), les études présentées dans cette thèse reposent sur un indicateur (les poids de l'ACI) permettant de mettre en évidence un phénomène caché (la stratégie d'écoute de l'auditeur). La mesure de cet indicateur impose, d'une part, un paradigme expérimental particulier (tâche de catégorisation forcée dans le bruit) et, d'autre part, un cadre statistique précis (modèle de l'observateur linéaire et GLM pénalisé). Les contraintes théoriques et pratiques qui en découlent restreignent nécessairement le champ des questions susceptibles d'être abordées au moyen de cette méthode. La mise en place d'une expérimentation fondée sur les ACIs et l'interprétation des résultats obtenus n'allant pas de soi, il paraît indispensable de mener un examen approfondi des problèmes méthodologiques posés par cette nouvelle technique.

Les limites théoriques de la méthode sont la conséquence directe des hypothèses du modèle statistique de l'observateur linéaire décrit par les équations 7.7 et 7.8. Celui-ci suppose notamment la linéarité des processus et leur stationnarité au cours de l'expérience (Kingdom & Prins, 2010; Wichmann & Hill, 2001). Nous avons également présumé que le profil de poids utilisé est relativement lisse (régularisation par lissage, paragraphe 7.7.2), une hypothèse qui mérite d'être discutée ici. Certaines contraintes pratiques sont également imposées par le protocole expérimental, notamment le choix de la tâche et la durée de l'expérience. Ces différents aspects seront pris en compte par la suite pour tenter de définir un cadre d'application pour la méthode des ACIs.

14.1.1. Non-stationnarité des traitements : fatigue mentale et apprentissage perceptuel

Le nombre élevé d'essais nécessaires au calcul d'une image individuelle constitue une première limitation évidente de la méthode. En pratique, les volontaires ayant pris part à nos études ont accompli chacun 10000 catégorisations phonémiques, ce qui représente environ 3 à 4 heures de passation pour chaque participant. Une telle durée d'expérience pose problème car elle risque de modifier le traitement des stimuli. Cette influence peut se révéler négative pour les performances – c'est le cas d'un effet de fatigue mentale ou de lassitude vis-à-vis d'une tâche longue et répétitive – mais elle peut également être positive si, par exemple, un apprentissage perceptuel se met en place. Nous cherchons ici à éliminer les deux phénomènes car ils enfreignent une hypothèse de notre modèle : la stationnarité des processus tout au long de l'expérience (i.e. l'auditeur ne modifie pas sa stratégie d'écoute au cours de la passation). De plus, ils nous éloignent tous deux de la situation dite « naturelle » de perception de la parole, remettant en question la validité des indices acoustiques dévoilés : s'agit-il des informations réellement utilisées dans la situation courante de perception de la parole ou d'une stratégie différente, acquise au fil de l'expérience ?

À nombre d'essais constant, l'une des premières possibilités pour prévenir la fatigue mentale, envisagée par les développeurs de la méthode des CIs et reprise dans nos études en ACIs, consiste à diviser le temps du travail cognitif en séquences plus courtes entrecoupées de pauses (Gold et al., 2000; Knoblauch & Maloney, 2012; Murray et al., 2002). Une autre solution, que nous avons également implémentée dans le protocole, consiste à « compenser » les pertes d'attention en adaptant continuellement la difficulté de la tâche aux performances immédiates du participant, par exemple au moyen d'un algorithme de *staircase* (Solomon, 2002; Watson & Rosenholtz, 1997). D'autres solutions pourraient encore être explorées, comme l'ajout d'un feedback (Brohé et al., 2012) ou une présentation plus « stimulante » de l'expérience au niveau de l'interface et/ou des consignes.

Le même problème se pose pour contrer un possible apprentissage perceptuel. En effet, il a été démontré que l'intelligibilité de la parole dégradée s'améliore avec l'expérience, par exemple pour les sons noise-vocodés (Davis et al., 2005; Hervais-Adelman et al., 2008) ou prononcés avec un accent étranger (Van Engen & Peelle, 2014). Les passations longues et répétitives, comme celles requises par notre méthode, sont particulièrement propices à engendrer un apprentissage perceptuel. En effet, on a constaté que les participants sont susceptibles de mettre en place des stratégies d'écoute spécifiques aux cibles qu'ils cherchent à détecter. Dans le cadre des CIs, ce phénomène se traduit par un transfert des poids perceptuels d'une région de l'image vers une autre (Gold et al., 2004; Kurki & Eckstein, 2014; Li et al., 2004)[17].

[17] On peut noter cependant, en comparant ces études, que la réallocation des ressources attentionnelles consiste, dans un cas, en une focalisation sur des éléments saillants du stimulus (Gold et al., 2004) et, dans

L'établissement d'un apprentissage perceptuel dans des tâches parolières telles que celles employées dans cette thèse ne semble pouvoir être contrecarré qu'au prix d'une multiplication des cibles et/ou d'une réduction de la durée de la passation. Nous reviendrons sur ce point critique au paragraphe 14.1.5.

Autant il est permis, à terme, d'espérer parvenir à satisfaire la condition de stationnarité des processus, autant leur linéarité reste une hypothèse fondamentalement théorique et impossible à vérifier en pratique.

14.1.2. Non-linéarité des traitements

Comme nous l'avons déjà souligné dans nos articles, les traitements auditifs et cognitifs impliqués par la tâche de catégorisation phonémique dans le bruit sont en grande partie non linéaires (Allen, 2008; Goldstein, 1967; Moore, 2002). Il est donc peu probable que le système considéré dans son ensemble puisse être reproduit fidèlement par un modèle linéaire. C'est pourtant l'une des hypothèses sous-jacentes de l'équation 7.7. D'après celle-ci, le stimulus entrant est tout d'abord transformé (par un ensemble de prétraitements regroupés dans la fonction φ, voir paragraphe 7.5.2), puis converti linéairement en une variable de décision unique. Ici, tous les traitements cognitifs sont donc supposés assimilables à une pondération linéaire. Le même problème se pose dans le domaine visuel et, comme nous l'avons déjà relevé, certains chercheurs ont montré qu'il était possible de mettre cette hypothèse en défaut en comparant les CIs calculées séparément pour une cible et pour une autre (Abbey & Eckstein, 2006; Ahumada & Lovell, 1971; Solomon, 2002).

Il a été reproché à la méthode des ACIs le fait que le modèle de l'observateur linéaire est trop réducteur pour représenter correctement les traitements effectués par le cerveau de l'auditeur. En réponse à cette objection, on peut faire remarquer qu'il ne s'agit pas de proposer ici un modèle complet du système auditif, ni même seulement de l'interface acoustico-phonémique. La méthode des ACIs ne prétend qu'à déterminer les régions temps-fréquence critiques pour distinguer deux phonèmes. Notre argument peut ici être éclairé par une analogie avec la neuro-imagerie. Le raisonnement le plus couramment utilisé dans ce domaine, appelé soustraction cognitive (Friston, 2003), consiste à mesurer les activités cérébrales dans deux conditions expérimentales A et B puis à en prendre la différence pour mettre en évidence un changement significatif d'activation dans une zone définie. Cette méthode permet donc d'isoler une ou plusieurs régions cérébrales critiques, impliquées différemment dans les conditions A et B. Bien que le modèle simple utilisé (ici la soustraction) soit linéaire, les chercheurs ne supposent en aucune façon que le cerveau lui-même opère de manière linéaire. Pareillement, dans le cas des ACIs envisagé ici, même si le GLM ne constitue pas un

l'autre, en une intégration de l'information sur une région plus étendue pour s'approcher de l'observateur idéal (Kurki & Eckstein, 2014; Li et al., 2004).

modèle biologiquement plausible du système auditif, il reste du moins capable de prédire de manière suffisamment fiable les réponses (et les erreurs) d'un observateur. On peut donc raisonnablement penser qu'il se base sur les mêmes éléments d'information pour prendre sa décision.[18]

Ainsi, la non-linéarité des processus impliqués dans le traitement n'est ainsi pas un obstacle rédhibitoire pour la méthode des ACIs. Il convient néanmoins d'observer qu'elle entraîne un certain nombre de biais lors de l'estimation de l'ACI, ainsi qu'une dépendance du résultat relativement à l'ensemble des signaux employés pour l'expérience. Il est donc inexact de dire que cette technique « révèle la stratégie du participant » : rigoureusement parlant, l'ACI correspond à la pondération linéaire modélisant au mieux (d'après les critères du MAP) les réponses de l'auditeur pour l'ensemble des stimuli considérés. En d'autres termes, cette méthode fournit seulement une « projection », dans l'espace des pondérations linéaires, de la stratégie non linéaire du participant. Deux conséquences importantes doivent alors être prises en compte :

(1) des stratégies d'écoute différentes sont susceptibles de produire une ACI identique (Murray, 2011)

(2) inversement, une même stratégie appliquée à des stimuli différents peut conduire à des ACIs différentes (Christianson et al., 2008).

Par conséquent, si les régions temps-fréquence associées à des poids élevés marquent effectivement la présence d'indices acoustiques utilisés par l'auditeur, les valeurs exactes de ces poids doivent néanmoins être interprétées avec précaution. Il est toutefois possible de contourner ces contraintes en utilisant un protocole expérimental particulier : la comparaison des ACIs, estimées sur un même ensemble de sons de parole, entre deux groupes de participants ou deux facteurs contextuels. C'est la procédure que nous avons suivie dans les études 3 et 4.

Enfin, plusieurs propositions ont été faites pour tenter d'enrichir le modèle avec des composantes non linéaires. Nous avons déjà mentionné le prétraitement φ qui peut recouvrir une transformation temps-fréquence ou un passage à l'énergie. À partir de notre deuxième article, nous avons basé l'estimation sur une représentation davantage biologiquement inspirée, le cochléogramme, de manière à incorporer les traitements qui ont lieu au niveau de la cochlée. Il serait envisageable de pousser plus loin ce raisonnement et de dériver les ACIs de représentations « haut-niveau ». La

[18] Comme il a été souligné par Guy Tiberghien, ce type de raisonnement concernant la neuro-imagerie a ensuite été fréquemment utilisé de façon erronée en prétendant localiser dans le cerveau une fonction cognitive impliquée dans la condition A mais pas dans la condition B (Tiberghien, 2007). La même faute logique dans le cas des ACIs conduirait à conclure que les indices acoustiques mis en évidence sont les seules informations utilisées par le système auditif pour la tâche demandée, c'est-à-dire que la suppression du reste du stimulus n'influerait pas sur la réaction des participants. Une telle assertion est bien évidemment fausse : par exemple, le stimulus /aba/ de l'étude 1, limité aux seules attaque et fin de F2, ne serait même pas perçu comme un son de parole. Il serait donc impossible à catégoriser.

représentation multi-résolution pourrait alors constituer un bon candidat, et permettrait de localiser les indices acoustiques dans l'espace des STRFs (voir paragraphe 3.2.4). La littérature concernant les CIs visuelles comporte également quelques tentatives intéressantes de modélisation de l'incertitude de l'observateur concernant la position de la cible en incluant dans le prétraitement un filtre gaussien plus ou moins large (Abbey & Eckstein, 2006; Barth et al., 1999; Murray et al., 2005). Cette idée pourrait être directement transposée à la technique des ACIs pour modéliser une incertitude fréquentielle et/ou temporelle. Une autre voie a été envisagée par certaines équipes de recherches pour sonder les non-linéarités du processus de décision (Joosten & Neri, 2012; Knoblauch & Maloney, 2008; Neri & Heeger, 2002). Partant du constat que l'équation 7.7 n'est plus valable lorsque la transformation effectuée par le système n'est pas linéaire, ils proposèrent d'utiliser une forme de développement en séries de Taylor pour obtenir une approximation au 2ème degré de la variable interne, en ajoutant un terme correspondant à la convolution entre le « gabarit de deuxième ordre » $\underline{B}$ et le carré du stimulus $\underline{S}^2$ (équation 14.1).

$$d = \underline{S}^2 * \underline{B} + \underline{S} * \underline{\beta} + \varepsilon \qquad 14.1$$

En pratique, cette démarche revient à effectuer l'estimation d'une CI de la variance, représentant les non-linéarités de deuxième ordre impliquées dans le traitement. Ici encore, l'étude des ACIs pourrait tirer bénéfice de ce type d'analyses.

14.1.3. Choix du régularisateur

La description de la méthode des ACIs dans la première partie de cette thèse n'envisage qu'un seul type de régularisation, la régularisation par lissage, qui pénalise les CIs abruptes (cf. paragraphe 7.7.2). Dans le cadre des expérimentations menées ici, nous n'avons appliqué que ce régularisateur, caractérisé par 1 ou 2 hyperparamètres. Cette technique, combinée avec un MAP, a l'avantage de ne pas imposer arbitrairement le degré de filtrage appliqué à l'ACI. Néanmoins, d'autres formes de régularisations existent (Wu et al., 2006), par exemple la régularisation parcimonieuse (*sparseness prior*), aussi appelée Lasso ou régularisation L1, qui suppose que la plupart des poids perceptuels sont nuls (Schönfelder & Wichmann, 2012). Le choix d'une régularisation particulière dépend donc de nos connaissances a priori concernant le gabarit sous-jacent. De prime abord, ce choix a priori peut sembler en contradiction avec notre objectif, qui était de caractériser les règles de décisions appliquées par l'auditeur de la manière la plus objective possible. Mais on peut remarquer que l'absence de régularisation correspond elle aussi à l'introduction d'un a priori dans l'équation : celui de l'indépendance des poids de l'ACI.

Le choix d'un régularisateur particulier traduit donc les attentes de l'expérimentateur concernant l'ACI. Il dépend également de l'objectif de l'étude : chaque type de régularisation introduisant un biais particulier dans l'estimation, il est important

de s'assurer que ce biais ne modifie pas les caractéristiques testées. Par exemple, dans le cadre d'une étude où seule la position des indices acoustique dans l'espace temps-fréquence nous intéresserait (et non leur forme ou leur pondération exacte), il serait judicieux d'appliquer une régularisation parcimonieuse sur une base gaussienne (*sparseness prior on a smooth basis*). Cet a priori très fort – selon lequel le gabarit estimé est composé d'un nombre restreint de noyaux gaussiens plus ou moins larges (Mineault et al., 2009) – permettrait alors de réduire sensiblement le nombre d'essais requis pour l'estimation. En revanche, il contraint fortement le profil de l'ACI résultante.

Outre les limites théoriques de la méthode, qui découlent du modèle de l'observateur linéaire et du type d'estimation utilisé, d'autres contraintes doivent également être prises en compte. En effet, le calcul d'une ACI impose un paradigme expérimental particulier qui n'est pas nécessairement compatible avec l'étude de la compréhension de la parole.

14.1.4. Choix de la tâche

La très grande majorité des études sur les CIs visuelles s'appuie sur un paradigme de choix forcé dans lequel un nombre limité de cibles différentes sont présentées au participant qui n'a le choix qu'entre un nombre limité de réponses prédéfinies (Ahumada, 2002; Gold et al., 2000; Thomas & Knoblauch, 2005). Comme nous l'avons montré, ce dispositif expérimental est également transposable dans la modalité auditive : les 4 études réalisées dans le cadre de cette thèse reposent sur une tâche de catégorisation phonémique dans le bruit mettant en jeu un ensemble très restreint d'enregistrements paroliers (2 ou 4).

Par conséquent, ce choix théorique ne correspond pas à la situation naturelle de compréhension de la parole qui consiste la plupart du temps en une identification ouverte de phonèmes et/ou de mots. Aucun cadre statistique n'a pour l'instant été proposé pour estimer une CI dans ce genre de situations. En revanche, d'autres types de tâches ont été explorés dans le domaine visuel. Nous les répertorions ici en tentant d'esquisser un parallèle avec la modalité auditive :

- Le choix forcé à deux alternatives (*2 alternatives forced choice*, 2AFC) (Abbey & Eckstein, 2001, 2002, 2006; Pritchett & Murray, 2015; Solomon, 2002). À chaque essai, deux images bruitées sont présentées au participant qui doit identifier celle contenant le signal cible. Le 2AFC est un paradigme déjà utilisé dans le domaine auditif et sa transposition aux ACIs ne devrait donc pas poser de difficultés particulières. Ce type de tâche suggère le calcul d'ACIs « de discrimination » (plutôt que d'ACIs « de catégorisation »).

- L'identification à alternatives multiples (Dai & Micheyl, 2010; Murray, 2011) pour laquelle plus de deux réponses sont possibles. L'expérience (simulée) de Dai

et Micheyl porte, par exemple, sur 5 cibles auditives de structures harmoniques différentes. Un tel paradigme est extrêmement intéressant dans le cadre d'une étude de la compréhension de la parole. En effet, l'utilisation de plusieurs signaux pourrait être une solution efficace pour prévenir la mise en place d'un apprentissage perceptuel au cours de l'expérience. De plus, l'étude se rapprocherait ainsi davantage de la situation de choix ouvert. En revanche, l'augmentation du nombre de cibles démultiplie à proportion le nombre d'essais nécessaires (50000 essais pour Dai et Micheyl).

- La classification de bruit (Liu et al., 2014; Vondrick et al., 2014) : certaines études récentes proposèrent des tâches de catégorisation portant sur un stimulus composé uniquement de bruit (i.e. sans signal cible). De telles expériences sont relativement difficiles à mettre en place car elles nécessitent que les fluctuations aléatoires du bruit soient interprétées par l'observateur comme un objet cohérent. Pour cela, l'équipe de Liu diminua progressivement le SNR jusqu'à disparition complète du signal, tout en s'assurant que les participants continuaient à réaliser la tâche. Vondrick et ses collaborateurs adoptèrent quant à eux une approche différente en utilisant un bruit dimensionnel (cf. paragraphe 7.6.2) dans l'espace objet (*feature space*) qui engendre des formes structurées dans l'espace visuel. L'avantage évident de la classification de bruit tient au fait qu'elle prémunit des biais d'estimation dus aux caractéristiques des stimuli. On pourrait imaginer une classification de bruit auditif en s'inspirant des expériences de restauration phonémique, dans lesquelles un phonème supprimé et remplacé par du bruit blanc au sein d'un phrase est néanmoins clairement perçu par l'auditeur (Warren, 1970).

Il est important de noter que le choix de la tâche peut influencer le gabarit utilisé pour la décision. Ainsi, lorsqu'Abbey et Eckstein comparèrent les CIs visuelles obtenues pour trois tâches à deux alternatives (détection, discrimination et identification) dans lesquelles la différence entre les deux signaux cibles était toujours identique, ils constatèrent que le gabarit estimé n'était pas identique dans chacun des cas (Abbey & Eckstein, 2006). Ce résultat démontre que le participant adapte sa stratégie, non seulement aux stimuli mais, également, au type de tâche demandée.

14.1.5. Nombre de signaux utilisés

Nous avons souligné précédemment que les expériences répétitives et portant sur un nombre restreint de signaux différents étaient propices à l'apprentissage perceptuel. Ainsi, il est possible que les participants réalisent la tâche de catégorisation phonémique demandée non par la détection des primitives auditives « générales » de la parole mais, plutôt, en s'appuyant sur des différences acoustiques spécifiques aux enregistrements utilisés comme cible, si cette seconde stratégie s'avère moins coûteuse et/ou plus efficace. Ce phénomène d'apprentissage constitue un point critique puisqu'il

remet en cause la naturalité des indices acoustiques identifiés par la méthode des ACIs. On peut donc se demander combien de signaux sont nécessaires pour limiter cet effet d'apprentissage perceptuel. Un ensemble de données non publiées permet d'éclairer cette question.

Pour vérifier la validité des résultats obtenus dans la tâche de catégorisation à 4 signaux (/alda/, /alga/, /aɹda/ et /aɹga/), nous avons réalisé une seconde expérience, similaire en tout point, excepté que deux signaux seulement étaient présentés (/alda/ et /alga/). Les 3 participants impliqués dans ce post-test obtinrent des performances bien supérieures à celles observées dans l'étude 2 : en moyenne ils réalisèrent la tâche avec un SNR de -15.8 dB contre -11.8 dB (± 0.9 dB S.D.) précédemment. Ce résultat suggérait qu'ils s'appuyaient sur une stratégie d'écoute différente. Effectivement, l'exploration des ACIs obtenues semble confirmer cette hypothèse (Figure 26). Outre l'aspect bruité, dû au faible nombre de participants, la première image laisse apparaître un profil de poids différent : en plus des indices acoustiques déjà observés dans les études 2, 3 et 4 (attaques de F1, F2 et F3 dans la deuxième syllabe), les trois participants fondaient également leur décision sur des indices situés dans la première syllabe (correspondant probablement à la durée de l'intervalle entre les deux syllabes et au contour de f_0 au début du stimulus).

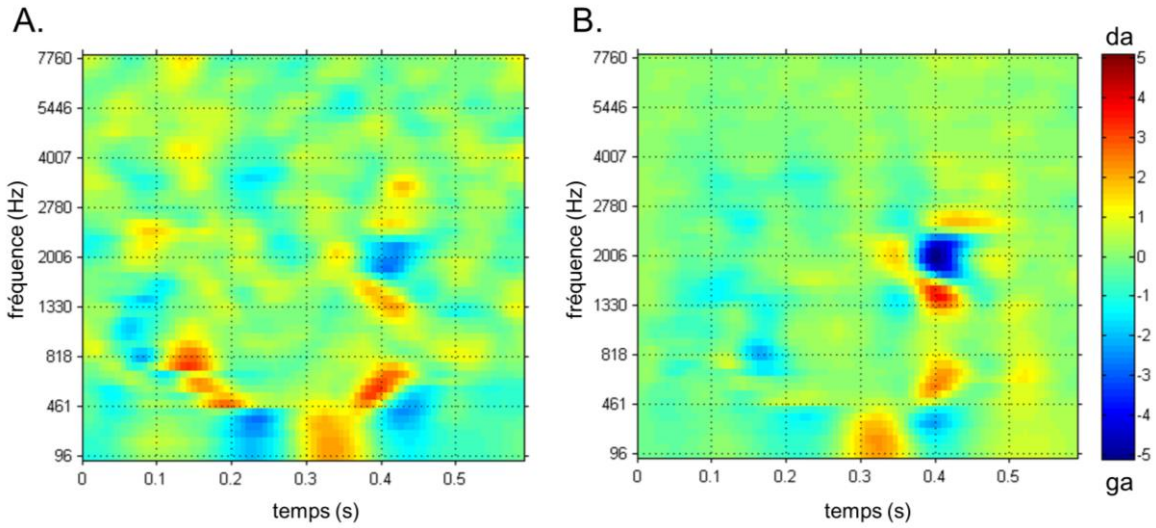


Figure 26 – ACIs moyennes pour la tâche de catégorisation /da/-/ga/. A. Expérience à 2 signaux (/alda/ et /alga/) (N=3 participants). B. Expérience à 4 signaux (/alda/, /alga/, /aɹda/ et /aɹga/) (N=19 participants).

Il semble donc que, dans la tâche portant sur 2 signaux seulement, les participants parviennent à s'appuyer sur des indices acoustiques spécifiques aux enregistrements utilisés, au contraire de la tâche portant sur 4 signaux. La comparaison

entre les résultats des deux expériences suggère donc que l'emploi de 4 signaux différents pourrait suffire à prévenir la mise en place de stratégies non linguistiques.

14.1.6. Comparaison avec les autres méthodes psycholinguistiques et champ d'application des ACIs

Dans la première partie de cette thèse, nous avons dressé une liste détaillée des différentes méthodes psychoacoustiques exploitables pour identifier les indices acoustiques impliqués dans une tâche de catégorisation phonémique (chapitre 4). Nous présentons un résumé des caractéristiques de ces différentes approches dans la Table 1 page suivante.

Comme c'est le cas pour les différentes techniques d'imagerie cérébrales, qui possèdent chacune des avantages et des inconvénients (Zani et al., 2003), on constate dans la Table 1 qu'aucune des techniques psychoacoustiques envisagées – pas même l'ACI – n'est satisfaisante sur tous les aspects. Le choix d'une approche particulière dépend donc nécessairement de l'objectif de l'étude. Pour ne pas en rester à cette conclusion trop générale, nous tentons ci-dessous de délimiter le champ d'application des ACIs, par comparaison avec les autres méthodes disponibles. Plusieurs aspects doivent être considérés :

- Une contrainte pratique souvent reprochée à la méthode des ACIs tient à sa durée d'expérimentation. Celle-ci doit cependant être relativisée au regard d'autres techniques exigeant jusqu'à 15 heures de passation par participant (la 3DDS combine même les résultats de 3 tâches longues, accomplies par des participants différents, pour un total d'environ 19 heures). L'ACI n'en demeure pas moins une technique coûteuse en termes de durée. À ce titre, elle est donc d'un usage moins « immédiat » que le continuum de parole.

- Aucune des approches présentées ici ne garantit absolument la naturalité des indices acoustiques identifiés, puisque toutes sont basées sur des stimuli synthétiques ou dégradés (ajout de bruit, filtrage, troncature), toutes opérations qui peuvent susciter un phénomène d'adaptation au niveau du système auditif. Comme nous l'avons déjà noté, cependant, l'ACI est plutôt défavorable de ce point de vue, notamment par rapport à la 3DDS, puisque le nombre limité d'enregistrements utilisés comme stimuli peut faciliter la mise en place de stratégies alternatives. Nous avons déjà apporté, dans ce chapitre, quelques pistes de solutions à ce problème.

- Dans l'attente de nouveaux développements méthodologiques, le calcul des ACIs est restreint au cadre d'une tâche de catégorisation en choix fermé. Les situations en choix ouvert, quant à elles, requièrent l'utilisation de la 3DDS ou des bulles auditives.

Méthode	*Première utilisation référencée*	*Durée de passation*	*Stimuli naturels*	*Stimuli non-dégradés*	*Tâche en choix ouvert*	*Caractérisation des indices*	*Distinction des indices multiples*	*Rôle des connaissances a priori*	*Combinaison avec neuro-imagerie*
Continuum de parole	(Liberman et al., 1952)	~ 30 mn	Non	Oui	Non	Précise	Parfois	Indispensables (a priori sur l'indice)	Oui
Signal progressivement dégradé	(Fletcher, 1922)	~ 30 mn	Parfois	Non	Oui	Très imprécise	Non	Important (a priori sur la dimension)	Oui
Profils de confusion	(Allen, 2005)	15 h	Oui	Non	Oui	Très imprécise	Non	Faible, mais nécessite une interprétation	Oui
3DDS	(Li et al., 2010)	4 h + 4 h + 15 h	Oui	Non	Oui	Relativement précise	Difficile	Faible, mais nécessite une interprétation	Non
Fonctions de pondération	(Ahumada & Lovell, 1971)	3 h	Oui	Non	Oui	Très imprécise	Non	Important (choix des fréquences de coupure)	Envisageable
Bulles auditives	(Mandel, 2013)	1 h	Oui	Non	Oui	Moyenne	Non	Faible (choix de la taille et de la forme des bulles)	Envisageable
ACIs	(Varnet et al., 2013b)	3 h	Oui	Non	Non	Très précise	Oui	Faible (choix du type de régularisation)	Envisageable

Table 1 – Tableau récapitulatif des différentes méthodes psychoacoustiques d'identification des indices acoustiques.

- En revanche, l'ACI présente à l'évidence un point fort, par rapport à toutes les autres techniques mentionnées ici, qui réside dans la précision de sa description de la stratégie employée. Elle permet en effet de localiser, avec une excellente résolution temps-fréquence, n'importe quel nombre d'indices acoustiques et, ce, sans exiger de connaissances a priori (sinon la supposition que le gabarit sous-jacent est relativement lisse), ni introduire une trop grande part d'interprétations.

- Autre avantage de ce point de vue : contrairement à la plupart des autres méthodes (signal progressivement dégradé, profils de confusion, 3DDS, fonctions de pondération et bulles auditives) qui fondent leurs calculs sur l'exactitude de la réponse du participant à chaque essai, les données recueillies pour l'ACI concernent le phonème répondu (que cette réponse soit correcte ou non). Ainsi l'ACI comporte plusieurs clusters de poids, en faveur d'une réponse ou de l'autre, organisés horizontalement et/ou verticalement, ce qui permet de déterminer précisément l'effet de chaque indice sur la catégorisation.

Compte tenu de ces éléments de comparaison, nous préconiserons donc d'utiliser la technique des ACIs dans les deux contextes expérimentaux suivants :

1) Dans le cadre d'expériences visant à identifier précisément les indices acoustiques mis en jeu lors d'une tâche de catégorisation particulière. Cette situation correspond aux études 1 et 2, avec la recherche des indices acoustiques des catégorisations /ba/-/da/ et /da/-/ga/. Il serait également intéressant d'identifier de cette manière les indices acoustiques employés pour la segmentation de la parole, par exemple pour différencier « l'ami » et « la mie » (Pota et al., 2012), ou dans la compréhension d'autres signaux de communication, tels que les langues sifflées (Meyer, 2015).

2) Dans le cadre d'expériences étudiant l'effet d'un facteur susceptible d'entraîner des modifications relativement fines de la stratégie de catégorisation employée par les auditeurs. Ce deuxième cas est illustré par les études 3 et 4, qui visaient à caractériser l'impact de l'entraînement musical et de la dyslexie sur les indices acoustiques utilisés dans une tâche de catégorisation. D'autres facteurs inter-sujets pourraient être explorés par le même procédé, notamment l'adaptation liée à une surdité ou l'apprentissage perceptuel développé suite à une implantation cochléaire (voir paragraphe 14.2.4). L'étude peut également porter sur des facteurs contextuels : par exemple, l'effet du contexte acoustique ou phonétique, du contexte lexical ou du contexte visuel.[19] Jusqu'à présent, seules les méthodes du continuum de parole (Mann,

[19] Ces trois facteurs sont mis en évidence par trois protocoles expérimentaux particuliers, respectivement (1) une tâche de catégorisation d'une cible ambigüe dont l'interprétation dépend des sons ou phonèmes adjacents (Liberman et al., 1952), (2) une tâche de catégorisation d'un phonème initial biaisée par le statut lexical (« barrique ») vs. non lexical (« darrique ») de la cible (Ganong, 1980), (3) une tâche opposant des stimuli

1980) et des fonctions de pondération (Gilbert et al., 2002) ont été mises en œuvre dans ce type d'études mais, en raison de leur faible résolution, les résultats obtenus se sont avérés parfois difficiles à interpréter. Du fait de sa plus grande précision, l'approche par les ACIs paraît donc plus adaptée dans ce cas.

Pour terminer, deux autres effets de contexte, dans un sens plus large, mériteraient, eux aussi, d'être envisagés sous l'angle des ACIs.

Tout d'abord, l'influence du type de bruit sur l'utilisation d'un indice acoustique plutôt qu'un autre (voir paragraphe 5.1). L'un des avantages du GLM tient à ce qu'il n'impose pas de contrainte sur la distribution des prédicteurs, ouvrant ainsi la possibilité d'utiliser des bruits non gaussiens. On pourrait, en particulier, imaginer comparer les résultats de l'étude 2 avec ceux d'une expérience similaire réalisée dans un bruit stationnaire de spectre semblable à celui de la parole (Phatak & Allen, 2007), dans un bruit fluctuant (Ziegler et al., 2009) ou dans un masqueur parolier (Bronkhorst, 2000; Hoen et al., 2007; Varnet et al., 2012b).

Un second protocole expérimental qui pourrait, lui aussi, être combiné avec la méthode des ACIs est celui proposé par Dehaene-Lambertz et ses collaborateurs (Dehaene-Lambertz et al., 2005). Les auteurs utilisent ici une dégradation particulière de la parole, dite parole sinusoïdale (*sine-wave speech*). L'intérêt de ce type de sons est que l'auditeur naïf ne les interprète pas comme de la parole mais, plutôt, comme une superposition de sifflements. En revanche, lorsque l'on informe l'auditeur de la présence d'un contenu linguistique dans les stimuli, il lui devient difficile, sinon impossible, de ne pas les percevoir en tant que sons de parole. Tirant parti de ce phénomène, les chercheurs étudièrent les activités cérébrales des participants lors de tâches de discrimination, avant et après ce basculement de la perception de non parolière à parolière. De manière identique, il pourrait être intéressant de comparer les ACIs obtenues pour une même tâche de catégorisation en parole sinusoïdale, selon que les stimuli sont interprétés ou non comme de la parole. Une expérience semblable avec des CI visuelles a été réalisée sur la perception de visages à l'endroit et à l'envers (Sekuler et al., 2004).

14.2. Discussion psycholinguistique

La description des limites et contraintes de la méthode des ACIs, ainsi que sa comparaison avec les autres méthodes psychoacoustiques, nous a conduits à délimiter son champ d'application. Les quatre études présentées dans cette thèse offrent des exemples pratiques d'utilisation de cette technique dans le cadre de problèmes

audiovisuels cohérents ou incohérents de type McGurk (Fowler and Rosenblum, 2014; Johnson, 2011). Chacun de ces 3 protocoles expérimentaux pourraient être adaptés assez simplement au calcul des ACIs.

psycholinguistiques concrets. Dans cette partie, nous verrons dans quelle mesure les observations réalisées lèvent une partie du voile sur la compréhension de l'interface acoustico-phonétique. Cinq aspects sont particulièrement intéressants dans ce sens : les indices anticipatoires, l'importance des facteurs aléatoires internes, la variabilité limitée des cibles, la diversité des stratégies d'écoute et la similarité des ACIs avec les STRFs. Pour finir, cette analyse nous amènera à avancer l'idée que la méthodologie décrite ici pourrait constituer un nouveau pont entre les domaines de la neurobiologie et de la psycholinguistique.

14.2.1. Indices anticipatoires et coarticulation

L'un des phénomènes les plus frappants mis en évidence au moyen des ACIs est l'observation d'indices anticipatoires, c'est-à-dire de caractéristiques acoustiques précédant la cible mais influant pourtant sur sa catégorisation. L'interprétation que nous en avons faite dans l'étude 1 est assez succincte et demande donc à être approfondie ici.

Pour la première mise en application des ACIs, nous avons utilisé des sons de parole composés de deux syllabes produites séparément : un /a/, identique pour les deux cibles, suivi d'un /ba/ ou d'un /da/. De manière inattendue, cette étude révéla la présence d'un indice acoustique sur la fin du F2 de la 1ère syllabe, en sus de celui espéré sur l'attaque du F2 de la 2ème syllabe. Comme nous l'avons noté, ce premier indice est particulièrement intriguant car il signifie que la décision de l'auditeur est affectée par la distribution du bruit dans la première syllabe, bien que celle-ci ne contienne aucune information pertinente pour la tâche (puisqu'elle est identique pour les deux cibles).[20]

Une première intuition nous fit rapprocher cette configuration particulière des poids au sein de l'ACI de celle précédemment observée dans les CIs visuelles pour la tâche d'acuité de Vernier (Ahumada, 1996; Ahumada & Beard, 1998; Barth et al., 1999; Beard & Ahumada, 1998; Li et al., 2006). En effet, la décision concernant l'alignement ou non de deux segments – l'un restant toujours à la même position pour tous les essais et l'autre occupant une position variable selon les cibles – faisait, elle aussi, apparaître un indice attendu au niveau du segment mobile mais également un second indice au niveau du segment fixe (Figure 21). Ce résultat fut interprété comme une preuve attestant que la tâche sollicitait non pas un, mais deux, filtres de Gabor et que la décision de l'observateur portait en réalité sur la position relative des deux segments. De la même manière, on peut imaginer que la catégorisation /ba/-/da/ soit sous-tendue par une simple détection de l'alignement du F2 entre la consonne et la voyelle précédente. Cette explication revient à envisager une forme de théorie acoustique semblable à celles énoncées au paragraphe 3.1.2. Cette hypothèse est toutefois remise en question par la

[20] Par ailleurs – si l'on se place dans le cadre de la Théorie de combinaison des indices (Treisman, 1999) – la présence de cet indice non informatif dans la première syllabe pourrait expliquer la non-linéarité des fonctions psychométriques observées dans l'étude 1.

disposition des poids positifs et négatifs. Tandis que, dans l'ACI correspondant à la tâche de catégorisation phonémique, les deux indices possèdent la même structure (un groupe de poids négatifs au-dessus d'un groupe de poids positifs), dans la CI visuelle correspondant à la tâche d'alignement, le couple d'indice montre des structures opposées : poids positifs en haut et négatifs en bas pour l'un, et inversement pour l'autre. En effet, dans la tâche de Vernier, un non-alignement est détecté lorsque l'un des segments apparaît relativement haut tandis que l'autre semble relativement bas. Au contraire, le stimulus auditif a une probabilité maximale d'identification en tant que /aba/ (respectivement /ada/) lorsque les fréquences du F2 dans les deux syllabes sont conjointement perçues comme étant basses (respectivement élevées). La fréquence du F2 ne semble donc pas jouer le rôle de « point de repère » que nous lui avions prêté de prime abord. Par conséquent, cette tâche auditive ne se ramène pas à une simple estimation de la hauteur du F2 dans la seconde syllabe par rapport à la première.

Une autre explication de cet indice anticipatoire pourrait tenir à ce que l'auditeur tente d'extraire des informations de coarticulation de la première syllabe, même si, par construction du stimulus, celle-ci n'en contient aucune. En effet, dans des productions naturelles des sons /aba / et /ada/, la trajectoire des formants dans la voyelle initiale dépend de l'identité de consonne qui va suivre. En particulier, la fin du F2 dans la 1ère syllabe a une fréquence plus élevée lorsqu'elle est suivie de /d/ que lorsqu'elle est suivie de /b/ (Figure 4). Donc, en situation naturelle, le système auditif est ordinairement en mesure de prédire l'identité d'une consonne à partir du contenu spectral du contexte qui la précède (Mann, 1980). Dans la tâche de l'étude 1, cette anticipation est impossible, puisque la 1ère syllabe est absolument identique dans les deux cibles, mais l'auditeur demeure, malgré tout, sensible aux variations aléatoires introduites par le bruit dans la 1ère syllabe. Lorsque le F2 du /a/ initial semble perceptivement plus haut qu'il n'est en réalité, du fait de la répartition du bruit, la probabilité de la réponse /d/ est plus haute (et inversement pour /b/). Il s'agit là d'une démonstration directe de l'utilisation de l'information de coarticulation.

14.2.2. Bruit interne

Pour espérer mettre à jour les règles présidant à l'association entre stimulus présenté et percept produit, il est indispensable que celles-ci manifestent une certaine régularité, c'est-à-dire que les effets aléatoires n'aient qu'une incidence limitée sur la décision. Dès l'équation 7.7, nous avons reconnu que la décision psychoacoustique était gouvernée par deux types de facteurs : les facteurs externes et déterministes, traduits par le terme $\underline{S} * \underline{\beta}$ qui dépend uniquement du stimulus ; mais, aussi, les facteurs internes et aléatoires, comme l'état instantané du système nerveux de l'auditeur, l'historique de ses réponses passées, les pertes d'attention, etc…, qui sont rassemblés dans le terme ε. Néanmoins, il subsiste un doute quant à l'importance relative de ces deux effets. Pour

justifier que le gabarit $\underline{\beta}$ possède une influence réelle et déterminante sur la catégorisation, il faut s'assurer que le facteur aléatoire n'est pas trop prépondérant.

Le bruit interne ε permet de tenir compte du fait qu'un participant humain ne réagira pas systématiquement de la même manière en réponse à un même stimulus. Ainsi, dans la tâche de catégorisation phonémique, la présentation d'une cible donnée dans un bruit fixe donné, à deux reprises dans le cours de la passation, n'entraînera pas nécessairement deux réponses identiques. Cette observation conduit à une première technique d'estimation du bruit interne, appelée double-passation (*double-pass*). Elle consiste à évaluer la cohérence des réponses du participant entre deux sessions successives des mêmes N essais, présentés dans un ordre aléatoire. La première estimation, en 1964, dans le cadre d'une tâche auditive non parolière indiqua une correspondance de seulement 66 % entre les réponses aux deux sessions (ce qui correspond à des facteurs interne et externe de magnitudes égales) (Green, 1964). Des expériences postérieures dans le domaine visuel (Diependaele et al., 2012; Kurki & Eckstein, 2014) ou auditif (Joosten & Neri, 2012) conduisirent à des résultats du même ordre. Cette inconstance dans les réponses des participants impose donc naturellement une borne supérieure aux performances des algorithmes de prédiction.

Les études que nous avons réalisées n'incluant pas de double passation, elles ne permettaient pas d'estimer directement le bruit interne. Néanmoins, il est à noter que le taux de prédictions correctes de notre modèle (mesuré par validation croisée) dépasse ceux des études mentionnées précédemment. Nous sommes donc enclins à conclure que la tâche proposée, bien que très répétitive, n'engendre pas à une part trop importante du facteur aléatoire dans la décision, peut-être grâce à l'algorithme de *staircase* qui limite les chutes attentionnelles.

14.2.3. Variabilité de la parole

Il pourrait sembler, de prime abord, que les études présentées dans cette thèse ne permettent pas de prendre en compte la variabilité des productions parolières. En effet, elles comportent seulement 1 ou 2 enregistrements différents pour chaque cible, alors que la parole naturelle connaît une infinité de variations acoustiques à l'intérieur d'une même catégorie phonémique.

Néanmoins, par l'addition d'un bruit blanc, le protocole expérimental introduit de fait une certaine forme de variabilité dans les stimuli présentés. L'idée ingénieuse qui sous-tend la technique des CIs tient à ce que, lors de certains essais, la distribution aléatoire du bruit reproduit par hasard certaines caractéristiques similaires à l'une des cibles, augmentant ainsi la probabilité que le participant identifie le stimulus comme contenant cette cible. D'une certaine façon, le bruit blanc engendre donc une variabilité qui est parfois interprétée par l'auditeur comme une variabilité parolière. Le bruit peut ainsi faire paraître les formants plus élevés ou plus bas qu'ils ne sont en réalité,

induisant parfois des erreurs de catégorisation. Ce phénomène permet ainsi de mettre en évidence des effets relevant de la parole naturelle et variable : comme nous l'avons fait remarquer au paragraphe 14.2.1, les participants de l'étude 1 cherchaient à extraire des informations de la première syllabe des stimuli /aba/ et /ada/, bien que cette syllabe soit toujours parfaitement identique.

Les participants font donc preuve, pour la plupart, d'une intéressante absence d'adaptation face aux tâches de catégorisation extrêmement répétitives présentées ici. Les processus de traitement de la parole impliqués sont à un tel point automatisés que le système auditif ne parvient pas à prendre en compte la variabilité très limitée des cibles. Au contraire, il reste en *speech mode* tout au long de l'expérience, interprétant les stimuli présentés à la lumière de ses connaissances concernant la parole naturelle – notamment, le fait que celle-ci contient normalement une variabilité importante et des indices de coarticulation. Il en découle une conséquence directe très importante : il est possible, par le calcul des ACIs, de faire apparaître les indices auxquels les auditeurs s'attendent en situation ordinaire, même lorsque ces indices ne sont pas présents dans les signaux.

14.2.4. Diversité des stratégies d'écoute

Ce travail de thèse s'intéresse non seulement au fonctionnement général de l'interface acoustico-phonétique mais, également, à la variabilité présente au sein des stratégies d'écoute des différents participants. Nous avons distingué deux sources de variabilité : d'une part une variabilité inter-groupes (p.ex. les participants musiciens obtiennent une distribution des poids de l'ACI différente de celle des participants non-musiciens) et, d'autre part, une variabilité intra-groupe (p.ex. les participants dyslexiques présentent des stratégies d'écoute relativement hétérogènes).

L'étude 3 a permis de constater que l'expertise musicale se traduit par des performances supérieures lors de la catégorisation phonémique dans le bruit et, au niveau de l'ACI, par une focalisation accrue des poids perceptuels sur les régions du stimulus pertinentes pour cette tâche. Ainsi, même si la stratégie d'écoute des participants musiciens demeure identique à celle du groupe contrôle, le système auditif fait preuve d'une certaine plasticité au niveau des pondérations relatives des différents indices. En outre, dans notre étude 4, l'ACI d'un sous-groupe de 11 participants dyslexiques fait ressortir l'implication de deux indices supplémentaires dans la décision phonémique. Il s'agit donc bien, ici, d'une différence qualitative, et non quantitative, entre les stratégies d'écoute. Dans cet article, nous avons argumenté que le déficit phonologique contraint les auditeurs dyslexiques à s'adapter en s'appuyant, notamment, sur les redondances présentes dans les signaux. De façon plus générale, on peut se demander de quelle manière d'autres altérations des fonctions auditives non-langagières parviennent à influencer sur la compréhension de la parole. Certaines études, utilisant la technique des fonctions de pondération, indiquent que les participants souffrant de pertes auditives, dotés d'implants cochléaires ou non,

réallouent les poids perceptuels sur les différentes bandes de fréquence (Calandruccio & Doherty, 2008; Doherty & Lutfi, 1996; Gilbert et al., 2002). Là encore, l'approche des ACIs pourrait amener une analyse plus fine de ce phénomène d'adaptation des stratégies d'écoute. Toutefois, comme souligné dans les études 3 et 4, lors des comparaisons inter-groupes, il est nécessaire de prendre en compte le fait que les participants présentant un déficit auditif peuvent se montrer particulièrement sensibles au protocole expérimental employé, car celui-ci requiert une attention soutenue sur une tâche longue et répétitive.

A contrario, les ACIs individuelles des auditeurs contrôles et musiciens présentent visuellement une saisissante similarité. Précédemment, les études employant la méthode des fonctions de pondération sur des tâches de compréhension de la parole avaient déjà obtenu des stratégies d'écoute très similaires au sein de leurs groupes de participants (Apoux & Bacon, 2004; Doherty & Turner, 1996; Turner et al., 1998b). Ces résultats peuvent s'expliquer par le fait que les tâches de parole mettent en jeu des processus extrêmement automatisés et donc probablement identiques chez tous les auditeurs d'un même groupe. Comparativement, les ACIs des participants dyslexiques ayant participé à l'étude 4 présentent des distributions de poids beaucoup plus hétérogènes. Deux explications sont possibles pour ce phénomène : d'une part, ces pondérations déstructurées pourraient être le reflet des représentations phonologiques dégradées des auditeurs dyslexiques, se traduisant in fine par des performances plus faibles ; au contraire, il est possible que le SNR élevé auquel ces participants réalisent la tâche leur ouvre la possibilité d'utiliser des stratégies d'écoute plus diverses, en tirant parti de la redondance du signal de parole. Néanmoins, ces constatations ne reposant que sur un sous-groupe de 7 participants dyslexiques, une exploration complémentaire plus approfondie est indispensable pour interpréter ces stratégies alternatives. Comme suggéré plus haut, on peut sans doute s'attendre à observer également une grande variété parmi les ACIs individuelles dans le cadre d'une étude portant sur des participants souffrant de pertes auditives. En effet, dans une tâche d'identification de phonèmes, les matrices de confusion qu'obtiennent ces participants indiquent d'importantes différences perceptuelles (Phatak et al., 2009; Trevino & Allen, 2013).

Une conclusion temporaire semble donc se dessiner à partir des observations précédentes. La stratégie d'écoute mise en œuvre dans une tâche de catégorisation de phonèmes dans le bruit apparaît extrêmement similaire d'un auditeur à l'autre, excepté dans le cas de troubles particuliers nécessitant la mise en place de stratégies alternatives, comme le déficit phonologique associé à la dyslexie. Cependant, même chez un auditeur normo-lecteur, un entraînement intensif des fonctions auditives, comme celui procuré par la pratique musicale, peut amener à une plus grande focalisation de la stratégie qui se traduit par une pondération plus marquée des indices usuels.

14.2.5. Interprétation des ACIs en lien avec les STRFs

L'ACI est une méthode comportementale qui s'inscrit dans le cadre des recherches en psycholinguistique sur les primitives de la parole. Elle se fonde néanmoins sur le même cadre statistique (GLM) et le même espace de représentation (poids dans l'espace temps-fréquence) qu'une autre méthode couramment employée en neurobiologie pour caractériser les relations stimulus-réponse, cette fois-ci au niveau des neurones individuels : le STRF (voir paragraphe 6.1). La similitude entre les propriétés des ACIs et celles des STRFs est frappante (Figure 27). Les deux représentations mettent en évidence des zones activatrices flanquées de zones inhibitrices (Joosten & Neri, 2012; Neri & Levi, 2006). Ceci traduit le fait que dans les deux cas les processus opèrent non pas des détections absolues (i.e. l'énergie dépasse un certain seuil) mais plutôt des mesures relatives (i.e. l'énergie présente des variations importantes entre deux fréquences ou deux instants adjacents). On observe aussi, par les deux techniques, une relative adaptabilité des poids à la tâche (Fritz et al., 2003) ainsi qu'une certaine sensibilité au contexte (Woolley et al., 2006).

Dans un article de 2006, Neri et Levi réalisèrent une revue comparée entre, d'une part, des études utilisant la méthode des CIs visuelles et, d'autre part, des études estimant les champs récepteurs de neurones visuels par une méthode analogue aux STRFs (Neri & Levi, 2006). Les auteurs notèrent le plus souvent une remarquable ressemblance entre les résultats des études psychophysiques et leurs équivalents neurobiologiques, par exemple entre un champ récepteur d'une cellule de V1 détectant l'orientation et les CIs de tâches de discrimination d'orientation ou d'alignement de Vernier.

De même, les ACIs mesurées dans nos 4 études offrent des ressemblances frappantes avec les champs récepteurs des neurones auditifs mesurés chez l'animal, et particulièrement avec deux catégories de STRFs : ceux des neurones spectraux à bande étroite ou à bande large (Woolley et al., 2009). Ces STRFs sont le plus souvent caractérisés par 2 ou 3 zones excitatrices et inhibitrices alternées verticalement (ou parfois diagonalement). Dans la typologie proposée par Woolley et al. ces groupes sont notamment associés à la détection du timbre et seraient donc parfaitement adaptés à la catégorisation de phonèmes (Mesgarani et al., 2008). Si l'on regarde plus en détail la structure de nos ACIs (Figure 27), celle obtenue pour la tâche de catégorisation /aba/-/ada/ pourrait être vue comme une somme de deux STRFs spectraux, détectant la fréquence de la fin et de l'attaque du F2. L'ACI de la tâche de catégorisation /da/-/ga/ rappelle, elle aussi, des profils de poids observés lors de mesures de STRFs, avec une combinaison d'indices hautes- et basses-fréquences (Figure 27).

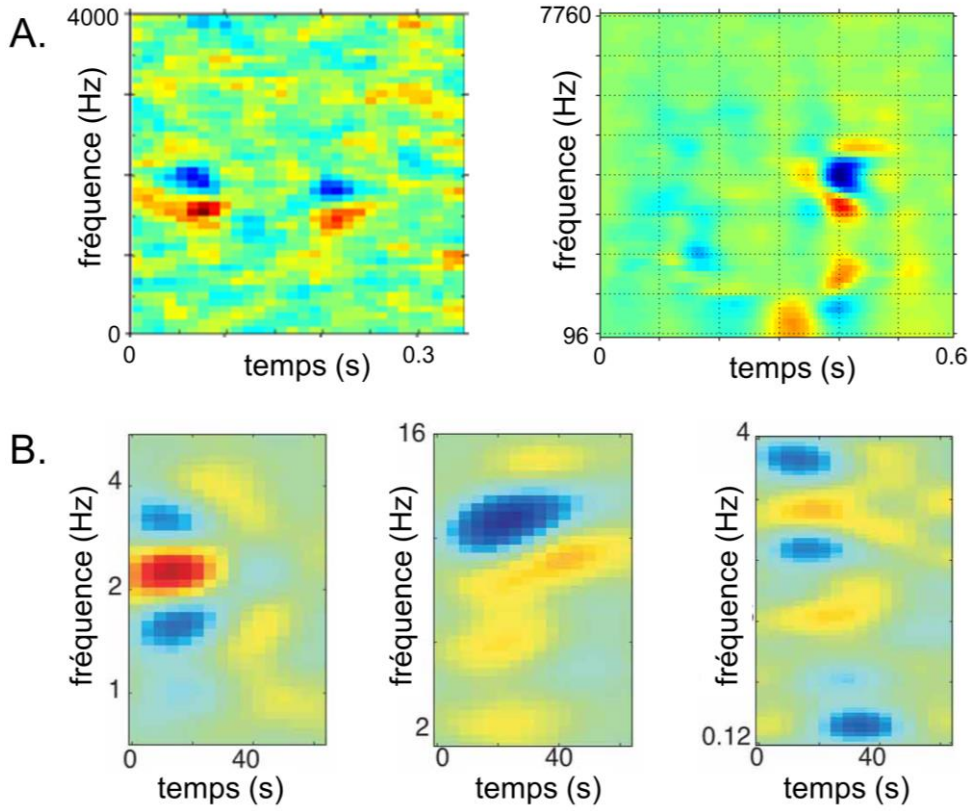


Figure 27 – Rapprochement entre les ACIs et les STRFs. A. Exemples d'ACIs issues des études précédentes (gauche : tâche de catégorisation /aba/-/ada/ ; droite : tâche de catégorisation /da/-/ga/ en contexte /al/ et /aʁ/). B. Exemples de STRFs de neurones auditifs du cortex auditif primaire obtenus pour des tâches d'écoute passives chez le furet, adaptés de (Fritz et al., 2003). L'échelle couleur représente la probabilité de réponse dans A. et la probabilité de décharge dans B. Tous les axes fréquentiels sont logarithmiques excepté celui de la première ACI.

Neri et Levi soulignent néanmoins que cette similarité ne doit pas être interprétée comme une correspondance directe entre les niveaux neurologique et comportemental (i.e. le processus de catégorisation ne se réduit pas à l'activation ou non d'un ou deux neurones spécialisés), mais plutôt comme une indication que le comportement de l'auditeur et du neurone sont caractérisés par des computations similaires. De plus, comme nous l'avons déjà fait remarquer précédemment, STRFs et ACIs ne sont que des estimations linéaires des processus : leur similarité apparente pourrait donc, éventuellement, masquer une grande dissemblance au niveau des traitements non linéaires mis en jeu. Il est cependant à noter que, si l'on se place dans le cadre du modèle de la représentation multi-résolution (voir paragraphe 3.2.4) dans lequel chaque phonème est caractérisé par l'activation d'un groupe particulier de neurones auditifs, l'ACI correspondant à la tâche de catégorisation entre deux phonèmes est effectivement équivalente à une somme pondérée des STRFs impliqués. Ce type de construction d'un champ récepteur complexe par sommation de plusieurs champs

récepteurs simples a par ailleurs été observé entre les neurones auditifs du thalamus et du cortex (Amin et al., 2010; Miller et al., 2001).

Suivant la même ligne de raisonnement, Mesgarani et ses collaborateurs proposèrent en 2008 de caractériser les STRFs d'une population de neurones encodant un phonème particulier (Mesgarani et al., 2008). Après avoir fait écouter à des furets des stimuli de parole anglaise tout en enregistrant les réponses de leurs neurones auditifs, ils calculèrent pour chaque phonème les BF, BS et BR moyens sur l'ensemble des neurones s'activant de manière maximale suite à la présentation de ce phonème. De manière intéressante, il apparaît, dans cette représentation, que les phonèmes /b/ et /d/ sont séparés essentiellement en termes de BF, tandis que /d/ et /g/ diffèrent principalement en BS. Ces observations recoupent les conclusions de nos études (la catégorisation /b/-/d/ est gouvernée par la hauteur d'un formant ; la catégorisation /d/-/g/ par les positions relatives de deux formants).

Pour aller au-delà de la simple constatation d'une ressemblance entre les CIs et les champs récepteurs des neurones visuels, un groupe de recherche a tenté récemment de relier les résultats comportementaux analysés par la méthode des CIs et l'activité d'une aire du cortex visuel mesurée par IRMf (Liu et al., 2014). L'application d'une telle combinaison avec la neuro-imagerie dans le contexte des ACIs est également tout à fait envisageable ; elle permettrait potentiellement d'établir un lien entre les prédictions des réponses comportementales des participants à un stimulus auditif et les prédictions de leurs réponses corticales.

14.2.6. L'ACI, un nouveau pont entre la neurobiologie et la psycholinguistique ?

Selon David Marr, l'analyse de toute activité cognitive peut être conduite à plusieurs niveaux différents : le niveau computationnel (description des fonctions que le système doit être en mesure d'accomplir pour effectuer une tâche), le niveau algorithmique ou *software* (description des processus par lesquels une fonction est réalisée, étapes de traitements, règles et représentations utilisées) et le niveau implémentationnel ou *hardware* (description exactes des computations réalisées et des structures matérielles supportant le calcul) (Marr & Poggio, 1976). Ainsi, la question de la compréhension de la parole peut être abordée du point de vue de son encodage neuronal ou de celui des fonctions linguistiques supportées (catégorisation, stockage, segmentation…), les deux aspects étant reliés par la description détaillée des opérations de traitement de l'information réalisées par le système.

À ces niveaux de description correspondent deux approches distinctes. D'une part, la psycholinguistique qui s'appuie sur des indicateurs comportementaux pour formuler et valider ses hypothèses quant aux traitements et représentations mis en jeu (par quels traitements l'information linguistique est-elle extraite du signal acoustique ?) ; d'autre part, la neurobiologie qui tend à reconstituer la machinerie

cognitive, par rétro-ingénierie (*reversed-engineering*), à partir de l'observation de son substrat neuronal (comment l'architecture neuronale du système auditif permet-elle le traitement du son en vue de sa compréhension ?). Pour prétendre à la complétude, une théorie auditive ne peut se limiter à une description en termes algorithmiques, car les fonctions peuvent être implémentées de différentes manières au niveau neuronal (Cariani & Micheyl, 2012) et, inversement, une description purement implémentationnelle – sous la forme d'une carte, même exhaustive, de l'architecture neuronale du système auditif – ne nous renseignera pas nécessairement sur les traitements effectués (Tiberghien, 2007).

De formidables avancées ont été réalisées récemment dans le domaine de la neurolinguistique et nous disposons à présent de cartographies du cerveau extrêmement précises, spécifiant au sein de quelles structures (hémisphères, aires cérébrales, voies, réseaux neuronaux…) les principales fonctions linguistiques (e.g. segmentation, traitement phonologique, accès lexical, analyse syntaxique…) sont accomplies (Friederici, 2012; Hickok, 2009; Hickok & Poeppel, 2007; Kotz & Schwartze, 2010). Cependant ces modèles restent relativement évasifs quant au détail des processus effectués au sein de ces structures (i.e. ils ne traitent pas du niveau algorithmique) (Poeppel et al., 2012).[21] L'un des enjeux des sciences cognitives actuelles est de parvenir à unifier les 3 niveaux de description en proposant des théories neurobiologiquement fondées qui explicitent la chaîne de traitements opérés par le système cognitif. Cet objectif est matérialisé par l'apparition d'un nouveau champ de recherche, la neurobiologie computationnelle, qui a déjà proposé plusieurs modèles du traitement de la parole par le cerveau (Giraud & Poeppel, 2012; Pasley & Knight, 2013). Ainsi les chercheurs espèrent-ils aujourd'hui comprendre de quelle manière les circuits neuronaux permettent l'implémentation des opérations qui soutiennent les fonctions linguistiques.

Dans ce cadre, l'une des principales difficultés s'opposant à la création d'un pont entre les données psycholinguistiques et neurobiologiques tient à leur l'incommensurabilité (*Ontological Incommensurability Problem*) (Embick & Poeppel, 2015). En effet, les deux champs de recherche fondent leurs réflexions sur des éléments conceptuels différents, ne présentant pas de relation biunivoque entre eux. Pour le linguiste, les unités élémentaires du langage sont les traits distinctifs, les phonèmes ou les syllabes. En revanche, le neurobiologiste réfléchit en termes de champs récepteurs, réseaux de neurones ou synchronisations de décharges. L'un des défis actuels pour la neurobiologie computationnelle est donc de décrire les règles de conversion entre ces deux niveaux d'abstraction (Cariani & Micheyl, 2012). De ce point de vue, la méthode de l'ACI pourrait être amenée à jouer un rôle important dans les recherches concernant la

[21] De plus, certains auteurs se sont interrogés sur la nature de la connaissance produite par cette méthodologie « localisationniste », en critiquant notamment l'imprécision de la définition des fonctions recherchées qui regroupent souvent un ensemble de processus très divers (Poeppel and Embick, 2005; Tiberghien, 2011).

compréhension de la parole car elle pourrait aider à établir le lien manquant entre les primitives linguistiques et neuronales.[22]

Le schéma Figure 28 illustre les 3 niveaux de représentations d'un son de parole en jeu : le stimulus acoustique (1), encodé sous la forme d'une réponse neuronale (2) et engendrant *in fine* un percept linguistique (3). Le problème revient donc à établir la relation entre les niveaux 2 et 3 du schéma, c'est-à-dire à décrypter la réponse neuronale encodant un percept linguistique donné.

L'approche suivie jusqu'à présent par la neurolinguistique (symbolisée par les flèches rouges) combine les méthodes neurophysiologiques et psycholinguistiques : une variation du stimulus acoustique entraîne, ou non, des variations conjointes aux niveaux perceptuel (mesurées par exemple par un pourcentage de catégorisation phonémique) et neuronal (visualisée par la neuro-imagerie). Cette solution interdisciplinaire espère ainsi pouvoir s'affranchir de la dimension acoustique afin d'étudier les corrélats neuronaux stables des percepts linguistiques (Bellier et al., 2013; Chang et al., 2010; Obleser et al., 2004). Plus récemment, l'approche en *reversed-engineering* de la neurobiologie computationnelle (correspondant à la flèche verte) a ouvert une voie plus directe : les réponses neuronales sont ici « décodées » afin de prédire le percept linguistique (en passant parfois par une étape acoustique). La reconstruction de la parole à partir de l'activité cérébrale permet non seulement de faire le lien entre les différentes unités dans ces deux niveaux mais, également, de mettre à jour le processus de traduction à l'œuvre entre eux (Mesgarani et al., 2014a; Pasley & Knight, 2013).

[22] Il a déjà été souligné que les méthodes de la famille des CIs peuvent être envisagées aux trois niveaux de description de Marr car elles explicitent la fonction de catégorisation, décrivent un processus plausible d'apprentissage de cette fonction et sont directement transposables sous la forme d'un réseau de neurones du type perceptron (Jäkel et al., 2009).

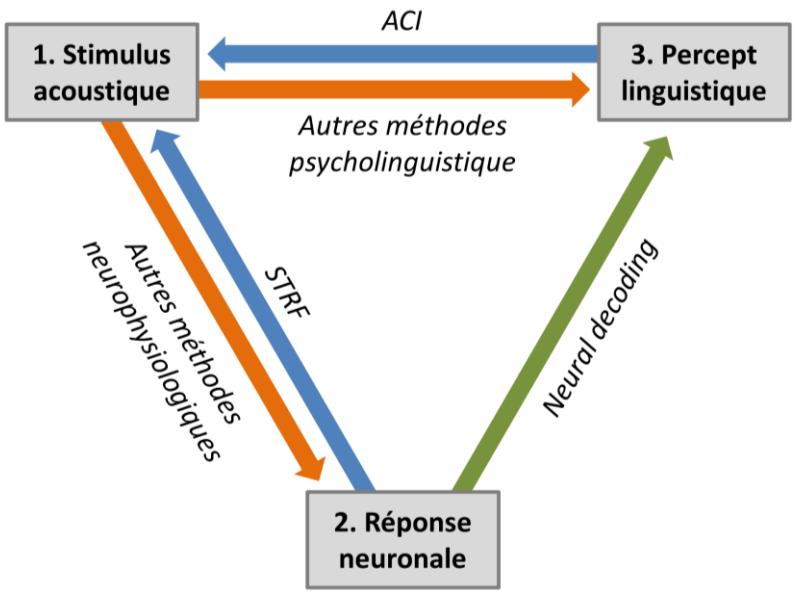


Figure 28 – Diagramme schématique des différentes méthodes psycholinguistiques et neurophysiologiques pour l'étude de la compréhension de la parole. Les 3 niveaux de représentation d'un son de parole sont présentés en gris : 1. Niveau acoustique, 2. Niveau neuronal et 3. Niveau linguistique. Les flèches indiquent les différentes méthodes reliant ces niveaux. Les approches permettant la mise en relation des niveaux neuronal et linguistique sont représentées par la couleur des flèches (voir texte pour une description plus détaillée).

La méthode des ACIs, combinée avec les STRFs, jette un nouveau pont entre réponse neuronale et percept linguistique (flèches bleues). En offrant une représentation commune dans l'espace des stimuli, elle permet une comparaison directe entre les unités élémentaires des deux niveaux (avec néanmoins les limitations évoquées au paragraphe précédent). De plus, la dérivation d'ACIs à partir de la représentation multi-résolution pourrait conduire au *mapping* direct du percept linguistique dans l'espace des réponses neuronales.

15. Conclusion générale

À l'issue de ce travail de recherche, nous avons pu élaborer et mettre en application l'ACI, un nouvel outil dans le champ de la psycholinguistique. Il implique un dispositif expérimental particulier (consistant essentiellement en une tâche de catégorisation de phonème dans le bruit) et un cadre statistique pour l'analyse des données recueillies (basé sur l'ajustement d'un GLM pénalisé). Ainsi, il devient possible de déterminer avec précision les indices acoustiques sur lesquels s'appuient les participants lors du décodage d'un signal parolier. Nous avons suggéré que cette nouvelle méthode peut trouver son utilité dans deux types de problématiques: la visualisation et la quantification des indices acoustiques soutenant une opposition phonémique particulière, d'une part, et la comparaison des stratégies d'écoutes employées par différents groupes d'auditeurs lors de la perception de phonèmes, d'autre part.

L'efficacité de cette approche a été illustrée à travers deux exemples de catégorisations : /aba/-/ada/ et /da/-/ga/ précédé de /al/ ou /aʁ/. Cette expérimentation a permis de confirmer l'implication des indices déjà décrits par les travaux précédents et, également, de mettre en évidence l'extraction d'autres indices, concomitants ou anticipatoires, non prévus initialement. Nous avons souligné la très bonne précision de cette méthode et le poids limité des connaissances a priori dans le calcul, qui lui offrent un champ d'application propre parmi les autres méthodes comportementales. La seconde phase de cette recherche avait pour objectif la mise en pratique de la nouvelle approche fournie par les ACIs pour comparer les stratégies d'écoutes de participants dyslexiques ou musiciens à celles d'un groupe contrôle. Pris dans leur ensemble, ces résultats suggèrent une importante plasticité des traitements effectués à l'interface acoustico-phonétique.

L'usage de cette technique est cependant limité par plusieurs facteurs, parmi lesquels la quantité de données nécessaires au calcul (de l'ordre de 6000 catégorisations par participant) et la non-stationnarité des traitements. Nous avons suggéré des voies à explorer pour réduire le temps de passation, notamment l'emploi d'autres types de régularisateurs. Nous avons également évoqué différentes modifications pouvant être apportées au paradigme expérimental, dans l'objectif de le rendre plus souple et, donc, potentiellement applicable dans des situations diverses, en variant le type de tâche, de bruit, ou encore le nombre de cibles. Enfin, les ACIs ont en outre l'avantage d'être directement comparables aux STRFs, tant du point de vue méthodologique qu'au niveau de l'espace de représentation, ce qui offre d'intéressantes perspectives pour de futurs travaux en combinaison avec la neuro-imagerie, avec l'ambition de relier les représentations linguistiques et neuronales.

Bibliographie

Abbey, C. K., & Eckstein, M. P. (2001). Theory for Estimating Human-Observer Templates in Two-Alternative Forced-Choice Experiments. In M. F. Insana & R. M. Leahy (éd.), *Information Processing in Medical Imaging* (p. 24‑35). Springer Berlin Heidelberg.

Abbey, C. K., & Eckstein, M. P. (2002). Classification image analysis: estimation and statistical inference for two-alternative forced-choice experiments. *Journal of Vision*, *2*(1), 66‑78.

Abbey, C. K., & Eckstein, M. P. (2006). Classification images for detection, contrast discrimination, and identification tasks with a common ideal observer. *Journal of Vision*, *6*(4), 335‑355.

Abdi, H. (2007). Signal Detection Theory (SDT). In *Encyclopedia of Measurement and Statistics*. Neil Salkind.

Ahumada, A. J., Jr. (1996). Perceptual classification images from vernier acuity masked by noise. In *ECVP'96 Abstracts.*

Ahumada, A. J., Jr. (2002). Classification image weights and internal noise level estimation. *Journal of Vision*, *2*(1), 121‑131.

Ahumada, A. J., Jr, & Beard, B. L. (1998). Response classification images in vernier acuity. *Investigative Ophthalmology and Visual Science*, *40*, S1109.

Ahumada, A. J., Jr, & Lovell, J. (1971). Stimulus Features in Signal Detection. *The Journal of the Acoustical Society of America*, *49*(6B), 1751‑1756.

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

Alexander, J. M., & Lutfi, R. A. (2004). Informational masking in hearing-impaired and normal-hearing listeners: sensation level and decision weights. *The Journal of the Acoustical Society of America*, *116*(4 Pt 1), 2234‑2247.

Allen, J. B. (1994). How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, *2*(4), 567‑577.

Allen, J. B. (1996). Harvey Fletcher's role in the creation of communication acoustics. *The Journal of the Acoustical Society of America*, *99*(4 Pt 1), 1825‑1839.

Allen, J. B. (2005). Consonant recognition and the articulation index. *The Journal of the Acoustical Society of America*, *117*(4 Pt 1), 2212‑2223.

Allen, J. B. (2006). *Articulation and Intelligibility*. Morgan & Claypool Publishers.

Allen, J. B. (2008). Nonlinear Cochlear Signal Processing and Masking in Speech Perception. In P. J. B. Dr, P. M. M. Sondhi, & P. Y. (Arden) H. Dr (éd.), *Springer Handbook of Speech Processing* (p. 27‑60). Springer Berlin Heidelberg.

Amin, N., Gill, P., & Theunissen, F. E. (2010). Role of the Zebra Finch Auditory Thalamus in Generating Complex Representations for Natural Sounds. *Journal of Neurophysiology*, *104*(2), 784‑798.

Apoux, F., & Bacon, S. P. (2004). Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise. *The Journal of the Acoustical Society of America*, *116*(3), 1671‑1680.

Apoux, F., & Healy, E. W. (2009). On the number of auditory filter outputs needed to understand speech: further evidence for auditory channel independence. *Hearing Research*, *255*(1-2), 99‑108.

Ardoint, M., Mamassian, P., & Lorenzi, C. (2007). Internal representation of amplitude modulation revealed by reverse correlation. In *In Abstract AR0 n 919. 30th ARO midwinter meeting, Feb 10-15, Denver, Colorado, USA*.

Atencio, C. A., Sharpee, T. O., & Schreiner, C. E. (2012). Receptive field dimensionality increases from the auditory midbrain to cortex. *Journal of Neurophysiology*, *107*(10), 2594‑2603.

Atiani, S., David, S. V., Elgueda, D., Locastro, M., Radtke-Schuller, S., Shamma, S. A., & Fritz, J. B. (2014). Emergent selectivity for task-relevant stimuli in higher-order auditory cortex. *Neuron*, *82*(2), 486‑499.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

Barth, E., Beard, B. L., & Ahumada, A. J., Jr. (1999). Nonlinear features in vernier acuity (Vol. 3644, p. 88‑96).

Beard, B. L., & Ahumada, A. J., Jr. (1998). Technique to extract relevant image features for visual tasks (Vol. 3299, p. 79‑85).

Beard, B. L., & Ahumada, A. J., Jr. (1999). Detection in fixed and random noise in foveal and parafoveal vision explained by template learning. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *16*(3), 755‑763.

Bellier, L., Mazzuca, M., Thai-Van, H., Caclin, A., & Laboissière, R. (2013). Categorization of speech in early auditory evoked responses. In *Proceeding of Interspeech 2013* (p. 911‑915).

Benkí, J. R. (2003). Analysis of English nonsense syllable recognition in noise. *Phonetica*, *60*(2), 129‑157.

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2010). Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction. *JSUR, 2010. 1(1):1-5 | Journal of Serendipitous and Unexpected Results*, *1*(1), 1‑5.

Bidelman, G. M., & Krishnan, A. (2010). Effects of reverberation on brainstem representation of speech in musicians and non-musicians. *Brain Research*, *1355*, 112‑125.

Boebinger, D., Evans, S., Rosen, S., Lima, C. F., Manly, T., & Scott, S. K. (2015). Musicians and non-musicians are equally adept at perceiving masked speech. *The Journal of the Acoustical Society of America*, *137*(1), 378‑387.

Boë, L.-J. (1997). Sciences phonétiques et relations forme/substance : 1. un siècle de ruptures, négociations et réorganisations. *Histoire Épistémologie Langage*, *19*(1), 5‑41.

Boë, L.-J., & Liénard, J. S. (1988). La communication parlée est-elle une science ? En doutiez-vous ? Éléments de discussion et de réflexion suivis de repères chronologiques. In *XVIIe JEP, GCP-SFA,* (p. 79‑92). Nancy, France.

Boets, B., Wouters, J., van Wieringen, A., & Ghesquière, P. (2006). Auditory temporal information processing in preschool children at family risk for dyslexia: relations with phonological abilities and developing literacy skills. *Brain and Language*, *97*(1), 64‑79.

Bogliotti, C., Serniclaes, W., Messaoud-Galusi, S., & Sprenger-Charolles, L. (2008). Discrimination of speech sounds by children with dyslexia: Comparisons with chronological age and reading level controls. *Journal of Experimental Child Psychology*, *101*(2), 137‑155.

Bouet, R., & Knoblauch, K. (2004). Perceptual classification of chromatic modulation. *Visual Neuroscience*, *21*(3), 283‑289.

Brady, S., Shankweiler, D., & Mann, V. (1983). Speech perception and memory coding in relation to reading ability. *Journal of Experimental Child Psychology*, *35*(2), 345‑367.

Brattico, E., Pallesen, K. J., Varyagina, O., Bailey, C., Anourova, I., Järvenpää, M., … Tervaniemi, M. (2009). Neural discrimination of nonprototypical chords in music experts and laymen: an MEG study. *Journal of Cognitive Neuroscience*, *21*(11), 2230‑2244.

Brett, M., Penny, W., & Kiebel, S. (2003). An Introduction to Random Field Theory. In *Human Brain Function* (Academic Press, 2nd edition.). R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny.

Brohé, S., Piccaluga, M., Delvaux, V., Huet, K., & Harmegnies, B. (2012). Orientation sélective de l'attention et apprentissage perceptuel.

Bronkhorst, A. W. (2000). The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions. *Acta Acustica united with Acustica*, *86*(1), 117‑128.

Calabrese, A., Schumacher, J. W., Schneider, D. M., Paninski, L., & Woolley, S. M. N. (2011). A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. *PloS One*, *6*(1), e16104.

Calandruccio, L., & Doherty, K. A. (2007). Spectral weighting strategies for sentences measured by a correlational method. *The Journal of the Acoustical Society of America*, *121*(6), 3827‑3836.

Calandruccio, L., & Doherty, K. A. (2008). Spectral weighting strategies for hearing-impaired listeners measured using a correlational method. *The Journal of the Acoustical Society of America*, *123*(4), 2367‑2378.

Cariani, P., & Micheyl, C. (2012). Towards a theory of information processing in the auditory cortex. In D. Poeppel, T. Overath, A. Popper, & R. R. Fay (éd.), *The Human Auditory Cortex*. New York: Springer.

Carré, R., Ainsworth, W. A., Jospa, P., Maeda, S., & Pasdeloup, V. (2001). Perception of vowel-to-vowel transitions with different formant trajectories. *Phonetica*, *58*(3), 163‑178.

Carré, R., & Divenyi, P. L. (2000). Modeling and perception of « gesture reduction ». *Phonetica*, *57*(2-4), 152‑169.

Chafcouloff, M. (2004). Voir la parole. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, *23*, 23‑65.

Chamak, B. (2011). Dynamique d'un mouvement scientifique et intellectuel aux contours flous : les sciences cognitives (États-Unis, France). *Revue d'histoire des sciences humaines*, *25*, 15‑36.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*(11), 1428‑1432.

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of Vision*, *5*(9), 659‑667.

Chechik, G., & Nelken, I. (2012). Auditory abstraction from spectro-temporal features to coding auditory entities. *Proceedings of the National Academy of Sciences*, *109*(46), 18968‑18973.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., & Shamma, S. (1999). Spectro-temporal modulation transfer functions and speech intelligibility. *The Journal of the Acoustical Society of America*, *106*(5), 2719‑2732.

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887‑906.

Christianson, G. B., Sahani, M., & Linden, J. F. (2008). The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *28*(2), 446‑455.

Clément, É., Linden, A.-C. J. V. der, Tiberghien, G., Linden, M. V. der, & Vivicorsi, B. (2013). Psychologie et économie de marché. *Bulletin de psychologie*, *Numéro 527*(5), 437‑440.

Coady, J. A., Kluender, K. R., & Rhode, W. S. (2003). Effects of contrast between onsets of speech and other complex spectra. *The Journal of the Acoustical Society of America*, *114*(4 Pt 1), 2225‑2235.

Cooper, F. S., Gaitenby, J. H., & Nye, P. W. (1984). Evolution of reading machines for the blind: Haskins Laboratories' research as a case history. *Journal of Rehabilitation Research and Development*, *21*(1), 51‑87.

Cooper, F. S., Liberman, A. M., & Borst, J. M. (1951). The Interconversion of Audible and Visible Patterns as a Basis for Research in the Perception of Speech. *Proceedings of the National Academy of Sciences of the United States of America*, *37*(5), 318‑325.

Dai, H., & Micheyl, C. (2010). Psychophysical reverse correlation with multiple response alternatives. *Journal of Experimental Psychology. Human Perception and Performance*, *36*(4), 976‑993.

Davidson, S. A., Gilkey, R. H., Colburn, H. S., & Carney, L. H. (2006). Binaural detection with narrowband and wideband reproducible noise maskers. III. Monaural and diotic detection and model results. *The Journal of the Acoustical Society of America*, *119*(4), 2258‑2275.

David, S. V., Mesgarani, N., & Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network (Bristol, England)*, *18*(3), 191‑212.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *23*(8), 3423‑3431.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1-2), 132‑147.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology. General*, *134*(2), 222‑241.

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *NeuroImage*, *24*(1), 21‑33.

Delattre, P. (1948). Un Triangle acoustique des voyelles orales du français. *French Review*, *21*(6), 477 ‑ 484.

Delattre, P. C. (1958). Acoustic cues in speech : first report. *Phonetica*, *2*, 108‑118, 226‑251.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic Loci and Transitional Cues for Consonants. *The Journal of the Acoustical Society of America*, *27*(4), 769‑773.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1964). Formant transitions and loci as acoustic correlates of place of articulation in american fricatives. *Studia Linguistica*, *16*(1-2), 104 ‑ 122.

Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word*, *8*(3), 195‑210.

Diependaele, K., Brysbaert, M., & Neri, P. (2012). How Noisy is Lexical Decision? *Frontiers in Psychology*, *3*, 348.

Doherty, K. A., & Lutfi, R. A. (1996). Spectral weights for overall level discrimination in listeners with sensorineural hearing loss. *The Journal of the Acoustical Society of America*, *99*(2), 1053‑1058.

Doherty, K. A., & Turner, C. W. (1996). Use of a correlational method to estimate a listener's weighting function for speech. *The Journal of the Acoustical Society of America*, *100*(6), 3769‑3773.

Dole, M., Hoen, M., & Meunier, F. (2012). Speech-in-noise perception deficit in adults with dyslexia: Effects of background type and listening configuration. *Neuropsychologia*, *50*(7), 1543‑1552.

Drullman, R., Festen, J. M., & Plomp, R. (1994a). Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, *95*(5 Pt 1), 2670‑2680.

Drullman, R., Festen, J. M., & Plomp, R. (1994b). Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America*, *95*(2), 1053‑1064.

Dufor, O., Serniclaes, W., Sprenger-Charolles, L., & Démonet, J.-F. (2009). Left premotor cortex and allophonic speech perception in dyslexia: a PET study. *NeuroImage*, *46*(1), 241‑248.

Eckstein, M. P., & Ahumada, A. J., Jr. (2002). Classification images: A tool to analyze visual strategies. *Journal of Vision*, *2*(1), i.

Eckstein, M. P., Ahumada, A. J., Jr, & Watson, A. B. (1997). Visual signal detection in structured backgrounds. II. Effects of contrast gain control, background variations, and white noise. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *14*(9), 2406‑2419.

Elhilali, M., Chi, T., & Shamma, S. A. (2003). A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech Communication*, *41*(2–3), 331‑348.

Eliason, S. R. (1993). *Maximum Likelihood Estimation: Logic and Practice*. SAGE Publications.

Embick, D., & Poeppel, D. (2015). Towards a computational(ist) neurobiology of language: Correlational, integrated, and explanatory neurolinguistics. *Language, Cognition and Neuroscience*, *30*(4), 357‑366.

Fletcher, H. (1922). The nature of speech and its interpretation. *Journal of the Franklin Institute*, *193*(6), 729‑747.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*(1), 3‑28.

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics*, *68*(2), 161‑177.

Fowler, C. A., Best, C. T., & McRoberts, G. W. (1990). Young infants' perception of liquid coarticulatory influences on following stop consonants. *Perception & psychophysics*, *48*(6), 559‑570.

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology. Human Perception and Performance*, *26*(3), 877‑888.

Fowler, C. A., & Rosenblum, L. D. (2014). The Perception of Phonetic Gestures. In *Modularity and the Motor theory of Speech Perception: Proceedings of A Conference To Honor Alvin M. Liberman*. Psychology Press.

Fox, J. (2008). Generalized linear models. In *Applied Regression Analysis and Generalized Linear Models Second* (Edition, SAGE Publications., p. 379‑424).

François, C., Jaillet, F., Takerkart, S., & Schön, D. (2014). Faster Sound Stream Segmentation in Musicians than in Nonmusicians. *PLoS ONE*, *9*(7), e101340.

Friederici, A. D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. *Trends in Cognitive Sciences*, *16*(5), 262‑268.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1‑22.

Friston, K. (2003). Introduction. Experimental design and Statistical Parametric Mapping. In *Human Brain Function* (R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny.). R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny.

Fritz, J., Shamma, S., Elhilali, M., & Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience*, *6*(11), 1216‑1223.

Fry, D., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The Identification and Discrimination of Synthetic Vowels. *Language and Speech*, *5*(4), 171‑189.

Fu, Q.-J., & Galvin, J. J., 3rd. (2003). The effects of short-term training for spectrally mismatched noise-band speech. *The Journal of the Acoustical Society of America*, *113*(2), 1065‑1072.

Furui, S. (1986). On the role of spectral transition for speech perception. *The Journal of the Acoustical Society of America*, *80*(4), 1016‑1025.

Gaab, N., Tallal, P., Kim, H., Lakshminarayanan, K., Archie, J. J., Glover, G. H., & Gabrieli, J. D. E. (2005). Neural correlates of rapid spectrotemporal processing in musicians and nonmusicians. *Annals of the New York Academy of Sciences*, *1060*, 82‑88.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361‑377.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, *6*(1), 110‑125.

Gaser, C., & Schlaug, G. (2003a). Brain Structures Differ between Musicians and Non-Musicians. *The Journal of Neuroscience*, *23*(27), 9240‑9245.

Gaser, C., & Schlaug, G. (2003b). Gray Matter Differences between Musicians and Nonmusicians. *Annals of the New York Academy of Sciences*, *999*(1), 514‑517.

Gilbert, G., & Micheyl, C. (2005). Influence of competing multi-talker babble on frequency-importance functions for speech measured using a correlational approach. *Acta Acustica United with Acustica*, *91*(1), 145‑154.

Gilbert, G., Micheyl, C., Berger Vachon, C., & Collet, L. (2002). Frequency-importance functions for speech in young and older listeners. In *Forum Acousticum*. Seville.

Gilkey, R. H., & Robinson, D. E. (1986). Models of auditory masking: a molecular psychophysical approach. *The Journal of the Acoustical Society of America*, *79*(5), 1499‑1510.

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511‑517.

Giraud, A.-L., & Ramus, F. (2013). Neurogenetics and auditory processing in developmental dyslexia. *Current Opinion in Neurobiology*, *23*(1), 37‑42.

Godement, R. (1992). Postface. Science, technologie, armement. In *Analyse mathématique II: Calcul différentiel et intégral, séries de Fourier, fonctions holomorphes* (2ème éd. corrigée 2003 edition.). Berlin ; New York: Springer.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251‑279.

Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology: CB*, *10*(11), 663‑666.

Gold, J. M., Sekuler, A. B., & Bennett, P. J. (2004). Characterizing perceptual learning with external noise. *Cognitive science*, *28*(2), 167‑207.

Goldstein, J. L. (1967). Auditory Nonlinearity. *The Journal of the Acoustical Society of America*, *41*(3), 676‑699.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research*, *41*(17), 2261‑2271.

Green, D. M. (1964). Consistency of auditory detection judgments. *Psychological Review*, *71*(5), 392‑407.

Grossetti, M., & Boë, L.-J. (2008). Sciences humaines et recherche instrumentale : qui instrumente qui ? *Revue d'Anthropologie des Connaissances*, *2*(1), 97‑114.

Harvey, L. O. (2004). *Detection Theory: Sensitivity and Response Bias*.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: effects of feedback and lexicality. *Journal of Experimental Psychology. Human Perception and Performance*, *34*(2), 460‑474.

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., & Carlyon, R. P. (2011). Generalization of perceptual learning of vocoded speech. *Journal of Experimental Psychology. Human Perception and Performance*, *37*(1), 283‑295.

Hickok, G. (2009). The functional neuroanatomy of language. *Physics of Life Reviews*, *6*(3), 121‑143.

Hickok, G. (2014). *The Myth of Mirror Neurons - The Real Neuroscience of Communication and Cognition*. New York: W. W. Norton & Company.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393‑402.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5 Pt 1), 3099‑3111.

Hoen, M., Meunier, F., Grataloup, C.-L., Pellegrino, F., Grimault, N., Perrin, F., … Collet, L. (2007). Phonetic and lexical interferences in informational masking during speech-in-speech comprehension. *Speech Communication*, *49*(12), 905‑916.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, *16*(4), 305‑312.

Holt, L. L. (2006a). Speech categorization in context: joint effects of nonspeech and speech precursors. *The Journal of the Acoustical Society of America*, *119*(6), 4016‑4026.

Holt, L. L. (2006b). The mean matters: effects of statistically defined nonspeech spectral distributions on speech categorization. *The Journal of the Acoustical Society of America*, *120*(5 Pt 1), 2801‑2817.

Holt, L. L., & Lotto, A. J. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research*, *167*(1-2), 156‑169.

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception & Psychophysics*, *72*(5), 1218‑1227.

Hoonhorst, I., Medina, V., Colin, C., Markessis, E., Radeau, M., Deltenre, P., & Serniclaes, W. (2011). Categorical perception of voicing, colors and facial expressions: A developmental study. *Speech Communication*, *53*(3), 417‑430.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences*, *13*(9), 381‑388.

Jakobson, R. (1961). *Preliminaries to Speech Analysis - The Distinctive Features & their Correlates*. Cambridge Mass.: MIT Press.

James E. Cutting, D. P. (1978). An information processing approach to speech perception. In *Speech and Language in the Laboratory, School, and Clinic* (p. 38‑72). J. F. Kavanagh, W. Strange.

Johnson, K. (2011). *Acoustic and Auditory Phonetics*. John Wiley & Sons.

Joosten, E. R. M., & Neri, P. (2012). Human pitch detectors are tuned on a fine scale, but are perceptually accessed on a coarse scale. *Biological Cybernetics*, *106*(8-9), 465‑482.

Kingdom, F. A. A., & Prins, N. (2010). *Psychophysics: a practical introduction*. London: Academic.

Kingston, J., Kawahara, S., Mash, D., & Chambless, D. (2011). Auditory contrast versus compensation for coarticulation: data from Japanese and English listeners. *Language and Speech*, *54*(Pt 4), 499‑525.

Kishon-Rabin, L., Amir, O., Vexler, Y., & Zaltz, Y. (2001). Pitch discrimination: are professional musicians better than non-musicians? *Journal of Basic and Clinical Physiology and Pharmacology*, *12*(2 Suppl), 125‑143.

Klatt, D. H. (1979). Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access. *Journal of Phonetics*, *7*, 279‑312.

Knoblauch, K., & Maloney, L. T. (2008). Estimating classification images with generalized linear and additive models. *Journal of Vision*, *8*(16), 10.1‑19.

Knoblauch, K., & Maloney, L. T. (2012). *Modeling Psychophysical Data in R*. Springer Science & Business Media.

Kontsevich, L. L., & Tyler, C. W. (2004). What makes Mona Lisa smile? *Vision Research*, *44*(13), 1493‑1498.

Kotz, S. A., & Schwartze, M. (2010). Cortical speech processing unplugged: a timely subcortico-cortical framework. *Trends in Cognitive Sciences*, *14*(9), 392‑399.

Krieger-Redwood, K., Gaskell, M. G., Lindsay, S., & Jefferies, E. (2013). The selective role of premotor cortex in speech perception: a contribution to phoneme judgements but not speech comprehension. *Journal of Cognitive Neuroscience*, *25*(12), 2179‑2188.

Kuhl, P. K. (1991). Human adults and human infants show a « perceptual magnet effect » for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93‑107.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831‑843.

Kuhl, P. K., Ramírez, R. R., Bosseler, A., Lin, J.-F. L., & Imada, T. (2014). Infants' brain responses to speech suggest Analysis by Synthesis. *Proceedings of the National Academy of Sciences*, *111*(31), 11238‑11245.

Kurki, I., & Eckstein, M. P. (2014). Template changes with perceptual learning are driven by feature informativeness. *Journal of Vision*, *14*(11), 6.

Kurki, I., Peromaa, T., Hyvärinen, A., & Saarinen, J. (2009). Visual Features Underlying Perceived Brightness as Revealed by Classification Images. *PLoS ONE*, *4*(10).

Kurki, I., Saarinen, J., & Hyvärinen, A. (2014). Investigating shape perception by classification images. *Journal of Vision*, *14*(12), 24.

Laudanski, J., Edeline, J.-M., & Huetz, C. (2012). Differences between Spectro-Temporal Receptive Fields Derived from Artificial and Natural Stimuli in the Auditory Cortex. *PLoS ONE*, *7*(11), e50539.

Law, J. M., Vandermosten, M., Ghesquiere, P., & Wouters, J. (2014). The relationship of phonological ability, speech perception, and auditory perception in adults with dyslexia. *Frontiers in Human Neuroscience*, *8*, 482.

Lehongre, K., Morillon, B., Giraud, A.-L., & Ramus, F. (2013). Impaired auditory sampling in dyslexia: further evidence from combined fMRI and EEG. *Frontiers in Human Neuroscience*, *7*.

Lehongre, K., Ramus, F., Villiermet, N., Schwartz, D., & Giraud, A.-L. (2011). Altered low-γ sampling in auditory cortex accounts for the three main facets of dyslexia. *Neuron*, *72*(6), 1080‑1090.

Lesica, N. A., & Grothe, B. (2008). Efficient Temporal Processing of Naturalistic Sounds. *PLoS ONE*, *3*(2), e1655.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431‑461.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology*, *65*(4), 497‑516.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, *68*(8), 1‑13.

Liberman, A. M., Harris, H. D., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358‑368.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1‑36.

Liberman, A. M., Mattingly, I. G., & Turvey, M. T. (1972). Language codes and memory codes. In *Coding Processes in Human Memory* (Winston and sons., p. 307‑334). Washington, DC: A. W. Melton and E. Martin.

Liberman, & Whalen. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, *4*(5), 187‑196.

Li, F., & Allen, J. B. (2009). Multiband product rule and consonant identification. *The Journal of the Acoustical Society of America*, *126*(1), 347‑353.

Li, F., & Allen, J. B. (2011). Manipulation of consonants in natural speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *19*(3), 496‑504.

Li, F., Menon, A., & Allen, J. B. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *The Journal of the Acoustical Society of America*, *127*(4), 2599‑2610.

Li, R. W., Klein, S. A., & Levi, D. M. (2006). The receptive field and internal noise for position acuity change with feature separation. *Journal of Vision*, *6*(4), 311‑321.

Li, R. W., Levi, D. M., & Klein, S. A. (2004). Perceptual learning improves efficiency by re-tuning the decision « template » for position discrimination. *Nature Neuroscience*, *7*(2), 178‑183.

Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, *33*(1).

Liu, J., Li, J., Feng, L., Li, L., Tian, J., & Lee, K. (2014). Seeing Jesus in toast: Neural and behavioral correlates of face pareidolia. *Cortex*, *53*, 60‑77.

Lobdell, B. E. (2009). *Models of human phone transcription in noise based on intelligibility predictors*. UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN.

Lobdell, B. E., Allen, J. B., & Hasegawa-Johnson, M. A. (2011). Intelligibility predictors and neural representation of speech. *Speech Communication*, *53*(2), 185‑194.

Loizou, P. C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *The Journal of the Acoustical Society of America*, *106*(4 Pt 1), 2097‑2103.

Lorenzi, C., Berthommier, F., Apoux, F., & Bacri, N. (1999). Effects of envelope expansion on speech recognition. *Hearing Research*, *136*(1-2), 131‑138.

Lotto, A. J. (2004). Perceptual compensation for coarticulation as a general auditory process. In *Proceedings of the 2003 Texas Linguistic Society Conference* (p. 42‑53). Sommerville, MA: A. Agwuele, W. Warren, & S-H. Park.

Lotto, A. J., & Holt, L. L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics*, *68*(2), 178–183.

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, *60*(4), 602–619.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica). *The Journal of the Acoustical Society of America*, *102*(2 Pt 1), 1134–1140.

Lotto, A. J., Sullivan, S. C., & Holt, L. L. (2003). Central locus for nonspeech context effects on phonetic identification (L). *The Journal of the Acoustical Society of America*, *113*(1), 53–56.

Machens, C. K., Wehr, M. S., & Zador, A. M. (2004). Linearity of cortical receptive fields measured with natural sounds. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *24*(5), 1089–1100.

Maddox, W. T., Molis, M. R., & Diehl, R. L. (2002). Generalizing a neuropsychological model of visual categorization to auditory categorization of vowels. *Perception & Psychophysics*, *64*(4), 584–597.

Magne, C., Schön, D., & Besson, M. (2006). Musician Children Detect Pitch Violations in Both Music and Language Better than Nonmusician Children: Behavioral and Electrophysiological Approaches. *Journal of Cognitive Neuroscience*, *18*(2), 199–211.

Mandel, M. I. (2013). Learning an intelligibility map of individual utterances. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

Mandel, M. I., Yoho, S. E., & Healy, E. W. (2014). Generalizing time-frequency importance functions across noises, talkers, and phonemes. In *Proceedings of Interspeech*.

Mangini, M. C., & Biederman, I. (2003). *Making the ineffable explicit: estimating the information employed for face classifications*.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–412.

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *The Journal of the Acoustical Society of America*, *69*(2), 548–558.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

Marr, D., & Poggio, T. A. (1976). From Understanding Computation to Understanding Neural Circuitry. *Neuroscience Research Program Bulletin*, *15*, 470–488.

Martin-Malivel, J., Mangini, M. C., Fagot, J., & Biederman, I. (2006). Do Humans and Baboons Use the Same Information When Categorizing Human and Baboon Faces? *Psychological Science*, *17*(7), 599–607.

Massaro, D. W., & Chen, T. H. (2008). The motor theory of speech perception revisited. *Psychonomic Bulletin & Review*, *15*(2), 453–457; discussion 458–462.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1‑86.

McGettigan, C., Rosen, S., & Scott, S. K. (2008). Investigating the perception of noise-vocoded speech - an individual differences approach. *Journal of the Acoustical Society of America*, *123*(5), 3330‑3330.

Mehr, M. A., Turner, C. W., & Parkinson, A. (2001). Channel weights for speech recognition in cochlear implant users. *The Journal of the Acoustical Society of America*, *109*(1), 359‑366.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental," in Pattern Recognition and Artificial Intelligence. In *Pattern recognition and artificial intelligence* (p. 374‑388). New-York: C. H. Chen, Ed.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014a). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, 1245994.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America*, *123*(2), 899‑909.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2014b). Mechanisms of noise robust representation of speech in primary auditory cortex. *Proceedings of the National Academy of Sciences*, *111*(18), 6792‑6797.

Mesgarani, N., Thomas, S., & Hermansky, H. (2010). A Multistream Multiresolution Framework for Phoneme Recognition. In *Proceedings of Interspeech*.

Mesgarani, N., Thomas, S., & Hermansky, H. (2011a). Adaptive Stream Fusion in Multistream Recognition of Speech. In *Proceedings of Interspeech, 2011*.

Mesgarani, N., Thomas, S., & Hermansky, H. (2011b). Toward optimizing stream fusion in multistream recognition of speech. *The Journal of the Acoustical Society of America*, *130*(1), EL14‑18.

Meyer, J. (2015). *Whistled Languages: A Worldwide Inquiry on Human Whistled Speech*. Springer.

Meyer, J., Dentel, L., & Meunier, F. (2010). Intelligibilité de la parole à plusieurs distances dans un bruit naturel. In *10ème Congrès Français d'Acoustique*.

Meyer, J., Dentel, L., & Meunier, F. (2013). Speech Recognition in Natural Background Noise. *PLoS ONE*, *8*(11), e79279.

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research*, *219*(1-2), 36‑47.

Miller, G. A., & Licklider, J. C. R. (1950). The Intelligibility of Interrupted Speech. *The Journal of the Acoustical Society of America*, *22*(2), 167‑173.

Miller, G. A., & Nicely, P. E. (1955). An Analysis of Perceptual Confusions Among Some English Consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338‑352.

Miller, L. M., Escabí, M. A., Read, H. L., & Schreiner, C. E. (2001). Functional convergence of response properties in the auditory thalamocortical system. *Neuron*, *32*(1), 151‑160.

Millman, R. E., Woods, W. P., & Quinlan, P. T. (2011). Functional asymmetries in the representation of noise-vocoded speech. *NeuroImage*, *54*(3), 2364‑2373.

Mineault, P. J., Barthelmé, S., & Pack, C. C. (2009). Improved classification images with sparse priors in a smooth basis. *Journal of Vision*, *9*(10), 17.1‑24.

Moore, B. C. J. (2002). Psychoacoustics of normal and impaired hearing. *British Medical Bulletin*, *63*, 121‑134.

Moreno, S., Marques, C., Santos, A., Santos, M., Castro, S. L., & Besson, M. (2009). Musical Training Influences Linguistic Abilities in 8-Year-Old Children: More Evidence for Brain Plasticity. *Cerebral Cortex*, *19*(3), 712‑723.

Morin Duchesne, X., Gosselin, F., Fiset, D., & Dupuis-Roy, N. (2014). Paper features: a neglected source of information for letter recognition. *Journal of Vision*, *14*(13), 11.

Murray, R. F. (2011). Classification images: A review. *Journal of Vision*, *11*(5).

Murray, R. F. (2012). Classification images and bubbles images in the generalized linear model. *Journal of Vision*, *12*(7), 2.

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2002). Optimal methods for calculating classification images: weighted sums. *Journal of Vision*, *2*(1), 79‑104.

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2005). Classification images predict absolute efficiency. *Journal of Vision*, *5*(2), 139‑149.

Nagai, M., Bennett, P. J., Rutherford, M. D., Gaspar, C. M., Kumada, T., & Sekuler, A. B. (2013). Comparing face processing strategies between typically-developed observers and observers with autism using sub-sampled-pixels presentation in response classification technique. *Vision Research*, *79*, 27‑35.

Nagai, M., Bennett, P. J., & Sekuler, A. B. (2008). Exploration of vertical bias in perceptual completion of illusory contours: Threshold measures and response classification. *Journal of Vision*, *8*(7), 25.1‑17.

Nandy, A. S., & Tjan, B. S. (2007). The nature of letter crowding as revealed by first- and second-order classification images. *Journal of vision*, *7*(2), 5.1‑526.

Nearey, T. M. (1997). Speech perception as pattern recognition. *The Journal of the Acoustical Society of America*, *101*(6), 3241‑3254.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, *135*, 370–384.

Nelken, I., Bizley, J., Shamma, S. A., & Wang, X. (2014). Auditory Cortical Processing in Real-World Listening: The Auditory System Going Real. *The Journal of Neuroscience*, *34*(46), 15135‑15138.

Neri, P., & Heeger, D. J. (2002). Spatiotemporal mechanisms for detecting and identifying image features in human vision. *Nature Neuroscience*, *5*(8), 812‑816.

Neri, P., & Levi, D. M. (2006). Receptive versus perceptive fields from the reverse-correlation viewpoint. *Vision Research*, *46*(16), 2465‑2474.

Neri, P., & Levi, D. M. (2008). Evidence for joint encoding of motion and disparity in human visual perception. *Journal of Neurophysiology*, *100*(6), 3117‑3133.

Neri, P., Parker, A. J., & Blakemore, C. (1999). Probing the human stereoscopic system with reverse correlation. *Nature*, *401*(6754), 695‑698.

Nève, F.-X. (2002). La phonologie a-t-elle inspiré le découpage « numérique » du langage humain ? *La linguistique*, *Vol. 38*(2), 89‑104.

Nickisch, H. (2012). glm-ie: generalised linear models inference & estimation toolbox. *J. Mach. Learn. Res.*, *13*, 1699–1703.

Noordenbos, M. W., Segers, E., Serniclaes, W., Mitterer, H., & Verhoeven, L. (2012a). Allophonic mode of speech perception in Dutch children at risk for dyslexia: A longitudinal study. *Research in Developmental Disabilities*, *33*(5), 1469‑1483.

Noordenbos, M. W., Segers, E., Serniclaes, W., Mitterer, H., & Verhoeven, L. (2012b). Neural evidence of allophonic perception in children at risk for dyslexia. *Neuropsychologia*, *50*(8), 2010‑2017.

Noordenbos, M. W., Segers, E., Serniclaes, W., & Verhoeven, L. (2013). Neural evidence of the allophonic mode of speech perception in adults with dyslexia. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *124*(6), 1151‑1162.

Noordenbos, M. W., & Serniclaes, W. (2015). The Categorical Perception Deficit in Dyslexia: A Meta-Analysis. *Scientific Studies of Reading*, *0*(0), 1‑20.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357‑395.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *The Behavioral and Brain Sciences*, *23*(3), 299‑325; discussion 325‑370.

Nuyts, J., & Fessler, J. A. (2003). A penalized-likelihood image reconstruction method for emission tomography, compared to postsmoothed maximum-likelihood with matched spatial resolution. *IEEE Transactions on Medical Imaging*, *22*(9), 1042‑1052.

Obleser, J., Eisner, F., & Kotz, S. A. (2008). Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *28*(32), 8116‑8123.

Obleser, J., & Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *NeuroImage*, *55*(2), 713‑723.

Obleser, J., Lahiri, A., & Eulitz, C. (2004). Magnetic brain response mirrors extraction of phonological features from spoken vowels. *Journal of Cognitive Neuroscience*, *16*(1), 31‑39.

Obleser, J., & Weisz, N. (2012). Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cerebral Cortex (New York, N.Y.: 1991)*, *22*(11), 2466‑2477.

O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1957). Acoustic cues for the perception of initial /w, j, r, l/ in English. *Word*, *13*(25-34).

Olasagasti, I., Bouton, S., & Giraud, A.-L. (2015). Prediction across sensory modalities: A neurocomputational model of the McGurk effect. *Cortex*, *68*, 61‑75.

Olman, C., & Kersten, D. (2004). Classification Objects, Ideal Observers & Generative Models. *Cognitive Science*, *28*(2), 227–239.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 156869.

Palm, G. (2012). *Novelty, Information and Surprise*. Springer Science & Business Media.

Paninski, L., Pillow, J., & Lewi, J. (2007). Statistical models for neural encoding, decoding, and optimal stimulus design. *Progress in Brain Research*, *165*, 493‑507.

Pantev, C., Oostenveld, R., Engelien, A., Ross, B., Roberts, L. E., & Hoke, M. (1998). Increased auditory cortical representation in musicians. *Nature*, *392*(6678), 811‑814.

Parbery-Clark, A., Skoe, E., & Kraus, N. (2009a). Musical Experience Limits the Degradative Effects of Background Noise on the Neural Processing of Sound. *The Journal of Neuroscience*, *29*(45), 14100‑14107.

Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009b). Musician enhancement for speech-in-noise. *Ear and Hearing*, *30*(6), 653‑661.

Parbery-Clark, A., Tierney, A., Strait, D. L., & Kraus, N. (2012). Musicians have fine-tuned neural distinction of speech syllables. *Neuroscience*, *219*, 111‑119.

Pascucci, D., Megna, N., Panichi, M., & Baldassi, S. (2011). Acoustic cues to visual detection: A classification image study. *Journal of Vision*, *11*(6), 7.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., … Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology*, *10*(1), e1001251.

Pasley, B. N., & Knight, R. T. (2013). Decoding speech for understanding and treating aphasia. *Progress in Brain Research*, *207*, 435‑456.

Patel, A. D. (2011). Why would Musical Training Benefit the Neural Encoding of Speech? The OPERA Hypothesis. *Frontiers in Psychology*, *2*, 142.

Pedersen, B., & Ellermeier, W. (2008). Temporal weights in the level discrimination of time-varying sounds. *The Journal of the Acoustical Society of America*, *123*(2), 963‑972.

Peterfalvi, J. M. (1966). La perception de la parole d'après les expériences de synthèse acoustique. *L'année psychologique*, *66*(2), 559‑577.

Phatak, S. A., & Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise. *The Journal of the Acoustical Society of America*, *121*(4), 2312‑2326.

Phatak, S. A., Lovitt, A., & Allen, J. B. (2008). Consonant confusions in white noise. *The Journal of the Acoustical Society of America*, *124*(2), 1220‑1233.

Phatak, S. A., Yoon, Y., Gooler, D. M., & Allen, J. B. (2009). Consonant recognition loss in hearing impaired listeners. *The Journal of the Acoustical Society of America*, *126*(5), 2683‑2694.

Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Mozart effect–Shmozart effect: A meta-analysis. *Intelligence*, *38*(3), 314‑323.

Pillow, J. W. (2007). Likelihood-Based Approaches to Modeling the Neural Code. In *Bayesian Brain: Probabilistic Approaches to Neural Coding* (K Doya, S Ishii, A Pouget & R Rao. MIT press., p. 53‑70).

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., & Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, *454*(7207), 995‑999.

Plas, R. (2011). La psychologie cognitive française dans ses relations avec les neurosciences. Histoire, enjeux et conséquences d'une alliance. *Revue d'histoire des sciences humaines*, (25), 125‑142.

Poeppel, D., Emmorey, K., Hickok, G., & Pylkkänen, L. (2012). Towards a new neurobiology of language. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(41), 14125‑14131.

Pota, S., Spinelli, E., Boulenger, V., Ferragne, E., Varnet, L., Hoen, M., & Meunier, F. (2012). La mie de pain n'est pas une amie : une étude EEG sur la perception de différences infraphonémiques en situation de variations. In *Actes de la conférence conjointe JEP-TALN-RECITAL*.

Pritchett, L. M., & Murray, R. F. (2015). Classification images reveal decision variables and strategies in forced choice tasks. *Proceedings of the National Academy of Sciences*, *112*(23), 7321‑7326.

Rabinowitz, N. C., Willmore, B. D. B., King, A. J., & Schnupp, J. W. H. (2013). Constructing Noise-Invariant Representations of Sound in the Auditory Pathway. *PLoS Biol*, *11*(11), e1001710.

Rammsayer, T., & Altenmüller, E. (2006). Temporal Information Processing in Musicians and Nonmusicians. *Music Perception: An Interdisciplinary Journal*, *24*(1), 37‑48.

Ramus, F. (2003). La dyslexie développementale : déficit phonologique spécifique ou trouble sensori-moteur global ? *Médecine & Enfance*, *23*(4), 255‑258.

Ramus, F., Rosen, S., Dakin, S. C., Day, B. L., Castellote, J. M., White, S., & Frith, U. (2003). Theories of developmental dyslexia: insights from a multiple case study of dyslexic adults. *Brain: A Journal of Neurology*, *126*(Pt 4), 841‑865.

Régnier, M. S., & Allen, J. B. (2008). A method to identify noise-robust perceptual features: application for consonant /t/. *The Journal of the Acoustical Society of America*, *123*(5), 2801‑2814.

Repp, B., & Liberman, A. (1984). Phonetic category boundaries are flexible. In *Collection* (p. 89‑112).

Rieth, C. A., Lee, K., Lui, J., Tian, J., & Huber, D. E. (2011). Faces in the mist: illusory face and letter detection. *i-Perception*, *2*(5), 458‑476.

Russ, B. E., Lee, Y.-S., & Cohen, Y. E. (2007). Neural and behavioral correlates of auditory categorization. *Hearing Research*, *229*(1-2), 204‑212.

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, *398*(6730), 760.

Sahani, M., & Linden, J. (2003). Evidence optimization techniques for estimating stimulus-response functions. In *Advances in Neural Information Processing Systems* (p. 301‑308). MIT Press.

Schellenberg, E. G., & Peretz, I. (2008). Music, language and cognition: unresolved issues. *Trends in Cognitive Sciences*, *12*(2), 45‑46.

Schneider, P., Scherg, M., Dosch, H. G., Specht, H. J., Gutschalk, A., & Rupp, A. (2002). Morphology of Heschl's gyrus reflects enhanced activation in the auditory cortex of musicians. *Nature Neuroscience*, *5*(7), 688‑694.

Schönfelder, V. H., & Wichmann, F. A. (2012). Sparse regularized regression identifies behaviorally-relevant stimulus features from psychophysical data. *The Journal of the Acoustical Society of America*, *131*(5), 3953‑3969.

Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *25*(5), 336‑354.

Seck, M. (2012). *Développement de la perception catégorielle des consonnes et implications pour l'étude de la dyslexie* (Thèse de doctorat). Université Paris Diderot - Paris 7, France.

Segui, J., Frauenfelder, U., & Mehler, J. (1981). Phoneme monitoring, syllable monitoring and lexical access. *British Journal of Psychology*, *72*(4), 471‑477.

Sekuler, A. B., Gaspar, C. M., Gold, J. M., & Bennett, P. J. (2004). Inversion leads to quantitative, not qualitative, changes in face processing. *Current Biology: CB*, *14*(5), 391‑396.

Serniclaes, W. (2000). La perception de la parole. In *La parole, des modèles cognitifs aux machines communicantes* (Hermès., p. 159‑190). Paris: P. Escudier, G. Feng, P. Perrier, J.-L. Schwartz, Eds.

Serniclaes, W. (2011). Features are phonological transforms of natural boundaries. In *Where Do Phonological Features Come From?: Cognitive, physical and developmental bases of distinctive speech categories* (John Benjamins Publishing Company., p. 237–258). Clements, G. Nick and Rachid Ridouane.

Serniclaes, W., & Arrouas, Y. (1995). Perception des traits phonétiques dans le bruit. *Verbum*, (2), 131‑144.

Serniclaes, W., & Seck, M. (2013). Perception of subphonemic segments: a new instance of allophonic perception in dyslexia. In *SSSR 20th annual meeting abstracts*. Hong Kong.

Serniclaes, W., Van Heghe, S., Mousty, P., Carré, R., & Sprenger-Charolles, L. (2004). Allophonic mode of speech perception in dyslexia. *Journal of Experimental Child Psychology*, *87*(4), 336‑361.

Serniclaes, W., & Wajskop, M. (1992). Chapter 5 : Phonetic versus acoustic account of feature integration in speech perception. In *Analytic Approaches to Human Cognition.* (Amsterdam: Elsevier Science Publishers.). Alegria, J., Holender, D., Junça De Morais, J., Radeau, M. (Eds.).

Shahin, A., Roberts, L. E., Pantev, C., Trainor, L. J., & Ross, B. (2005). Modulation of P2 auditory-evoked responses by the spectral complexity of musical sounds. *Neuroreport*, *16*(16), 1781‑1785.

Shankweiler, D., & Fowler, C. A. (2015). Seeking a reading machine for the blind and discovering the speech code. *History of Psychology*, *18*(1), 78‑99.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science*, *270*(5234), 303‑304.

Shinn, T. (2000). Formes de division du travail scientifique et convergence intellectuelle. La recherche technico-instrumentale. *Revue française de sociologie*, *41*(3), 447‑473.

Shub, D. E., & Richards, V. M. (2009). Psychophysical spectro-temporal receptive fields in an auditory task. *Hearing Research*, *251*(1-2), 1‑9.

Singh, S., & Singh, K. (2005). *Phonetics: Principles and Practices, Third Edition* (3rd edition.). San Diego: Plural Publishing, Inc.

Slifka, J. (2005). Acoustic Cues, Landmarks, and Distinctive Features: a Model of Human Speech Processing. In *Proceedings of the 6th International Symposium on Natural Language Processing*. Chang Rai, Thailand.

Solomon, J. A. (2002). Noise reveals visual mechanisms of detection and discrimination. *Journal of Vision*, *2*(1), 105‑120.

Sprenger-Charolles, L., Colé, P., Lacert, P., & Serniclaes, W. (2000). On subtypes of developmental dyslexia: evidence from processing time and accuracy scores. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Expérimentale*, *54*(2), 87‑104.

Stephens, J. D. W., & Holt, L. L. (2003). Preceding phonetic context affects perception of nonspeech. *The Journal of the Acoustical Society of America*, *114*(6 Pt 1), 3036‑3039.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, *111*(4), 1872‑1891.

Strait, D. L., Kraus, N., Parbery-Clark, A., & Ashley, R. (2010). Musical experience shapes top-down auditory mechanisms: evidence from masking and auditory attention performance. *Hearing Research*, *261*(1-2), 22‑29.

Strait, D. L., Parbery-Clark, A., Hittner, E., & Kraus, N. (2012). Musical training during early childhood enhances the neural encoding of speech in noise. *Brain and Language*, *123*(3), 191‑201.

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *The Journal of the Acoustical Society of America*, *90*(3), 1309‑1325.

Theunissen, F. E., & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Reviews Neuroscience*, *15*(6), 355‑366.

Thomas, J. P., & Knoblauch, K. (2005). Frequency and phase contributions to the detection of temporal luminance modulation. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *22*(10), 2257‑2261.

Thorsten Hansen, K. R. G. (2005). Classification images for chromatic signal detection. *Journal of the Optical Society of America. A, Optics, image science, and vision*, *22*(10), 2081‑9.

Tiberghien, G. (2007). Entre neurosciences et neurophilosophie: La psychologie cognitive et les sciences cognitives. *Psychologie Française*, *52*, 279‑297.

Tiberghien, G., & Jeannerod, M. (1995). Pour la science cognitive. La métaphore cognitive est-elle scientifiquement fondée? *Revue internationale de psychopathologie*, (18), 173‑203.

Toscano, J. C., & Allen, J. B. (2014). Across- and within-consonant errors for isolated syllables in noise. *Journal of Speech, Language, and Hearing Research: JSLHR*, *57*(6), 2293‑2307.

Treisman, M. (1999). There are two types of psychometric function: A theory of cue combination in the processing of complex stimuli with implications for categorical perception. *Journal of Experimental Psychology. General*, *128*(4), 517‑546.

Trevino, A., & Allen, J. B. (2013). Within-consonant perceptual differences in the hearing impaired ear. *The Journal of the Acoustical Society of America*, *134*(1), 607‑617.

Turner, C. W., Kwon, B. J., Tanaka, C., Knapp, J., Hubbartt, J. L., & Doherty, K. A. (1998a). Frequency-weighting functions for broadband speech as estimated by a correlational method. *The Journal of the Acoustical Society of America*, *104*(3 Pt 1), 1580‑1585.

Turner, C. W., Kwon, B. J., Tanaka, C., Knapp, J., Hubbartt, J. L., & Doherty, K. A. (1998b). Frequency-weighting functions for broadband speech as estimated by a correlational method. *The Journal of the Acoustical Society of America*, *104*(3 Pt 1), 1580‑1585.

van Boxtel, J. J. A., & Lu, H. (2015). Joints and their relations as critical features in action discrimination: evidence from a classification image method. *Journal of Vision*, *15*(1), 15.1.20.

Van Engen, K., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in Human Neuroscience*, *8*(577).

Van Tasell, D. J., Greenfield, D. G., Logemann, J. J., & Nelson, D. A. (1992). Temporal cues for consonant recognition: training, talker generalization, and use in evaluation of cochlear implants. *The Journal of the Acoustical Society of America*, *92*(3), 1247‑1257.

Varnet, L., Knoblauch, K., Meunier, F., & Hoen, M. (2013a). Show me what you listen to! Auditory classification images can reveal the processing of fine acoustic cues during speech categorization. In *Proceeding of Interspeech 2013* (p. 3167‑3171).

Varnet, L., Knoblauch, K., Meunier, F., & Hoen, M. (2013b). Using auditory classification images for the identification of fine acoustic cues used in speech perception. *Frontiers in Human Neuroscience*, *7:*, 865.

Varnet, L., Knoblauch, K., Serniclaes, W., Meunier, F., & Hoen, M. (2015a). A Psychophysical Imaging Method Evidencing Auditory Cue Extraction during Speech Perception: A Group Analysis of Auditory Classification Images. *PLoS ONE*, *10*(3), e0118009.

Varnet, L., Meunier, F., & Hoen, M. (2012a). Oscillations corticales et intelligibilité de la parole dégradée. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012* (p. 673–680).

Varnet, L., Meyer, J., Hoen, M., & Meunier, F. (2012b). Phoneme resistance during speech-in-speech comprehension. In *Proceeding of Interspeech 2012*.

Varnet, L., Serniclaes, W., Knoblauch, K., Meunier, F., & Hoen, M. (2014a). Identification of functional acoustic cues involved in speech perception: recent advances using Auditory Classification Images. In *Conference Proceedings* (p. 201). Amsterdam: Cambridge University Press.

Varnet, L., Trollé, G., Serniclaes, W., Knoblauch, K., Meunier, F., & Hoen, M. (2014b). Auditory Classification Images: How noise can reveal the acoustic cues used in phoneme categorization. In *Proceedings of the 6th SpiN Workshop*. Marseille, France.

Varnet, L., Wang, T., Peter, C., Meunier, F., & Hoen, M. (2015b). How musical expertise shapes speech perception: evidence from auditory classification images. *Scientific Reports*, *5*, 14489.

Vidne, M., Ahmadian, Y., Shlens, J., Pillow, J. W., Kulkarni, J., Litke, A. M., … Paninski, L. (2012). Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. *Journal of Computational Neuroscience*, *33*(1), 97‑121.

Viswanathan, N., Fowler, C. A., & Magnuson, J. S. (2009). A critical examination of the spectral contrast account of compensation for coarticulation. *Psychonomic Bulletin & Review*, *16*(1), 74‑79.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2010). Compensation for coarticulation: disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology. Human Perception and Performance*, *36*(4), 1005‑1015.

Vondrick, C., Pirsiavash, H., Oliva, A., & Torralba, A. (2014). Acquiring Visual Classifiers from Human Imagination. Massachusetts Institute of Technology.

Wang, T., Varnet, L., Peter, C., Estivalet, G., Meunier, F., & Hoen, M. (2014). How does musical expertise shape speech perception? Visual evidence from Auditory Classification Images. In *Conference Proceedings* (p. 202). Amsterdam: Cambridge University Press.

Warren, R. M. (1970). Perceptual Restoration of Missing Speech Sounds. *Science*, *167*(3917), 392‑393.

Watson, A. B., & Rosenholtz, R. (1997). A Rorschach test for visual classification strategies. *Investigative ophthalmology & visual science*, *38*(4), 2‑2.

Weiss, M. W., & Bidelman, G. M. (2015). Listening to the Brainstem: Musicianship Enhances Intelligibility of Subcortical Representations for Speech. *The Journal of Neuroscience*, *35*(4), 1687‑1691.

Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., & Kollmeier, B. (2005). Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines. In *in Proceedings of Interspeech* (p. 1273–1276).

Wichmann, F. A., Graf, A. B. A., Simoncelli, E. P., Bülthoff, H. H., & Schölkopf, B. (2005). Machine Learning Applied to Perception: Decision-Images for Gender Classification. In *Advances in Neural Information Processing Systems* (Eds. Lawrence K. Saul and Yair Weiss and Léon Bottou., Vol. 17).

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293‑1313.

Wild, C. J., Davis, M. H., & Johnsrude, I. S. (2012). Human auditory cortex is sensitive to the perceived clarity of speech. *NeuroImage*, *60*(2), 1490‑1502.

Willmore, B., & Smyth, D. (2003). Methods for first-order kernel estimation: simple-cell receptive fields from responses to natural scenes. *Network (Bristol, England)*, *14*(3), 553‑577.

Wolfe, J. (2002). Speech and music, acoustics and coding, and what music might be « for ». In *Proceedings of the 7th International Conference on Music Perception and Cognition, Sydney, 2002* (p. 10‑13).

Woodford, C., & Phillips, C. (2011). *Numerical Methods with Worked Examples: Matlab Edition*. Springer Science & Business Media.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36.

Woolley, S. M. N., Gill, P. R., Fremouw, T., & Theunissen, F. E. (2009). Functional Groups in the Avian Auditory System. *The Journal of Neuroscience*, *29*(9), 2780‑2793.

Woolley, S. M. N., Gill, P. R., & Theunissen, F. E. (2006). Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *26*(9), 2499‑2512.

World Health Organization. (2010). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines.* (World Health Organization., Vol. 2). Geneva.

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, *29*, 477‑505.

Xu, L., Thompson, C. S., & Pfingst, B. E. (2005). Relative contributions of spectral and temporal cues for phoneme recognition. *The Journal of the Acoustical Society of America*, *117*(5), 3255‑3267.

Zani, A., Biella, G., & Proverbio, A. M. (2003). Appendix F - Brain Imaging Techniques: Invasiveness and Spatial and Temporal Resolution. In A. Z. M. P. I. Posner (éd.), *The Cognitive Electrophysiology of Mind and Brain* (p. 417‑422). San Diego: Academic Press.

Zhao, L., & Zhaoping, L. (2011). Understanding Auditory Spectro-Temporal Receptive Fields and Their Changes with Input Statistics by Efficient Coding Principles. *PLoS Comput Biol*, *7*(8), e1002123.

Ziegler, J. C., Castel, C., Pech-Georgel, C., George, F., Alario, F.-X., & Perry, C. (2008). Developmental dyslexia and the dual route model of reading: simulating individual differences and subtypes. *Cognition*, *107*(1), 151‑178.

Ziegler, J. C., Pech-Georgel, C., George, F., & Lorenzi, C. (2009). Speech-perception-in-noise deficits in dyslexia. *Developmental Science*, *12*(5), 732‑745.

Ziegler, J. C., Pech-Georgel, C., George, F., & Lorenzi, C. (2011). Noise on, voicing off: Speech perception deficits in children with specific language impairment. *Journal of Experimental Child Psychology*, *110*(3), 362‑372.

Table des figures

Annexes

1. Stimuli utilisés pour l'expérience /alda/-/alga/-/aʁda/-/aʁga/

Tous les enregistrements sont au format .wav avec une fréquence d'échantillonnage de 48 kHz.

- Signaux-cibles : https://zenodo.org/record/12300
- Signaux de bruit du groupe contrôle (études 2, 3 et 4) : https://zenodo.org/record/23064
- Signaux de bruit du groupe musicien (étude 3) : https://zenodo.org/record/19104
- Signaux de bruit du groupe dyslexique (étude 4) : https://zenodo.org/record/19102

2. Données récoltées dans l'expérience /alda/-/alga/-/aʁda/-/aʁga/

Chaque fichier .mat contient toutes les données comportementales pour un participant sur les 10000 essais :

- n_signal : signal présenté (1:Alda, 2:Alga, 3:Arda, 4:Arga)
- correct_answer : 1 si le participant identifie correctement la cible (da ou ga), 0 dans le cas contraire.
- SNR : rapport signal sur bruit auquel la cible est présentée
- date : date de la passation de l'essai (année, mois, jour, heure, minute, seconde)
- stim_order : ordre (aléatoire) de présentation des signaux de bruit.

- Données des participants contrôle (études 2, 3 et 4): https://zenodo.org/record/21808
- Données des participants musiciens (étude 3) : https://zenodo.org/record/19134
- Données des participants dyslexiques (étude 4) : https://zenodo.org/record/19129

3. Caractéristiques individuelles des participants, et résultats des tests annexes

Sujet	Groupe	Sexe	Age	Edinburgh	Score	SNR moyen (dB)	RT moyen (s)
S6	NT	F	22	80	0,7796	-10,255772	1,26234679
S11	NT	M	20	90	0,7859	-11,282398	1,03901875
S14	NT	F	20	90	0,7859	-11,719925	1,15795582
S16	NT	M	21	50	0,786	-11,6313	1,33428151
S17	NT	F	22	100	0,7907	-11,642687	1,12805014
S18	NT	F	21	100	0,7881	-11,947949	1,21138326
S19	NT	F	24	80	0,787	-12,03746	1,37570125
S21	NT	F	33	90	0,7906	-13,033203	1,26894693
S22	NT	M	35	-100	0,7946	-12,807765	1,60299328
S23	NT	M	19	100	0,7916	-11,03773	1,14859132
S25	NT	F	21	60	0,785	-11,528701	1,26088368
S26	NT	F	20	100	0,7891	-11,600695	1,34775308
S27	NT	F	20	80	0,7874	-10,918121	1,27800557
S28	NT	F	22	-60	0,7796	-11,105786	1,20110146
S29	NT	F	21	50	0,7923	-13,988872	1,25519255
S30	NT	M	20	50	0,7891	-10,897883	1,47066213
S31	NT	F	23	80	0,7906	-12,51431	1,33487785
S32	NT	M	27	90	0,7955	-14,264122	1,42933584
SD1	DYS	F	24	100	0,782	-7,5681522	1,4200769
SD2	DYS	F	44	90	0,7945	-11,324068	1,58800275
SD3	DYS	F	24	90	0,788	-10,873686	1,39197093
SD5	DYS	F	18	80	0,787	-9,2748444	1,39492416
SD6	DYS	M	20	60	0,7806	-7,5915757	1,30129579
SD7	DYS	F	19	100	0,7834	-7,9880045	1,38316294
SD8	DYS	F	18	40	0,7899	-11,970656	1,36222689
SD9	DYS	F	18	60	0,7879	-12,14018	1,24070981
SD10	DYS	M	23	60	0,7881	-11,30083	1,21552445
SD11	DYS	F	20	100	0,7889	-13,143949	1,48001283
SD12	DYS	M	20	100	0,7903	-11,073469	1,37528783
SD13	DYS	F	23	80	0,7895	-11,607291	1,21840579
SD14	DYS	M	21	-20	0,7889	-9,9409067	1,40802691
SD15	DYS	F	19	80	0,7862	-9,9208986	1,35992752
SD16	DYS	M	20	80	0,7924	-13,12634	1,34993915
SD18	DYS	M	32	40	0,7883	-9,9932246	1,40322523
SD19	DYS	M	29	100	0,7914	-13,31807	1,48351284
SD21	DYS	F	19	80	0,7808	-8,4649466	1,38156612
ST1	M	M	21	85	0,7902	-12,29367	1,10429364
ST2	M	M	26	100	0,7992	-14,832448	1,35147683
ST3	M	F	19	100	0,7966	-17,183129	1,07769612
ST4	M	M	21	100	0,79	-11,773793	1,24779513
ST5	M	F	22	90	0,7945	-13,181471	1,41010204
ST6	M	F	22	90	0,7848	-12,515747	1,11476056
ST7	M	M	31	100	0,797	-13,4672	1,3238759
ST8	M	F	21	-75	0,7899	-13,357419	1,37525151
ST9	M	M	24	100	0,7938	-13,970307	1,46228478
ST10	M	F	22	100	0,7887	-12,101202	1,08446214
ST11	M	M	27	90	0,7968	-14,113291	1,63680275
ST12	M	F	20	100	0,7915	-12,848288	1,23630583
ST13	M	F	21	95	0,7918	-12,710219	1,58362289
ST14	M	F	21	-100	0,7944	-12,183375	1,6489431
ST15	M	F	26	70	0,7923	-13,608396	1,48507873
ST17	M	M	23	80	0,7911	-13,32782	1,46002857
ST18	M	M	22	100	0,7913	-12,76381	1,42620175
ST19	M	M	24	80	0,7946	-13,637769	1,61237132
ST20	M	F	24	100	0,7962	-14,237508	1,43772846

Table 2 – Résumé des caractéristiques individuelles des participants et résultats moyens à l'expérience principale

Sujet	Raven (s/60)	Alouette (MCLM)	Alouette (nbr erreurs)	Alouette (tps)	Suppression phono. (s/10)	Suppression phono. (tps)	Contrepèteries (s/20)	Contrepèteries (tps)	Dictée texte (usage, s/10)	Dictée texte (accords, s/10)	Lect. mots réguliers (s/20)	Lect. mots réguliers (tps)	Lect. mots irréguliers (s/20)	Lect. mots irréguliers (tps)	Lect. non-mots (s/20)	Lect. non-mots (tps)	Dict. mots réguliers (s/20)	Dict. mots réguliers (tps)	Dict. mots irréguliers (s/10)	Dict. mots irréguliers (tps)	Dict. de non-mots (s/10)	Dict. non-mots (tps)	Répétition non-mots (s/20)	Mém. de W, endroit (s/9)	Mem. de W, envers (s/9)
S6	51	156	4	100	10	22	20	52	9	10	20	14	20	14	20	23	9	44	6	43	10	48	20	8	8
S11	58	180	3	87	10	26	16	58	9	9	20	10	20	9	18	14	8	52	7	47	10	42	20	7	6
S14	48	171	2	92	5	35	20	63	10	8	20	15	18	16	20	16	9	40	8	42	9	40	19	8	6
S16	54	226	1	70	10	27	20	47	10	9	20	9	18	14	20	13	10	31	10	32	8	41	20	9	8
S17	46	167	2	94	10	25	19	126	9	9	20	15	19	11	20	21	9	36	8	36	9	46	20	6	5
S18	54	177	5	88	10	30	19	67	8	8	20	12	18	13	10	21	10	41	6	67	10	52	20	8	6
S19	49	183	14	82	10	34	19	93	9	8	20	13	18	15	18	22	9	37	9	34	10	42	19	7	5
S21	50	187	5	83	10	24	18	82	10	8	20	12	20	13	19	18	9	32	8	31	10	40	19	6	5
S22	42	152	1	104	10	24	20	64	10	9	19	16	19	13	20	20	9	41	9	41	10	50	18	6	5
S23	57	197	8	78	10	26	19	65	9	9	20	12	19	10	18	22	8	45	5	48	10	49	19	7	5
S25	47	176	9	87	10	22	19	48	10	9	19	12	19	11	19	21	10	39	9	49	10	37	20	9	8
S26	49	174	4	90	10	38	20	59	9	8	20	11	20	10	20	19	9	36	8	52	9	39	20	6	6
S27	49	231	3	68	10	31	19	85	10	8	19	10	18	12	18	16	10	44	9	47	8	51	19	7	6
S28	50	207	5	75	10	20	19	69	10	8	20	12	20	11	19	17	10	41	8	45	10	49	18	6	4
S29	56	206	7	75	9	23	20	38	9	8	19	10	20	10	20	15	8	39	6	44	10	47	20	8	8
S30	51	160	1	99	10	26	19	70	9	9	19	12	19	16	20	20	8	39	8	31	4	49	20	7	6
S31	48	175	4	89	9	44	20	79	9	10	20	11	19	11	19	22	9	32	8	36	6	52	20	8	6
S32	56	237	0	67	3	23	20	81	8	9	20	9	20	10	20	13	8	28	9	32	5	40	20	7	7
SD1	52	126	14	119	7	39	18	146	7	4	18	23	15	26	16	42	8	50	4	67	6	70	18	4	3
SD2	56	109	5	142	10	40	19	139	8	9	19	18	20	17	18	61	8	57	5	60	8	90	20	6	5
SD3	42	119	16	125	6	69	17	167	6	8	18	17	19	13	17	28	8	45	4	65	8	57	18	6	5
SD5	47	120	20	113	4	38	15	120	10	4	19	25	20	26	16	39	9	45	7	48	8	53	20	7	5
SD6	47	120	11	126	5	37	16	101	4	1	19	14	19	11	18	19	6	56	4	76	4	99	17	7	4
SD7	45	103	3	148	5	45	16	241	6	7	19	38	15	31	18	69	8	53	3	58	9	68	16	5	4
SD8	53	154	8	140	9	43	18	133	2	8	19	22	16	23	19	34	7	50	4	63	8	54	19	6	5
SD9	56	65	5	180	7	68	4	298	6	7	18	39	18	36	6	63	7	47	3	41	7	59	20	6	3
SD10	49	157	13	96	3	49	16	82	7	3	20	14	20	14	15	19	6	39	4	36	9	37	20	7	3
SD11	53	96	35	143	9	38	11	189	4	3	19	29	15	25	14	39	5	61	0	67	7	66	20	8	7
SD12	53	143	4	109	10	36	16	80	8	5	20	14	20	16	20	26	7	44	5	44	8	48	17	7	7
SD13	55	151	22	87	6	51	19	104	7	8	18	24	18	27	20	34	7	49	5	47	6	49	19	6	4
SD14	44	93	24	155	4	50	7	169	5	4	19	30	15	26	17	35	4	55	2	46	7	50	18	6	3
SD15	44	94	11	162	9	50	16	122	7	9	19	27	19	25	19	45	9	51	4	53	8	61	18	7	5
SD16	51	132	8	116	6	65	18	116	7	7	20	14	20	13	20	29	7	41	4	55	9	60	20	6	6
SD18	51	144	7	107	6	40	16	162	7	7	20	12	18	13	16	20	6	39	7	43	4	46	18	4	4
SD19	44	170	4	92	8	30	20	78	8	9	20	10	20	9	20	18	9	40	9	39	6	51	20	6	4
SD21	52	137	9	11	3	36	17	104	9	8	20	15	18	13	20	27	8	49	5	52	5	52	19	7	5

Table 3 – Résultats des groupes contrôle et dyslexique aux tests de lecture

Sujet	Alerting effect	Orienting effect	Conflict effect
S6	41	36	128
S11	45	80	105
S14	85	21	158
S16	44	34	97
S17	-5	26	123
S18	26	36	132
S19	43	46	134
S21	67	78	132
S22	28	42	70
S23	18	36	245
S25	41	13	148
S26	37	18	180
S27	21	36	119
S28	26	85	130
S29	-5	59	103
S30	10	31	129
S31	47	12	120
S32	26	10	95
SD1	-10	28	191
SD2	30	65	114
SD3	80	65	107
SD5	70	72	148
SD6	72	52	109
SD7	0	47	252
SD8	34	44	164
SD9	-3	67	113
SD10	39	25	169
SD11	54	44	147
SD12	42	49	229
SD13	13	62	199
SD14	5	64	133
SD15	65	42	117
SD16	21	31	117
SD18	23	52	176
SD19	24	62	125
SD21	33	83	187
ST1	-4	29	82
ST2	3	65	120
ST3	20	21	82
ST4	36	49	161
ST5	59	70	189
ST6	16	28	115
ST7	41	28	145
ST8	86	55	78
ST9	2	39	111
ST10	44	49	197
ST11	5	33	117
ST12	44	46	160
ST13	16	49	93
ST14	41	59	117
ST15	34	49	121
ST17	39	34	119
ST18	18	68	114
ST19	51	39	102
ST20	18	65	280

Table 4 – Résultats du test ANT

4. Scripts et fonctions Matlab

Certains des scripts et fonctions utilisés durant cette thèse sont disponibles à l'adresse https://zenodo.org/record/29426, sous la forme d'une *toolbox* Matlab simplifiée. Celle-ci permet la mise en place et la passation d'une expérience de type ACI, l'analyse des réponses des participants, et finalement le calcul et l'affichage des ACIs.

Brève description des scripts :

1) Script de génération des stimuli (Script1_Initialisation.m)

Mise en place de la passation pour un participant: création d'un dossier contenant les signaux à catégoriser et d'un dossier contenant les bruits.

2) Script de passation (Script2_Passation.m)

Paramétrage de l'expérience et lancement de la passation. Permet également de simuler les réponses d'un participant par *template-matching*.

3) Script d'analyse des réponses (Script3_Analyse_reponses.m)

Analyse des réponses d'un groupe de participants (scores, SNR, fonctions psychométriques,...)

4) Script de calcul de l'ACI (Script4_Calcul_ACI.m)

Algorithme de calcul et d'affichage de l'ACI pour un participant, basé sur le GLM pénalisé. Un certain nombre de paramètres peuvent être ajustés : représentation en spectrogramme ou en cochléogramme, images dérivées des bruits ou des stimuli complets…