Modélisation de l'attention sélective auditive à l'aide de réseaux de neurones artificiels.

Auteur : Balla, Marion
Promoteur(s) : Sougné, Jacques
Faculté : þÿ F a c u l t é d e P s y c h o l o g i e , L o g o p é d i e e t S c i e n c e s d e l E d u c a t i o n
Diplôme : Master en sciences psychologiques, à finalité spécialisée
Année académique : 2024-2025
URI/URL : http://hdl.handle.net/2268.2/24741

Modélisation de l'attention sélective auditive à l'aide de réseaux de neurones artificiels

BALLA Marion

Mémoire de fin d'études en vue de l'obtention d'un master en sciences psychologiques, à finalité spécialisée

Promoteur : Jacques Sougné

Lecteurs : Robert French et John Read

Faculté de Psychologie, Logopédie et Sciences de l'éducation

Résumé

La modélisation de l'attention sélective auditive a été peu investiguée dans la littérature en psychologie cognitive. Le *cocktail party effect*, c'est-à-dire le fait que notre attention soit captée involontairement par un mot prononcé dans une autre conversation que celle à laquelle on participe, constitue un défi pour les modèles actuels. La difficulté réside dans la création d'un modèle gérant les phénomènes *top-down* et *bottom-up* en contexte bruyant. Ce mémoire vise à étudier comment différentes architectures de réseaux de neurones (LSTM vs. BiLSTM) modélisent ce phénomène, notamment face à une distraction sémantique.

Nous avons implémenté deux architectures : le modèle ASAM (Xu et al., 2018) basé sur un réseau BiLSTM, et une variante LSTM standard. Cette comparaison visait à évaluer le gain en performance d'un modèle complexe (BiLSTM), théoriquement supérieur, face à un modèle plus simple (LSTM) dont le traitement causal est conceptuellement plus proche de la cognition humaine. Les modèles ont été entraînés sur un jeu de données original conçu pour manipuler la familiarité de mots spécifiques. Lors de la phase de test, les mots rendus familiers étaient intégrés dans le discours du distracteur afin de mesurer leur potentiel de distraction. Les performances ont été évaluées à l'aide de métriques de séparation de sources et analysées par une MANOVA.

Les résultats révèlent deux conclusions majeures. Premièrement, le modèle LSTM a surpassé de manière contre-intuitive le modèle BiLSTM, un résultat probablement attribuable à un sur-apprentissage du modèle plus complexe sur notre jeu de données de taille limitée. Deuxièmement, bien que les deux architectures aient apparement manifesté un effet *cocktail party* (une dégradation des performances en présence de distracteurs), cet effet n'était pas modulé par la fréquence d'exposition. Cette découverte suggère que l'interférence était principalement due à un masquage énergétique (basé sur les propriétés acoustiques des mots) plutôt qu'à un masquage sémantique ou informationnel (basé sur la familiarité et le sens des mots).

Ce mémoire met donc en lumière une limite critique des modèles computationnels actuels de l'attention auditive, suggérant qu'ils reposent davantage sur l'apprentissage de motifs acoustiques sophistiqués que sur un véritable traitement sémantique. Il appelle à une investigation plus critique sur la manière d'intégrer et de tester les influences sémantiques dans les modèles de l'attention sélective auditive.

Remerciements

Au terme de ces deux années de travail consacrées à la réalisation de ce mémoire, je souhaite adresser mes plus sincères remerciements à toutes les personnes qui ont contribué, de près ou de loin, à son aboutissement.

Je tiens à exprimer ma plus profonde gratitude à Monsieur Sougné, mon promoteur de mémoire. Son accompagnement constant, sa disponibilité et ses conseils avisés tout au long de ces deux années ont été une source de motivation et de rigueur indispensable à la conduite de ce projet. Sa confiance et ses orientations précieuses ont guidé ma réflexion et m'ont permis de mener ce travail à son terme.

Je remercie également les membres du jury d'avoir accepté d'évaluer ce travail et pour le temps qu'ils y consacreront.

Mes remerciements s'adressent également à Monsieur Huckvale, pour m'avoir généreusement donné accès aux codes de son logiciel Audio3DServer. Cette ressource a été fondamentale pour le développement du jeu de données présenté dans ce mémoire.

Je souhaite remercier chaleureusement Jean-Philippe pour son aide précieuse face aux complexités du modèle informatique. C'est également son soutien infaillible et sa patience durant ces deux années qui m'ont donné la force de persévérer.

Je remercie tendrement mes parents pour leur soutien constant, leurs encouragements et leur persévérance à tenter de comprendre les subtilités de mon travail.

Enfin, une pensée spéciale est destinée à mes animaux de compagnie, qui m'ont efficacement empêchée de procrastiner en s'endormant sur moi pendant des heures lorsque j'étais assise à mon bureau. Leur contribution au respect des délais fut, à leur manière, essentielle.

Table des matières

Introduction

> *"Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought... It implies withdrawal from some things in order to deal effectively with others."*
>
> - William James

L'attention est l'un des objets d'étude les plus anciens des sciences humaines. Dès l'Antiquité, Aristote et Platon s'interrogeaient sur la manière dont l'esprit sélectionne certaines informations sensorielles au détriment d'autres. Au fil du temps, la psychologie expérimentale, la psychologie cognitive ou encore les neurosciences ont approfondi notre compréhension de l'attention, en explorant ses sous-bassement cérébraux ainsi que son influence sur le comportement humain dans divers contextes. La vision, dominant notre perception et notre action, s'est imposée comme modalité privilégiée dans la recherche expérimentale sur l'attention. En comparaison, l'attention auditive est restée moins étudiée. Pourtant, la sélection auditive fait partie intégrante de nos expériences sociales, les environnements dans lesquels nous évoluons étant souvent bruyants et multi-sources. Comprendre comment l'humain isole une source d'intérêt dans un contexte sonore encombré est ainsi crucial.

De ce fait, le *cocktail party effect* illustre bien ce défi. Il désigne la capacité à suivre un interlocuteur spécifique dans un environnement bruyant, tout en étant parfois distrait involontairement par des éléments saillants des discours environnants. Ce phénomène articule d'un côté les mécanismes de sélection guidés par nos buts personnels, de l'autre les mécanismes involontaires, et incarne la tension centrale entre les deux. Les modèles théoriques de l'attention ont structuré la réflexion afin d'expliquer ce phénomène. Malgré leur intérêt, ces modèles restent descriptifs et peinent à simuler les dynamiques fines du traitement attentionnel dans des contextes complexes. À l'inverse, les modèles computationnels fondés sur les réseaux de neurones artificiels permettent de tester des hypothèses de manière opérationnelle, en simulant l'émergence de comportements attentionnels dans des environnements contrôlés.

Ce mémoire vise à articuler ces approches en modélisant l'attention sélective auditive à l'aide de réseaux de neurones artificiels. Plusieurs modèles seront appliqués à une tâche de type *cocktail party*, afin d'identifier celui reproduisant au mieux les dynamiques de la sélection attentionnelle humaine. Une analyse complémentaire portera sur les facteurs sémantiques susceptibles d'influencer l'émergence du *cocktail party effect*.

Part I

Revue de littérature

1 L'attention sélective auditive

1.1 Définitions

L'attention est un processus cognitif particulier, en ce sens qu'il ne permet pas le traitement d'un type d'information spécifique, mais qu'il aide à choisir l'objet de traitement (Banich and Compton, 2023). L'attention sélective est la capacité à se concentrer sur une tâche ou un type d'information en ignorant le reste (Goldstein, 2015). Il s'agit d'un processus dit *top-down*, ou endogène, car notre objet d'attention dépend de nos buts et motivations personnels, et est donc choisi de façon volontaire. Par exemple, l'attention sélective intervient lorsque l'on se trouve dans un environnement bruyant et que l'on souhaite suivre une conversation particulière. Notre objectif est de capter le discours de notre interlocuteur, tout en ignorant les sons environnants.

Ce processus d'organisation perceptuelle de notre environnement auditif est nommé *auditory scene analysis* (ASA), et a été décrit par Albert Bregman en 1994. Il s'agit de notre capacité à séparer les sons environnants en leurs différentes sources, puis de les intégrer afin de former des stimuli significatifs. La théorie de Bregman décrit l'existence de deux systèmes distincts. Le *primitive auditory scene analysis* permet de créer des liens entre les stimuli acoustiques qui auraient pu provenir d'une même source, en se basant sur les caractéristiques sensorielles et avant tout traitement attentionnel, contrôlé ou automatique. Le *schema-based organization* permet de tenir compte de l'influence des connaissances préalables et des expériences auditives passées sur la perception auditive. La combinaison de ces deux processus permet de retracer l'origine des sons, et ainsi, réaliser efficacement une tâche d'écoute sélective.

Malgré nos excellentes capacités d'attention sélective auditive, il peut arriver qu'une situation devienne plus complexe. Notamment, dans un environnement bruyant, il est possible que notre attention soit captée involontairement par un mot prononcé dans une autre conversation que celle à laquelle on participe. Ce phénomène est connu dans la littérature sous le nom de *cocktail party effect* (Cherry,

1953). Il s'agit d'un processus dit *bottom-up*, ou exogène, car notre attention est involontairement détournée par un élément de l'environnement. Des stimuli saillants, comme notre nom, sont perçus dans les discours alentours et attirent notre attention de façon automatique.

La littérature distingue deux grands types de caractéristiques des stimuli saillants entraînant une capture involontaire de l'attention : acoustiques et sémantiques. Les caractéristiques acoustiques concernent les aspects de forme et de fréquence du son, tandis que les caractéristiques sémantiques concernent la signification du son, et sont liées à des aspects plus abstraits. Par exemple, Ylinen et al. (2022) ont mis en évidence que, lorsqu'on présente deux sons différents à des participants, un dans chaque oreille, la qualité de l'audio influence la capacité à se focaliser sur le stimulus cible. Il s'agit d'une caractéristique purement acoustique du son qui interfère avec les capacités d'attention sélective des sujets. À l'inverse, Moray (1959) a mis en évidence que les caractéristiques sémantiques d'un son, c'est-à-dire la signification des mots présentés, peuvent influencer les performances des sujets. Ces deux dimensions, acoustique et sémantique, sont en réalité profondément intriquées, comme en témoignent les résultats de Har-shai Yahav et al. (2024) mettant en évidence l'influence d'une voix familière, mobilisant à la fois des indices sensoriels et des connaissances mémorielles, sur les capacités d'attention sélective.

Comprendre l'attention sélective auditive nécessite ainsi de prendre en compte l'ensemble des mécanismes endogènes (*top-down*) et exogènes (*bottom-up*) intervenant dans un contexte de compétition auditive, ainsi que les propriétés des stimuli et de l'environnement susceptibles de moduler la sélection attentionnelle.

1.2 Substrats neuronaux

1.2.1 Régions cérébrales

La littérature a mis en évidence plusieurs régions cérébrales associées à l'attention sélective. Au niveau visuel, le colliculus supérieur joue un rôle dans l'orientation automatique de l'attention via le contrôle des saccades oculaires (Banich and Compton, 2023). Le *Frontal Eye Field* (FEF), responsable du contrôle volontaire des saccades (Vernet et al., 2014), envoie des feedbacks au colliculus supérieur afin de l'aider à diriger les mouvements oculaires dans la direction des objets d'intérêt attentionnel. Au niveau auditif, les recherches mettent en évidence un rôle similaire du colliculus inférieur dans l'orientation automatique

de l'attention. Cependant, il semblerait que la distinction entre les deux modalités ne soit pas aussi évidente, les deux colliculi étant impliqués dans l'orientation attentionnelle vers les stimuli saillants, toutes modalités confondues (Hu and Dan, 2022).

Comme autre structure importante, on retrouve le thalamus qui permet de filtrer l'information sensorielle pertinente (Saalmann and Kastner, 2014) via deux sous-régions clés : le corps genouillé latéral, qui permet de renforcer l'activation des informations pertinentes et de supprimer les informations non pertinentes en fonction des buts actuels, ainsi que le pulvinar, qui permet de filtrer l'information distractrice en aidant à la synchronisation de l'activité des régions impliquées dans la sélection de cibles attentionnelles (Banich and Compton, 2023).

On peut également citer le lobe pariétal qui remplit deux rôles principaux : l'allocation des ressources attentionnelles, et la sélection des informations pertinentes grâce à l'intervention de trois régions distinctes (Banich and Compton, 2023). Le lobe pariétal supérieur permet de diriger l'attention de façon *top-down* sur les objets d'intérêt (Shomstein and Yantis, 2006). Le lobe pariétal inférieur, à l'inverse, est le siège de l'attention *bottom-up*, notamment via la jonction temporo-pariétale (Corbetta et al., 2008). Le sillon intra-pariétal, quant à lui, joue un rôle d'intégration entre les informations *bottom-up* et *top-down* par la création de cartes de saillance ('*salience maps*') visant à prioriser les informations les plus importantes à chaque instant (Ptak, 2012).

Finalement, le cortex préfrontal est le siège des décisions conscientes et permet de définir les objectifs qui vont guider l'attention *top-down* (Banich and Compton, 2023).

1.2.2 Modèles neuronaux

Au vu du grand nombre de régions cérébrales impliquées, différents auteurs ont proposé des modèles théoriques visant à scinder les processus attentionnels en sous-systèmes plus spécifiques, sous-tendus par des régions cérébrales plus localisées.

Notamment, Posner et Petersen (1990) décrivent l'attention comme un réseau neuronal à trois composantes fonctionnelles : l'alerte (ou vigilance), l'orientation et le contrôle exécutif. Bien que ces fonctions travaillent en collaboration dans la plupart des situations, les auteurs décrivent ces systèmes comme indépendants, et sous-tendus par des régions cérébrales distinctes (Petersen and Posner, 2012,

voir 1). Ce modèle souligne le rôle des interactions dynamiques entre différentes régions cérébrales dans l'attention, mettant en évidence que le contrôle attentionnel est distribué dans un réseau complexe et non pas localisé dans une zone particulière. Les recherches des régions cérébrales impliquées dans les différentes composantes du modèle se sont principalement focalisées sur l'attention visuelle.

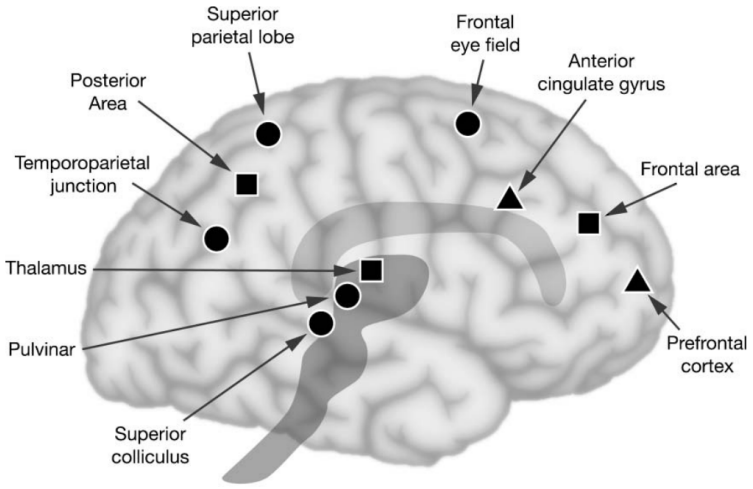


Figure 1: Modèle de Posner et Petersen du contrôle attentionnel. Les zones représentées par des carrés sous-tendent l'alerte, les zones représentées par des cercles représentent la fonction d'orientation, et le contrôle exécutif de l'attention est sous-tendu par les zones représentées par des triangles. Issu de Posner et Rothbart, 2007.

L'alerte est décrite comme la capacité d'accroître sa vigilance face à des stimuli imminents. Elle peut être de deux types: intrinsèque (ou tonique) lorsqu'elle concerne le contrôle cognitif des états de veille et d'éveil, ou phasique lorsqu'elle correspond à la capacité d'accroître l'état de préparation en réponse à un stimulus externe. La vigilance est associée aux régions cérébrales thalamiques, frontales et pariétales, et est latéralisée principalement dans l'hémisphère droit pour les processus intrinsèques, plus lents, et dans l'hémisphère gauche pour les processus phasiques (Posner and Petersen, 1990). De plus, cette fonction semble influencée par le système cérébral de norépinéphrine issu du locus coeruleus moyen. En effet, face à un signal d'alerte, on observe un pic d'activité dans cette zone, et donc une augmentation de la norépinéphrine (Petersen and Posner, 2012).

La fonction d'orientation permet la sélection d'informations spécifiques parmi un ensemble de stimuli sensoriels. Elle peut être automatique lorsqu'un stimulus saillant de l'environnement attire l'attention, ou volontaire lorsqu'on cherche notre environnement pour un stimulus particulier. Lorsque l'orientation implique un mouvement de la tête ou des yeux, on parle d'orientation ouverte, et lorsque ce n'est pas le cas, on parle d'orientation couverte.

Cette fonction d'orientation attentionnelle semble modulée par les systèmes cholinergiques cérébraux (Posner and Pertersen, 1990), impliqués dans les fonctions de localisation spatiale via lobe pariétal supérieur (Petersen and Posner, 2012). Notamment, les études utilisant un paradigme de détection indicée ont mis en évidence, tant chez l'humain que chez le primate non-humain, un impact des drogues influençant ce neurotransmetteur sur l'orientation de l'attention (Petersen and Posner, 2012).

La troisième composante du modèle de Posner et Petersen (1990) correspond au contrôle exécutif de l'attention et est principalement utilisé dans les situations nécessitant de la planification, une prise de décision, ou de la détection d'erreurs. Face aux capacités limitées du système attentionnel, notre cerveau doit prioriser et se focaliser sur les informations importantes, et ainsi effectuer un contrôle. Deux théories opposées tentent d'expliquer le contrôle exécutif de l'attention (Petersen and Posner, 2012). Dans la théorie cognitive, le cortex cingulaire antérieur serait responsable du traitement des conflits, en relation avec les aires frontales latérales. La seconde théorie postule l'existence de deux réseaux cérébraux distincts. D'un côté, le réseau fronto-pariétal serait impliqué lors des changements entre tâches et dans l'ajustement lors d'une tâche. De l'autre, le réseau cingulo-operculaire semblerait jouer un rôle de maintien du contrôle attentionnel. Bien que ces deux théories expliquent chacune de nombreuses données empiriques, Petersen and Posner (2012) mettent en évidence que la seconde semble plus pertinente, en regard notamment des études lésionnelles chez les humains et les animaux.

Un autre modèle, proposé par Corbetta et Schulman en 2002 (Figure 2), postule que le système d'orientation visuel est divisé en deux réseaux neuronaux distincts. Le réseau dorsal, ou *goal-directed attention*, reprend les régions pariétales supérieures et postérieures (dont le sillon intra-pariétal), ainsi que des régions frontales comme le *fontal eye field* (FEF). Il contrôle l'attention *top-down* qui permet de choisir l'objet de notre focus attentionnel en fonction de la tâche et des buts actuels. Le second réseau, dit ventral (*stimulus-driven attention*), comprend la jonction temporo-pariétale (TPJ) et les régions frontales ventrales et fronto-orbitaires. Il correspond à l'attention *bottom-up* qui permet de détecter les nouveaux stimuli saillants dans l'environnement, et est latéralisé principalement dans l'hémisphère droit (Petersen and Posner, 2012). Ces deux réseaux antagonistes travaillent en collaboration pour établir des *salience maps*. Le réseau dorsal représente l'information importante à un moment donné sous forme de carte des éléments temporairement importants et prioritaires, et le réseau ventral permet de mettre à jour ces cartes.

Au niveau auditif, peu de recherches ont été effectuées et les sous-bassement cérébraux sont encore

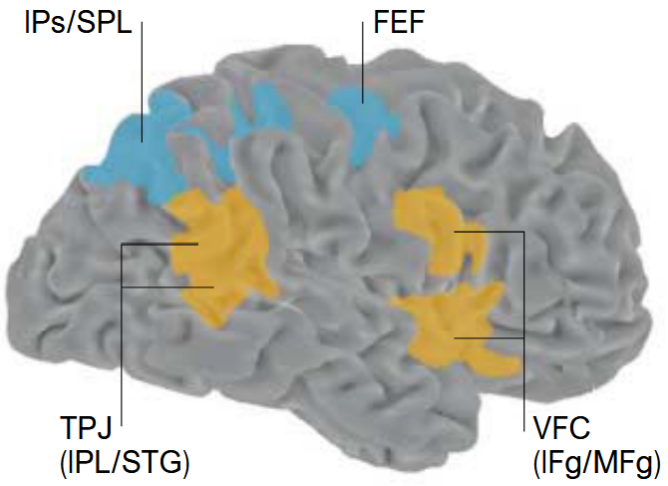


Figure 2: Modèle neuroanatomique du contrôle attentionnel de Corbetta et Schulman. Les aires en bleu indiquent le réseau fronto-pariétal dorsal, et les aires en orange indiquent le réseau fronto-pariétal ventral. IPs/SPL = sillon intra-pariétal/lobule pariétal supérieur; FEF = Frontal Eye Field; TPJ = jonction temporo-pariétale (IPL/STG = lobule pariétal inférieur/gyrus temporal supérieur); VVFC = cortex frontal ventral; (IFg/MFg = gyrus frontal inférieur/gyrus frontal médian). Issu de Corbetta and Shulman, 2002

mal compris. Cependant, des études mettent en évidence que le système attentionnel peut également être subdivisé en deux réseaux neuronaux distincts. Le réseau ventral "*what*" est composé du gyrus antérolatéral de Heschl, du gyrus temporal antérieur supérieur et du planum temporale postérieur, et s'occupe de l'identification des objets auditifs. Au sein de ce premier réseau, on peut identifier un sous-réseau (plan supratemporal et lobule pariétal inférieur) qui s'occupe de moduler l'attention sélective de façon dynamique. Le réseau dorsal, "*where*", en charge de la localisation de ces objets, comprend le planum temporale et le gyrus temporal supérieur postérieur (Häkkinen and Rinne, 2018; Higgins et al., 2017). Certains auteurs (Fu et al., 2020) suggèrent d'approfondir l'existence d'un réseau "*when*" responsable de la perception temporelle, comme décrit par Lu et al.(2017). Cet ajout s'avère pertinent au vu de l'importance de la cohésion temporelle dans la liaison et la séparation des différentes caractéristiques auditives d'un discours.

Des recherches en imagerie fonctionnelle mettent également en évidence un recouvrement entre les régions cérébrales associées à la fonction d'orientation visuelle de l'attention, et les régions impliquées dans l'orientation attentionnelle dans d'autres modalités sensorielles (Petersen and Posner, 2012). De plus, on observe une collaboration entre les différentes modalités. L'orientation de l'attention envers un certain endroit de l'espace augmente la priorité accordée aux stimuli apparaissant à cet endroit pour la

modalité concernée, ainsi que pour les autres.

1.2.3 Électrophysiologie

Outre le fait que les régions cérébrales impliquées dans la sélection attentionnelle sont distribuées dans le cerveau, elles sont également sollicitées à des moments distincts dans le temps (Banich and Compton, 2023). En effet, si la composante précoce ou tardive de la sélection attentionnelle a longtemps fait débat, les recherches en potentiels évoqués (*event-related potentials*, ERP) ont mis en évidence que le filtrage de l'information peut en réalité se faire de façon précoce ou tardive selon les besoins de la situation (Eimer, 2014).

En particulier, les connaissances actuelles sur l'activité électrochimique cérébrale ont permis de mettre en évidence une détection précoce des stimuli auditifs très saillants. La *mismatch negativity* (MMN) est une réponse électrophysiologique prenant place dans le cortex auditif primaire environ 100-150ms après l'apparition de stimuli déviants, ou inattendus (Näätänen et al., 1978). Il indique un traitement perceptuel qui permet la détection des changements dans l'environnement perceptif, comme l'apparition d'un stimulus inhabituel (Morrison et al., 2019). Cependant, la P3a (se produisant entre 250-300ms) est considérée comme le marqueur de la capture involontaire de l'attention par les stimuli saillants de l'environnement (Parmentier, 2014).

1.2.4 Oscillations neuronales

Les oscillations neuronales jouent un rôle crucial dans l'attention sélective auditive en modulant l'excitabilité neuronale et en synchronisant l'activité neuronale avec les stimuli pertinents (Schroeder and Lakatos, 2009). Les oscillations neuronales sont des modèles rythmiques d'activité électrique, observables dans différentes bandes de fréquences telles que thêta (4-8 Hz), alpha (8-12 Hz), bêta (12-30 Hz) et gamma (30-100 Hz). Elles facilitent la prise en compte des signaux auditifs pertinents, tout en supprimant le traitement des non pertinents. Les recherches sur le sujet ont montré que la phase de ces oscillations peut s'aligner sur le rythme des stimuli perçus, optimisant ainsi le traitement sensoriel et améliorant les résultats perceptifs (Lakatos et al., 2007).

Par exemple, il a été constaté que les oscillations de la bande delta (1-4 Hz) s'alignent sur la structure

temporelle des entrées auditives, ce qui améliore la réponse neuronale aux sons perçus (Schroeder and Lakatos, 2009). En outre, des études utilisant l'électroencéphalographie (EEG) et la magnétoencéphalographie (MEG) ont démontré que les oscillations thêta et gamma sont particulièrement impliquées dans le suivi et la compréhension de la parole, ce qui indique leur importance dans l'attention sélective auditive (Zion Golumbic et al., 2012). Ces résultats soulignent le rôle dynamique des oscillations neuronales dans la facilitation de l'attention sélective en coordonnant la synchronisation et la force des réponses neuronales aux stimuli auditifs.

2 La modélisation de l'attention en psychologie cognitive

2.1 La modélisation : qu'est-ce que c'est, et à quoi ça sert ?

Au vu de la complexité des processus étudiés en psychologie cognitive, le recours à la modélisation est essentiel pour comprendre les relations entre les différents mécanismes en jeu. Mais qu'entend-t-on exactement par "modèle" ? Si l'on consulte le Larousse, on se rend rapidement compte de la polysémie du terme : j'ai compté 15 définitions différentes du mot selon le contexte d'utilisation (Larousse, 2025). En sciences, cependant, on définit le mot "modèle" en suivant la tradition de la logique formelle : '*A model is a structure that makes all sentences of a theory true when its symbols are interpreted as referring to objects, relations, or functions of a structure.*' (Frigg and Hartmann, 2025). Autrement dit, un modèle est une représentation formelle d'un phénomène, c'est-à-dire exprimée dans un langage structuré, avec des règles précises dans le but de le décrire, l'expliquer, ou le prédire. En psychologie expérimentale, on peut citer plusieurs types de modèles, chacun offrant un cadre spécifique pour représenter et simuler les processus cognitifs étudiés. Après avoir exposé les différents types de modèles, nous passerons en revue les principaux modèles élaborés dans le cadre de la recherche sur l'attention auditive.

Les modèles boîtes-flèches sont parmi les plus répandus. Ils consistent en une représentation schématique des processus cognitifs sous forme de composantes interconnectées. Un exemple emblématique est le modèle de la mémoire de Baddeley (2000), qui la décrit comme un système modulaire avec différentes sous-composantes dédiées au stockage et à la manipulation de l'information. On peut également citer le modèle de Atkinson and Shiffrin (1968), ou encore celui de Patterson and Shewell (1987) sur le système langagier. Ces modèles permettent une vision synthétique des relations fonctionnelles entre les

différentes composantes cognitives, mais sont généralement purement descriptifs et dépourvus de formalisation précise. Ils ne permettent pas de modéliser le fonctionnement effectif du phénomène étudié, en particulier les interactions dynamiques et contextuelles des éléments qui le caractérisent (Spivey, 2023).

Les modèles mathématiques, quant à eux, reposent sur l'utilisation de formules et équations pour décrire un phénomène cognitif. Ils sont particulièrement utilisés dans l'étude de la prise de décision, du traitement sensoriel ou encore des mécanismes d'apprentissage. Un exemple classique est le modèle de diffusion utilisé pour modéliser les processus décisionnels en fonction du temps et du niveau de certitude (Ratcliff, 1978), ou encore le modèle de la mémoire à court terme de Schweickert and Boruff (1986) qui permet de prédire le rappel d'un item en fonction de sa position dans la séquence et de la vitesse de prononciation. L'avantage principal de ces modèles est qu'ils permettent une quantification précise des relations entre les composantes d'un phénomène. Cependant, ils sont souvent très simplifiés, ce qui peut amener à négliger certains aspects pourtant fondamentaux du phénomène, et empêcher la généralisation de leurs prédictions (McClelland, 2009).

Les modèles connexionnistes s'appuient sur des algorithmes d'apprentissage inspirés du fonctionnement cérébral, notamment via l'intelligence artificielle (IA). Ces modèles, largement influencés par les réseaux de neurones artificiels (*deep learning*), permettent de simuler des phénomènes cognitifs sans nécessairement suivre la structure biologique du cerveau. Par exemple, ils ont été utilisés pour modéliser le langage ou la reconnaissance visuelle, en observant l'émergence de ces capacités à partir de données brutes. Basés sur des règles d'apprentissage simples, ils permettent d'observer le développement auto-organisé (sans programmation explicite) de compétences cognitives complexes. Les différents types de modèles connexionnistes seront décrits en détail plus loin dans ce mémoire, mais leur mécanisme global est présenté dans l'Encadré 1 (page 24).

Enfin, les modèles neuronaux se distinguent des modèles connexionnistes par leur ancrage biologique. Alors que les réseaux de neurones artificiels sont souvent conçus pour optimiser des tâches spécifiques, les modèles neuronaux visent à reproduire fidèlement le fonctionnement des neurones biologiques, y compris leurs dynamiques électrochimiques. Ces modèles postulent une égalité entre le fonctionnement cognitif et le fonctionnement neuronal. Par exemple, les *spiking neural networks* (e.g. Izhikevich, 2003) sont largement utilisés dans la littérature, notamment dans les domaines comme la perception ou l'apprentissage. Ces approches sont particulièrement utiles en neurosciences computationnelles, où elles aident à mieux comprendre les bases biologiques de la cognition.

2.2 Les modèles boîtes flèches

2.2.1 Modèles classiques de l'attention

L'un des premiers modèles de l'attention a été celui de Donald Broadbent en 1958 (Figure 3). Il postule que notre système attentionnel utilise un filtre de sélection lorsqu'il se trouve face à deux stimuli en compétition. Ce modèle est souvent référé comme un modèle d'attention sélective précoce (ou *bottleneck model*) car le filtre éliminant l'information superflue agit dès l'entrée dans le système du flux d'informations (Goldstein, 2015).

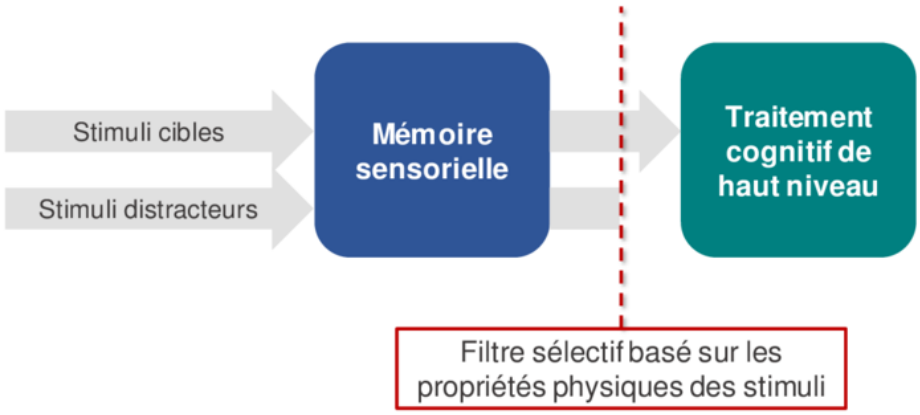


Figure 3: Modèle de Broadbent. Source : Techer, 2016. Les stimuli sensoriels sont d'abord stockés brièvement dans la mémoire sensorielle. Un filtre, basé uniquement sur les propriétés physiques des stimuli, laisse passer les informations pertinentes vers les étapes de traitement cognitif de haut niveau, tandis que les autres sont ignorées.

Les différents stimuli entrent tout d'abord dans la mémoire sensorielle qui permet d'analyser les caractéristiques acoustiques du son. Une fois que les différentes caractéristiques physiques des sons ont été extraites, le filtre identifie le message pertinent et l'envoie vers les processus de plus haut niveau de traitement de l'information. Les messages non pertinents sont simplement bloqués. Malheureusement, ce modèle ne permet pas de rendre compte du *cocktail party effect*. En effet, le modèle de Broadbent suggère que le filtre empêche une partie de l'information d'être traitée au niveau sémantique, or le *cocktail party effect* semble montrer qu'une information au départ non retenue par le filtre peut tout de même être captée et traitée par la suite.

A l'inverse, d'autres auteurs comme Deutsch et Deutsch (1963) ou Norman (1968) ont proposé des

modèles de sélection tardive. Ces modèles postulent que l'évaluation de la pertinence des stimuli et la sélection se font au moment où on analyse le contenu sémantique de l'information. Le processus de sélection se fait dans la mémoire de travail (Sieroff, 1992), après l'identification de l'information et du traitement perceptuel.

Anne Treisman (1964) a développé un modèle d'attention sélective par atténuation (Figure 4), dans lequel la sélection se fait en deux étapes, et non pas avec un filtre opaque, mais avec un atténuateur. Face à deux stimuli en compétition, l'atténuateur va les analyser sur base de leurs caractéristiques physiques (comme dans le modèle de Broadbent) afin de déterminer lequel est le message pertinent. La différence est qu'ici, les deux stimuli vont passer à l'étape suivante. Le message cible ne sera pas altéré, tandis que le message non pertinent va être atténué, et n'atteindra pas l'étape suivante avec autant de force. On parle parfois de "*leaky filter model*" car une partie du message non pertinent traverse l'atténuateur (Goldstein, 2015).

Une fois les messages passés dans l'atténuateur, ils sont envoyé dans le "*dictionnary unit*" qui va analyser leur contenu. Cette unité garde en mémoire des mots, chacun associé à un seuil d'activation. Les mots du message cible, étant ressortis de l'atténuateur avec toute leur "force", vont d'office être au dessus de leur seuil d'activation, et donc être captés. À l'inverse, les mots du message non pertinent ayant été atténués, la plupart ne seront pas suffisamment forts pour arriver au seuil d'activation. Seuls les mots fréquents et/ou signifiants (comme notre propre nom), présentant un seuil d'activation plus faible, seront captés (Goldstein, 2015). Ce modèle permet d'expliquer le *cocktail party effect*.

Navon, en 1989, a également tenté de dépasser cette dichotomie précoce/tardif. Il est parti du principe que la sélection attentionnelle, et plus généralement le traitement de l'information, ne serait pas une succession d'étapes de plus en plus spécialisées, mais plutôt un ensemble de processus avançant en parallèle. Tous les processus d'analyse commenceraient en même temps. Les analyses sur les traits plus perceptifs seraient plus rapides, tandis que les analyses sur les traits plus complexes (sémantiques) prendraient plus de temps. En fonction du critère de sélection (caractéristique physique ou sémantique de l'information), la sélection se ferait alors de façon plus ou moins rapide. On parle de modèle hiérarchisé, car il voit l'attention comme un réseau qui interviendrait à des niveaux plus précoces ou tardifs du traitement de l'information en fonction des caractéristiques d'une situation donnée (Sieroff, 1992).

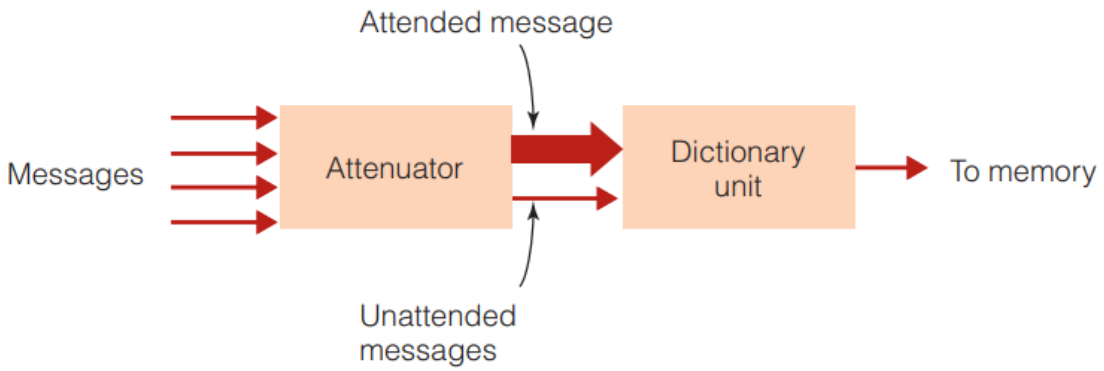


Figure 4: Modèle de Treisman. Source : Goldstein, 2015. Tous les messages rentrent dans l'atténuateur qui diminue (sans supprimer) l'intensité des messages non pertinents. Ces derniers peuvent encore être traités s'ils contiennent des éléments saillants dont le seuil d'activation est plus faible. Le traitement se poursuit ensuite dans une unité dictionnaire qui analyse le contenu sémantique et envoie les informations pertinentes vers la mémoire.

2.2.2 Load theory

Nilli Lavie, en 1995, a énoncé la théorie de la charge, ou *load theory*. Le postulat de base est que nous possédons des capacités limitées de traitement perceptuel, mais que le cerveau traite de façon automatique et obligatoire tous les stimuli perceptifs auxquels il peut accéder. On pourrait réaliser un parallèle avec une jauge : le cerveau possède une certaine capacité de traitement perceptuel, qu'il remplit de façon automatique et systématique avec le plus d'informations possible afin d'atteindre la capacité maximale. Face à une tâche avec une charge perceptuelle élevée, la totalité des capacités de traitement de notre cerveau sera allouée aux éléments de la tâche, et nous ne percevrons alors aucun stimuli extérieur. La jauge sera complète avec les stimuli de la tâche uniquement. À l'inverse, dans une tâche avec une charge perceptuelle faible, seule une partie de nos capacités de traitement sera allouée à la tâche en cours. Ainsi, le cerveau cherchera à remplir la jauge en traitant également les stimuli non-pertinents pour la tâche. Réaliser efficacement une tâche d'attention sélective et concentrée nécessite donc le maintien actif d'une priorité sur certains stimuli (contrôle *top-down*), mais également un niveau élevé de charge perceptuelle qui sollicitera toutes les capacités disponibles (Lavie et al., 2014). Cependant, l'applicabilité de cette théorie, initialement développée pour la modalité visuelle, au traitement auditif fait l'objet de débats, et la littérature sur le sujet conclut à des résultats mitigés (Murphy et al., 2017).

2.2.3 Free-Energy Model

Le modèle attentionnel de l'énergie libre (Feldman and Friston, 2010) s'inscrit dans le cadre théorique plus large du principe d'énergie libre (*Free Energy Principle*) développé par Karl Friston. Selon cette théorie, le cerveau est constamment en train de réaliser des prédictions sur le monde extérieur et compare ces prédictions avec les entrées sensorielles qu'il recoit. La différence entre les deux constitue l'énergie libre, ou l'entropie cérébrale. L'objectif est de la minimiser. Pour ce faire, on dispose de deux solutions : soit changer nos perceptions, nos représentations internes, soit agir pour modifier les stimuli sensoriels reçus afin qu'ils correspondent aux prédictions (Feldman and Friston, 2010).

Dans ce contexte, l'attention est vue comme un processus qui permet de prioriser les ressources cognitives afin de réduire efficacement l'énergie libre, en ajustant les prédictions en fonction des entrées sensorielles (Fu et al., 2020). Elle peut agir de trois façons (Feldman and Friston, 2010): 1. Augmenter le poids des informations sensorielles les plus pertinentes afin de mettre à jour nos représentations, 2. Mettre en évidence les erreurs de prédiction dans les voies sensorielles pertinentes pour permettre un apprentissage plus efficace, ou 3. Se diriger vers les stimuli pour lesquels des erreurs de prédictions entraîneraient la plus grande augmentation d'énergie libre, permettant ainsi une allocation plus optimale des ressources limitées. Malheureusement, ce modèle se concentre principalement sur l'attention *top-down*, et ne permet pas d'expliquer les phénomènes *bottom-up* tels que le *cocktail-party effect*.

2.3 Les modèles computationnels

2.3.1 Les types de modèles computationnels

Dans le cadre des approches connexionnistes, le choix de l'architecture du réseau neuronal a une influence déterminante sur la manière dont les données sont traitées, apprises, et généralisées. Nous allons aborder les principales architectures utilisées dans la modélisation en psychologie cognitive : les réseaux convolutionnels (CNN), les réseaux neuronaux récurrents (RNN), les réseaux de type *long-short term memory* (LSTM) ainsi que les transformeurs avec mécanismes d'attention.

Les réseaux convolutionnels (Lecun et al., 1998, Figure 5) sont des variants des réseaux neuronaux classiques. Ces réseaux sont souvent utilisés pour le traitement des images ou de spectrogrammes (pour

plus d'informations et une définition du spectrogramme, voir Annexe A), car ils y sont particulièrement performants. Par exemple, ils sont souvent utilisés dans la classification d'images (ex: chat vs. chien). Ces réseaux sont composés d'une alternance de couches convolutionnelles et de couches dites de "*pooling*". Les couches convolutionnelles ont pour objectif de réduire la taille de l'entrée en identifiant uniquement les caractéristiques importantes pour identifier l'image (e.g. les caractéristiques spécifiques aux chiens, et les caractéristiques spécifiques aux chats). Ces couches fonctionnent avec des filtres qui traitent, de façon sérielle, de petites portions de l'entrée. Ils encodent ensuite les données réduites dans ce qu'on appelle une "carte d'activation", ou *feature map*. Les couches de *pooling* diminuent ensuite la taille des cartes d'activation via la même procédure que précédemment. Leur but est de réduire la dimensionalité des données. Le résultat est l'obtention d'une *pooled feature map*, qu'on peut ensuite aplatir en un vecteur, qui passera à son tour dans un réseau de neurones classique (Mehrish et al., 2023; Mwiti, 2022).

Les CNN exploitent également le principe de la connectivité clairsemée (*sparse connectivity*) et du partage des poids, ce qui réduit le nombre de paramètres à entraîner et améliore la généralisation. Selon le type de données, ces réseaux peuvent être adaptés en une, deux ou trois dimensions : les CNN 1D sont souvent utilisés pour les signaux (par exemple, signaux EEG ou vocaux), les CNN 2D pour les images, et les CNN 3D pour les vidéos ou les volumes médicaux. En plus de la classification d'images, ils sont également employés dans des tâches complexes comme la détection d'objets, la segmentation d'image ou encore l'analyse de scènes en vision par ordinateur.

Les réseaux neuronaux récurrents (RNN, Elman, 1990, voir Figure 6) permettent de traiter des informations de façon séquentielle afin de prédire des valeurs dans une séquence. Ils sont particulièrement utilisés dans les domaines où les données ont une dimension temporelle ou ordonnée, comme la modélisation du langage naturel. Ils contiennent des boucles de rétroaction permettant de garder en mémoire les apprentissages précédents et d'analyser de nouveaux stimuli en tenant compte d'informations précédemment vues. Un problème bien connu des RNN classiques est qu'ils ont des difficultés à capturer les dépendances à long terme, c'est-à-dire créer des liens entre un stimulus vu au début de la séquence et un stimulus plus récent. À cause de leur structure, les boucles de rétroaction des RNN leur permettent de tenir compte du contexte proche (c'est-à-dire de l'itération précédente), tandis que le contexte plus lointain (les premières itérations de la séquence) est peu à peu dilué et l'information est perdue. On appelle cela le problème du '*vanishing or exploding gradient*' (Bengio et al., 1994). Ce phénomène survient lors de la rétropropagation du gradient à travers les nombreuses étapes temporelles du réseau,

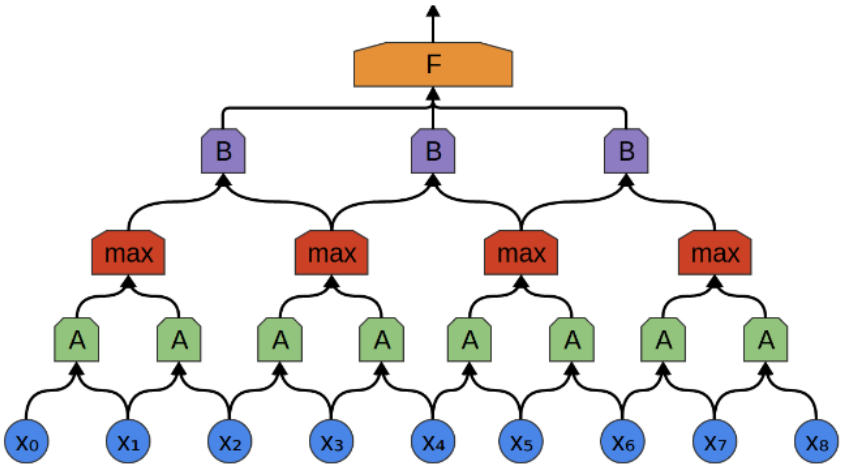


Figure 5: Architecture d'un réseau de neurones convolutionnel en une dimension. Des groupes d'entrées adjaçentes (e.g. X_0 et X_1) passent par un filtre de convolution (A) visant à extraires les caractéristiques pertinentes. La couche 'max' (*max pooling*) permet de réduire la dimensionnalité des données. Finalement, les couches supérieures (B et F) permettent l'agrégation des données afin de produire une sortie finale. Par exemple, dans le traitement du langage, les entrées X_0 à X_8 pourraient représenter les mots d'une phrase, et la sortie en serait la classification de sentiment (phrase positive vs. négative). Issu de "Conv Nets: A Modular Perspective - colah's blog", n.d.

ce qui peut conduire à des gradients extrêmement petits (*vanishing*) ou très grands (*exploding*), rendant l'entraînement instable ou inefficace.

Les modèles *long-short term memory* (Hochreiter and Schmidhuber, 1997, Figure 7) visent à répondre à ce problème grâce à l'intégration d'une "mémoire à long terme", moins sensible à la mise à jour effectuée après chaque nouvelle présentation de stimuli. Cette structure singulière leur permet d'analyser chaque nouvel élément au moyen d'informations stockées en mémoire à court terme (sur base du stimulus vu juste avant) et à long terme (prenant en compte les stimuli vu dès le début de la séquence). Le contexte gardé en mémoire est ainsi plus riche. Les LSTM se composent de chaînes de neurones artificiels composés de trois éléments. Tout d'abord, l'*input gate* (porte d'entrée) permet de déterminer quels nouveaux éléments (sur base du stimulus en cours de traitement) vont être mémorisés. Ensuite, la *forget gate* (porte d'oubli) permet de déterminer la quantité d'information stockée en mémoire à long terme qui doit être conservée. Enfin, l'*output gate* (porte de sortie) décide quelle information sera transmise à la cellule suivante. Ces trois composants travaillent conjointement pour permettre au réseau de décider dynamiquement quelles informations conserver, oublier ou transmettre à l'étape suivante. Cette flexibilité permet au modèle de mieux préserver les dépendances temporelles longues, en particulier dans des tâches comme la génération

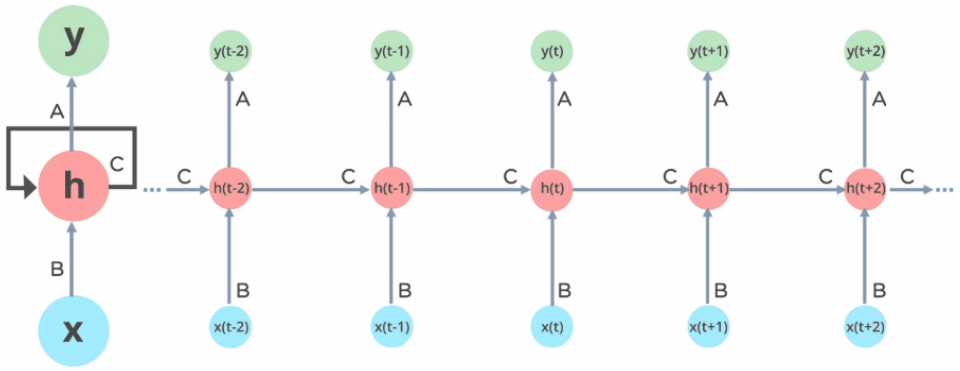


Figure 6: Schéma d'un réseau neuronal récurrent (RNN) illustrant le traitement séquentiel d'une entrée temporelle, avec boucle de rétroaction entre les étapes. A chaque étape, le RNN prend une entrée x_t, génère une sortie y_t, et met à jour son état caché h_t, qui est réutilisé à l'étape suivante. Cette boucle permet au réseau de conserver une forme de mémoire des étapes précédentes. Issu de Kalita, 2024.

de texte, la traduction automatique ou la reconnaissance vocale.

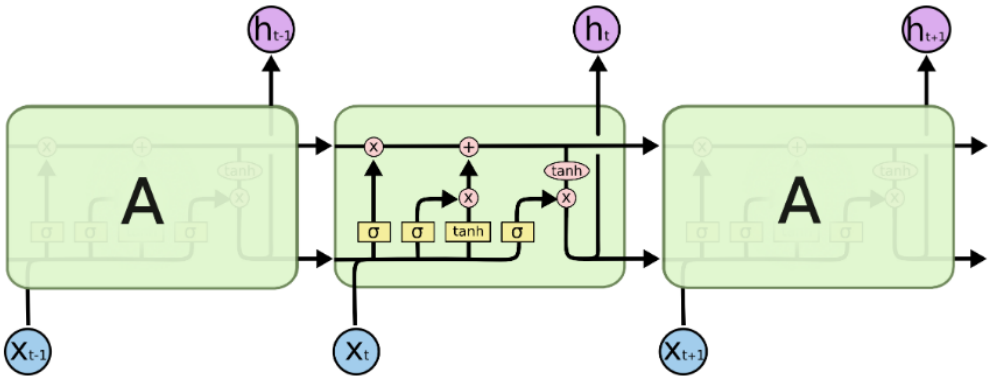


Figure 7: Schéma d'une architecture LSTM en trois blocs. Chaque bloc est composé de trois cellules comprenant trois portes de contrôle (*input*, *forget* et *output*) qui régulent le flux d'information à travers la mémoire. La ligne horizontale supérieure représente la mémoire à long terme, transmise d'une cellule à l'autre, tandis que la sortie h_t correspond à la mémoire à court terme. Les entrées X_t alimentent chaque cellule à chaque pas de temps. Issu de "Understanding LSTM Networks – colah's blog", n.d.

Une autre alternative aux RNN sont des modèles de type transformeurs (Vaswani et al., 2017) qui visent également à répondre au problème du *vanishing or exploding gradient* en incorporant un traitement non plus séquentiel, mais en parallèle, c'est-à-dire que le réseau dispose, à chaque instant, des informations de tous les éléments de la séquence. Une autre grande modification est l'ajout d'un mécanisme d'attention. Il s'agit d'un processus permettant d'établir des liens entre les différents éléments d'une séquence, comme par exemple entre les différents mots d'une phrase, en se concentrant sur les éléments les plus pertinents. Leur architecture repose généralement sur un empilement de couches

comprenant chacune une phase d'attention suivie d'un réseau de neurones entièrement connecté (voir Figure 8). Cette approche permet une modélisation plus efficace du contexte global d'une séquence, notamment dans des tâches complexes comme la traduction automatique, la réponse à des questions ou le résumé automatique de textes.

Prenons une phrase de quatre mots. Chaque mot est transformé en un vecteur sémantique, c'est-à-dire une représentation numérique de son sens. Ces vecteurs sont regroupés dans une matrice, à laquelle on ajoute ensuite des informations de position (appelées *positional encodings*), afin que le modèle puisse tenir compte de l'ordre des mots. Cette matrice est ensuite dupliquée en trois versions appelées Q (*Query*), K (*Key*) et V (*Value*). Le cœur du mécanisme d'attention repose alors sur le calcul suivant :

$$softmax(\frac{Q * K'}{\sqrt{d}}) * V$$

où d est la dimension des vecteurs. Ce calcul permet au modèle de pondérer l'importance de chaque mot en fonction de sa relation avec les autres, en combinant les informations des vecteurs de manière dynamique. On obtient ainsi une nouvelle représentation de chaque mot, enrichie de son contexte dans la phrase.

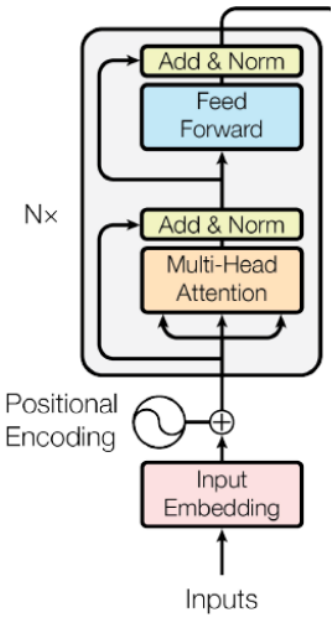


Figure 8: Bloc d'encodage d'un Transformer. Chaque mot est transformé en vecteur, auquel on ajoute des informations de position pour tenir compte de l'ordre de présentation. Ces représentations passent ensuite par plusieurs blocs identiques, composés d'un mécanisme d'attention (pour pondérer l'importance des mots entre eux, '*Multi-Head Attention*') et d'un réseau de neurones (pour affiner leur représentation, '*Feed Forward*'). Issu de Singh, 2025.

En résumé, les réseaux convolutionnels sont particulièrement adaptés aux données structurées spatialement, comme les images, tandis que les RNN et les LSTM sont conçus pour le traitement de données séquentielles. Les transformeurs, grâce à leur traitement parallèle de l'information et à leur mécanisme d'attention, surmontent plusieurs limitations structurelles des architectures récurrentes et s'imposent aujourd'hui comme une référence dans de nombreux domaines, notamment le traitement automatique du langage. Toutefois, bien que les transformeurs et les CNN soient très performants d'un point de vue computationnel, leur fonctionnement parallèle s'éloigne du mode de traitement humain, qui repose sur une perception séquentielle des stimuli. À l'inverse, les LSTM, en tenant compte des dépendances temporelles dans l'ordre de présentation, reproduisent plus fidèlement le traitement cognitif de l'information, notamment dans les tâches impliquant mémoire et raisonnement temporel.

Intelligence artificielle et apprentissage automatique

L'**intelligence artificielle** (IA) est une discipline qui vise à faire réaliser, par des machines, des tâches qui sont supposées nécessiter de l'intelligence humaine (Defays, 2023). Ces tâches peuvent concerner le traitement du langage, la résolution de problème, ou encore la reconnaissance de motifs (*patterns*). Il existe différentes sous-disciplines de l'IA. Notamment, le *machine learning* vise à créer des algorithmes capables d'apprendre directement à partir des données, et d'améliorer leurs performances via l'entraînement, sans qu'on ait besoin de les programmer de façon explicite. L'apprentissage est, ici, considéré comme la capacité de l'algorithme à transformer les données, de sorte à former des représentations facilitant leur traitement (Chollet, 2021).

Le ***deep learning*** est une sous-branche du *machine learning* qui utilise comme structure principale des "réseaux de neurones", c'est-à-dire des couches successives empilées les unes sur les autres, fournissant chacune des représentations de plus en plus significatives. Cette superposition de couches, appelé réseau de neurones artificiels (RNA), va progressivement simplifier ou "purifier" les données afin de créer des représentations se prêtant mieux à la tâche à réaliser (Chollet, 2021).

L'**apprentissage** repose sur un processus simple. L'algorithme va comparer ses résultats face à une tâche particulière aux résultats attendus. Par exemple, si on lui a demandé de labelliser des photos comme représentant un chien ou un chat, il va comparer ses classifications à celles escomptées. Pour ce faire, il aura besoin de trois éléments : les données d'entrée (les images), les sorties attendues (les labellisations correctes, chat ou chien) ainsi qu'une mesure d'écart avec la prédiction du modèle. Cette mesure est réalisée par la fonction de perte ('*loss function*') qui consiste à calculer une distance entre la prédiction du modèle et la sortie attendue, permettant d'indiquer au réseau s'il a atteint de bons résultats ou pas. Comme déjà expliqué, les RNA sont constitués de couches successives qui, chacune, transforment les données afin de faciliter le traitement. À chaque couche, la transformation effectuée est stockée dans les poids ('*weights*') ou paramètres. Dans ce contexte, la fonction de perte, et donc l'écart entre la prédiction et la sortie attendue, va servir de *feedback* pour ajuster les poids, et par conséquent la transformation effectuée, afin de rapprocher la sortie du modèle de la sortie attendue. L'ajustement des poids peut se faire de différentes façons, mais la plus courante en *deep learning* est de faire appel à un algorithme de *backpropagation*. (Rumelhart et al., 1986).

2.3.2 Dans le contexte de l'attention sélective auditive

Les phénomènes d'attention sélective auditive ont été peu modélisés dans la littérature. Or, très tôt (1953), Cherry apportait déjà la réflexion sur la potentielle modélisation de ce processus : *'On what logical basis could one design a machine ("filter") for carrying out such an operation [dichotic listening in cocktail party environment] ?'*. La capacité à reconnaître ce qu'une personne dit alors que d'autres parlent en même temps est un phénomène complexe, et la modélisation informatique permet d'apporter des éléments de réponse sur les organisations fonctionnelles permettant l'émergence d'un tel phénomène.

La plupart des modèles existants ont été développés dans le cadre de la recherche en robotique, afin de créer des robots capables de converser avec des personnes. Par exemple, l'article de Geva et al. (2024) propose un modèle CNN combinant les domaines temporel et fréquentiel (spectrogrammes et audio bruts) pour la localisation binaurale des sources sonores. Bien que leur objectif n'est pas de modéliser l'attention sélective, leur modèle peut être intégré dans des systèmes de modélisation cognitive, afin de mieux comprendre comment les humains filtrent et se concentrent sur des sources sonores spécifiques dans des environnements complexes. De plus, la combinaison des domaines temporel et fréquentiel est essentielle, car l'évolution dans le temps constitue une caractéristique fondamentale des stimuli sonores (Kaya and Elhailali, 2016). Cependant, l'utilisation d'un CNN ne permet pas la modélisation adéquate d'une architecture cognitive humaine, étant donné que ceux-ci traitent l'ensemble d'une séquence en parallèle.

Cependant, malgré un focus sur la recherche en robotique et depuis l'avènement des techniques de *deep learning*, plusieurs modèles computationnels distincts ont été proposés en sciences cognitives afin de directement simuler l'attention sélective auditive.

Xu et al. (2018) ont proposé un modèle unifié de l'attention sélective auditive en prenant en compte les aspects attentionnels, mais également les aspects mnésiques qui permettent l'émergence du *cocktail party effect*. Pour ce faire, ils ont entraîné un modèle de type BiLSTM supplémenté d'un module d'attention et l'ont soumis à deux tâches distinctes. La première, *top-down*, consistait à se concentrer sur un seul interlocuteur dans une entrée composée de 3 interlocuteurs. Dans la deuxième, *bottom-up*, le modèle devait attirer son attention sur le discours d'un distracteur lorsqu'il prononcait un mot spécifique appris en amont. La phase d'entraînement permettait au modèle d'apprendre à identifier la cible et le discours spécifique, puis de l'intégrer dans une mémoire à long terme (via le BiLSTM). Grâce à cette architecture

innovante, leur modèle réussit les deux tâches distinctes et ses performances sont généralisables à des environnements réalistes comprenant du bruit de fond.

Cependant, malgré ces excellentes performances, leur modèle comprend plusieurs limites. Tout d'abord, l'utilisation d'un BiLSTM, bien qu'efficace, ne permet pas de modéliser exactement un fonctionnement humain. En effet, les BiLSTM sont des algorithmes qui, comme les CNN, ont accès à l'ensemble de la séquence à tout instant du traitement (Schuster and Paliwal, 1997). Leur architecture est composée de deux LSTM travaillant en sens inverse, l'un traitant la séquence du début vers la fin, et l'autre traitant la séquence de la fin vers le début. Or, cette organisation n'est pas réaliste dans le cadre de l'attention humaine. Ensuite, bien que les deux tâches proposées permettent d'évaluer les deux composantes de l'attention sélective auditive (*top-down* et *bottom-up*), elles ne permettent pas de montrer l'émergence du *cocktail party effect* (influence *bottom-up*) dans une tâche a priori *top-down*. Leur modèle n'a en effet pas été testé dans une tâche intégrant ces deux défis de façon simultanée. Finalement, les jeux de données utilisés par les auteurs comprennent des fichiers audios simplifiés qui ne permettent pas une évaluation contextualisée, notamment avec des informations de localisation spatiale. Ils ont utilisés les jeux de données WSJ0 (*Wall Street Journal corpus*) et THCHS-30 (*Tsinghua Chinese 30 hour database*) composés chacun de nombreuses heures de discours parlé en anglais ou en chinois, qu'ils ont ensuite combinés de façon linéaire afin d'obtenir des fichiers audios mixés.

Des modèles biologiques ont été proposés, notamment celui de Wrigley et al. (2004), qui ont développé un modèle capable de reproduire les mécanismes de l'attention sélective auditive en se basant sur les principes neurobiologiques et cognitifs de l'attention, plus précisément sur le *neural oscillation model*. L'architecture est basée sur un réseau appelé LEGION ('*Locally Excitatory Globally Inhibitory Oscillator Network*') développé par Wang et al. (1995). Il intègre des composants tels que la séparation de source sonore, la modélisation des mécanismes attentionnels et l'utilisation de filtres adaptatifs pour imiter la capacité du cerveau à isoler des sons spécifiques. Les résultats obtenus montrent que le modèle est capable de sélectionner efficacement les stimuli auditifs pertinents tout en supprimant les interférences sonores. Les simulations démontrent une performance élevée du modèle, avec une précision dans la détection des sons cibles comparable à celle observée chez les humains dans des environnements bruyants. Cependant, ce modèle n'a pas encore été évalué quant à sa capacité à rendre compte des processus *bottom-up*.

3 Objectifs

L'objectif de ce mémoire est de modéliser, à l'aide d'un modèle connexionniste, l'attention sélective auditive. Pour ce faire, nous allons adapter le modèle ASAM de Xu et al. (2018) afin de remplacer le module BiLSTM par un LSTM unidirectionnel, et ainsi obtenir un modèle conceptuellement plus proche d'un fonctionnement humain. Nous allons l'entraîner sur une tâche d'attention sélective auditive, avec un jeu de données que nous avons créé, pour permettre l'intégration d'informations sur la localisation spatiale des interlocuteurs. Nous allons ensuite évaluer la capacité du modèle à atteindre des performances proches de celles rapportées dans la littérature humaine, et nous allons notamment observer ses patterns d'erreur afin de détecter un potentiel *cocktail party effect*. Pour ce faire, nous allons évaluer l'impact de la présence, dans le discours distracteur, de mots 'saillants', c'est-à-dire, appris comme pertinents par le modèle lors de la phase d'entraînement.

Ensuite, nous comparerons les performances des deux modèles (ASAM et ASAM modifié) sur la tâche d'attention sélective et leur capacité à rendre compte des processus *top-down* et *bottom-up*.

En analyse complémentaire, nous allons également observer l'influence de l'exposition aux mots saillants sur le *cocktail party effect*. Précisément, le modèle sera entraîné sur un jeu de données contenant différents mots présentés à différentes fréquences (certains très fréquents, certains moins fréquents), et nous observerons l'influence de ces différentes fréquences sur la capacité d'attention sélective lorsqu'ils sont placés dans le discours distracteur.

Nous formulons les hypothèses suivantes :

1. Le modèle ASAM atteindra de meilleures performances que le modèle ASAM-modifié concernant le processus *top-down* (concentration sur la cible);
2. Les modèles démontreront un *cocktail party effect* : la présence de mots saillants dans le discours distracteur diminuera leurs performances comparé à une condition sans distracteur pertinent;
3. Des mots plus saillants (plus grande fréquence d'apparition dans le dataset d'entraînement) entraîneront un plus grand *cocktail party effect* que des mots moins saillants (fréquence d'apparition plus faible);
4. Le modèle ASAM-modifié présentera un *cocktail party effect* plus important que le modèle ASAM.

Part II

Modélisation

4 Tâche

Nous avons soumis les modèles à une tâche d'attention sélective : le réseau devait écouter deux personnes parlant simultanément et se focaliser uniquement sur la voix cible. Concrètement, à partir d'un signal binaural issu d'un enregistrement mélangeant deux locuteurs répartis aléatoirement dans l'espace, le réseau devait reconstruire le discours de la source cible.

L'entrée du modèle se compose du spectrogramme du mélange, noté $\mathbf{X}_{\text{mix}}(f, \tau)$. La sortie attendue est le spectrogramme de la source cible (avant le mélange), noté $\mathbf{Y}_{\text{cible}}$. Pendant l'apprentissage, le modèle compare la sortie attendue à sa prédiction

$$\widehat{\mathbf{Y}}(f, \tau) \;=\; f_\theta\big(\mathbf{X}_{\text{mix}}\big),$$

où f_θ représente la fonction apprise par le réseau, le 'filtre' qu'il apprend au fil de l'entraînement. Le modèle a pour objectif de minimiser la fonction de perte (erreur quadratique moyenne, MSE) :

$$\mathcal{L} \;=\; \big\|\, \mathbf{Y}_{\text{cible}} \;-\; \widehat{\mathbf{Y}} \,\big\|_2^2 \;=\; \sum_{f,\tau} \Big(Y_{\text{cible}}(f, \tau) \;-\; \widehat{Y}(f, \tau) \Big)^2.$$

Le modèle est soumis à la tâche en trois phases distinctes. D'abord, une phase d'entraînement permet d'estimer les paramètres θ de la fonction f_θ afin de prédire correctement le spectrogramme cible. Ensuite, une phase de validation évalue la capacité du modèle à généraliser sur un jeu de données inédit. Pour ce faire, nous utilisons 10% du jeu de données d'entraînement pour la phase de validation. Enfin, lors de la phase d'évaluation, nous testons l'hypothèse d'un effet de type *cocktail party*. Pour cela, nous introduisons dans le discours distracteur des mots qui étaient présents dans les cibles lors de l'entraînement, tandis que le discours cible contient exclusivement des mots neutres et nouveaux. Cette dernière étape permet d'évaluer l'influence du contenu sémantique appris (et notamment la fréquence d'exposition) sur les erreurs d'attention du modèle.

5 Données

Les données utilisées dans ce travail sont des fichiers audio binauraux (.wav) contenant deux locuteurs (une source cible et une source distracteur) simultanés, positionnés à deux emplacements distincts et aléatoires de l'espace. La procédure complète de génération des données se compose de trois étapes.

5.1 Génération des phrases textuelles

Les phrases (discours) ont été générées automatiquement à l'aide de ChatGPT (version 4). Nous avons demandé à ChatGPT de produire 1150 phrases simples, toutes en anglais et d'une longueur similaire, afin d'assurer une durée audio homogène. Parmi ces 1150 phrases :

- 150 contiennent le mot *baker* (boulanger), par exemple "*The baker refuses to use a microwave.*",
- 150 contiennent le mot *book* (livre), par exemple "*An old book rests on the dusty shelf.*",
- 150 contiennent le mot *hospital* (hopital), par exemple "*She works as a nurse at the hospital.*",
- 150 contiennent le mot *room* (salon), par exemple "*The living room is decorated with modern furniture.*",
- 550 sont des phrases dites "neutres" (ne contenant aucun des mots ci-dessus), par exemple "*She bought plants to decorate her balcony.*".

La place des mots signifiants dans la phrase devait également être variée, et ce de façon similaire pour tous les mots. Ces phrases textuelles ont ensuite été stockées dans cinq fichiers CSV intermédiaires (un fichier par catégorie de mot-clé), disponibles sur le dépôt GitHub suivant : https://github.com/frog-93/master_thesis.

5.2 Conversion *text-to-speech*

La synthèse vocale a été réalisée via le logiciel *Kokoro* (Hexgrad, 2025), un modèle *text-to-speech* (TTS, conversion texte en audio) en accès libre disponible sur HuggingFace. Nous avons récupéré le code Python d'origine et l'avons adapté pour automatiser la génération des fichiers audio, compte tenu du volume important (1150 phrases réparties sur 2 locuteurs). Les paramètres de synthèse utilisés étaient

fixés par défaut.

Les voix utilisées étaient issues des voix proposées par Kokoro, c'est à dire la voix américaine féminine "Heart" pour la cible, et la voix américaine masculine "Puck" pour le distracteur.

Le script Python modifié nous a permis, pour chaque fichier csv contenant les phrases générées précédemment, d'invoquer le TTS de Kokoro pour chaque phrase et de stocker le résultat dans un dossier créé pour l'occasion. La génération des données a fourni :

- **pour la cible** :
  - 50 fichiers contenant le mot "*baker*",
  - 50 fichiers contenant le mot "*book*",
  - 50 fichiers contenant le mot "*hospital*",
  - 50 fichiers contenant le mot "*room*",
  - 200 fichiers neutres;
- **pour le distracteur** :
  - 100 fichiers contenant le mot "*baker*",
  - 100 fichiers contenant le mot "*book*",
  - 100 fichiers contenant le mot "*hospital*",
  - 100 fichiers contenant le mot "*room*",
  - 350 fichiers neutres.

5.3 Génération d'audio spatiaux

Le logiciel *Audio3D* (Huckvale, 2012, version serveur) a servi à spatialiser, dans un environnement virtuel, les deux locuteurs simultanés afin de simuler une situation de "*cocktail party*". Il s'agit d'un système de simulation audio spatiale qui permet d'imiter une situation dans laquelle une ou plusieurs personnes parlent dans une pièce, tout en générant l'audio binaural entendu par un écouteur passif (qui ne parle pas) à un certain endroit de cette pièce (voir Figure 9). Pour chaque paire de phrases (cible / distracteur), nous avons généré un fichier binaural (.wav) à 2 canaux (gauche/droite) d'une durée de 5 secondes.

Pour chaque fichier final, la source cible et la source distracteur sont positionnées de façon aléatoire

autour de l'auditeur. L'auditeur (écouteur passif) est placé selon les paramètres par défaut du logiciel, c'est-à-dire, à la position (2, 1.5, 1.25), exprimée en mètres. Nous avons gardé l'intégralité des paramètres par défaut du logiciel concernant les dimensions de la pièce (4m x 3m x 2.5m), les coefficients de réverbération sur les murs (0.9), le sol (0.7) et le plafond (0.7), et le volume (en décibel) de chaque source sonore, et avons gardé ces paramètres constants pour toutes les générations effectuées. Audio3D génère automatiquement l'impulsion de tête (HRTF) correspondant à l'emplacement de chaque locuteur, puis somme les signaux dans le domaine temporel pour produire un fichier à deux canaux.

À l'aide de scripts MATLAB (récupérés auprès de l'auteur et adaptés pour notre usage), nous avons généré de façon automatique 1000 fichiers binauraux, chacun contenant un discours cible et un discours distracteur. Ainsi, ont été générés :

- Données d'entraînement : 500 fichiers binauraux (`.wav`), correspondant à :
  - 125 paires `cible - baker ; distracteur - neutre`,
  - 100 paires `cible - book ; distracteur - neutre`,
  - 75 paires `cible - hospital ; distracteur - neutre`,
  - 50 paires `cible - room ; distracteur - neutre`,
  - 150 paires `cible - neutre ; distracteur - neutre`;
- 100 fichiers contenant des paires `cible - neutre ; distracteur - neutre` afin de tester la généralisabilité de l'entraînement sur de nouvelles données;
- Données d'évaluation : 400 fichiers binauraux (`.wav`) répartis en quatre sous-datasets :
  - 100 paires `cible - neutre ; distracteur - baker`,
  - 100 paires `cible - neutre ; distracteur - book`,
  - 100 paires `cible - neutre ; distracteur - hospital`,
  - 100 paires `cible - neutre ; distracteur - room`,

Il est à noter que des fichiers sources différents ont été utilisés dans la création du dataset d'entraînement, du dataset visant à tester la généralisabilité et des quatre sous-datasets de test. Par exemple, un fichier audio de la cible prononçant une phrase neutre et utilisé dans le dataset d'entraînement ne pouvait pas se retrouver également dans un dataset de test.

Pour chaque fichier binaural généré, et afin de faciliter le traitement ultérieur des données, nous avons automatiquement créé une ligne dans un fichier `data_configurations.csv` contenant les informations suivantes : le nom du fichier binaural généré, le nom du fichier source cible, les informations de position

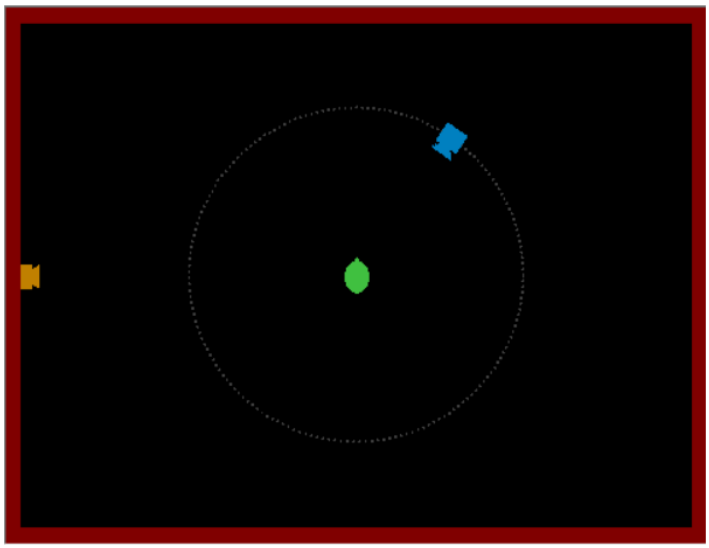


Figure 9: Exemple de placement spatial généré sur Audio3DServer. Le point vert représente le placement des écouteurs binauraux, recevant le signal audio en provenance des interlocuteurs bleu et orange placés à différents endroits de la pièce.

de la cible par rapport au locuteur, le nom du fichier source distracteur et la position du distracteur par rapport au locuteur.

Tous les scripts de génération (Python pour Kokoro, MATLAB pour Audio3D) sont disponibles via https://github.com/frog-93/master_thesis. Les fichiers audio audio simples (générés par Kokoro) et les fichiers binauraux peuvent être demandés par email.

6 Modèles

6.1 Auditory Selection with Attention and Memory (ASAM) de Xu et al. (2018)

Le modèle *Auditory Selection with Attention and Memory* (ASAM) a été développé par Xu et al. (2018) afin de simuler une situation de *cocktail party* où un locuteur doit se concentrer sélectivement sur certaines voix. Pour ce faire, son architecture repose sur deux principes fondamentaux : une mémoire à long terme pour stocker les caractéristiques des voix et un mécanisme d'attention sélective pour filtrer le son et isoler la cible. Concrètement, le modèle passe d'abord par une phase de mémorisation des voix cibles afin de les reconnaître dans un environnement contenant plusieurs locuteurs. L'architecture complète du modèle se trouve dans la Figure 10.

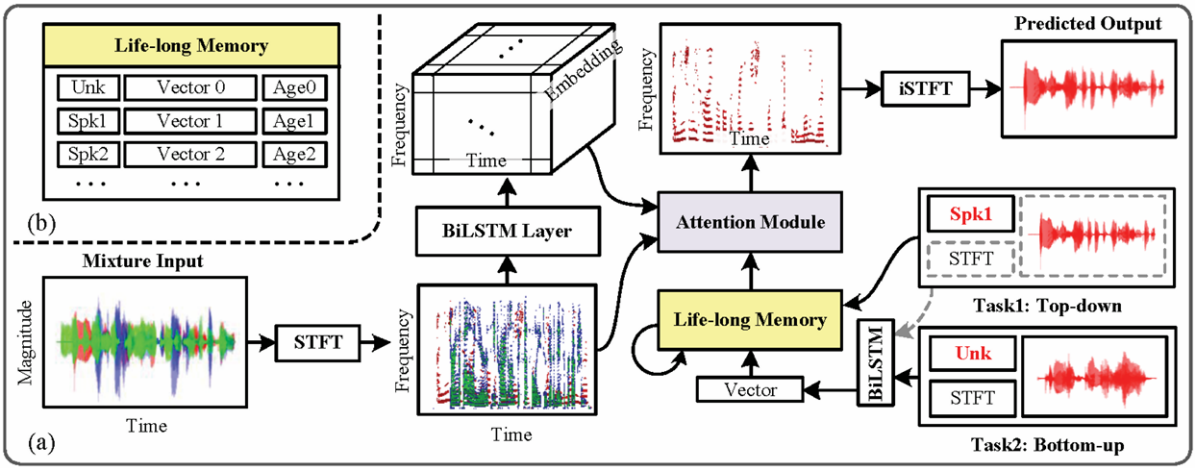


Figure 10: Architecture du modèle ASAM issu de Xu et al. (2018), composé d'un module BiLSTM ainsi que d'un mécanisme d'attention. Durant la phase d'entraînement, le modèle crée un répertoire des différents interlocuteurs qu'il utilise ensuite pendant la phase d'évaluation afin de filtrer sélectivement le discours mixte.

Spécifiquement, le modèle ASAM commence par convertir le signal audio mixte $x(t)$ en spectrogramme temps-fréquence par transformée de Fourier à court terme

$$X_{t,f} = \text{STFT}\{x(t)\},$$

où chaque trame temporelle est encodée par un BiLSTM. Ses sorties $\overrightarrow{h}_t$ et $\overleftarrow{h}_t$ sont ensuite sommées pour former h'_t puis projetées en *embeddings* temps-fréquence $h_{t,f} \in \mathbb{R}^d$. Ces vecteurs représentent les caractéristiques acoustiques de chaque point temps-fréquence du mélange audio.

Au cœur de l'architecture se trouve une mémoire externe conçue pour accumuler des connaissances sur les différents locuteurs. Le modèle crée une sorte de "carnet d'adresses" des voix, afin de les reconnaître par la suite. Cette mémoire est formalisée par $M = (K, V, A)$, où K stocke les identifiants de locuteurs (de qui s'agit-il), V leurs vecteurs acoustiques (les caractéristiques acoustiques propres à chacun), et A leur « âge » (nombre d'itérations depuis la dernière mise à jour).

Durant l'entraînement, le modèle met à jour cette mémoire après chaque nouvelle présentation d'un audio mixé, afin de raffiner ses connaissances des locuteurs (leur signature vocale). Si le locuteur est déjà connu, sa signature en mémoire est affinée en moyennant l'ancienne et la nouvelle information. Si le locuteur est inconnu, le modèle ajoute sa signature au carnet d'adresses, en remplaçant l'entrée la plus ancienne. D'abord, il extrait un vecteur de caractéristiques acoustiques v de la cible à l'aide d'un petit encodeur BiLSTM suivi d'un *pooling*. Si l'identité du locuteur p est déjà présente dans la mémoire K à la position n, alors la signature acoustique et l'âge sont mis à jour comme suit :

$$V[n] \leftarrow \frac{V[n] + v}{\|V[n] + v\|}, \qquad A[n] \leftarrow 0,$$

tandis que tous les autres $A[i]$ sont incrémentés de 1. Si le locuteur est inconnu, le modèle remplace l'entrée la plus ancienne (celle avec le $A[i]$ le plus élevé) par le nouveau triplet $(p, v, 0)$. Lors de l'inférence, on récupère directement le vecteur v associé à p.

Ce vecteur sonde v, extrait de la mémoire, agit comme un « attracteur attentionnel » pour guider le filtrage. Le modèle compare cette signature vocale à chaque point du spectrogramme mixte afin d'isoler le locuteur cible. Spécifiquement, pour chaque unité temps-fréquence du spectrogramme mixte, le modèle calcule un score d'attention bilinéaire entre l'*embedding* du mélange $h_{t,f}$ et la sonde v:

$$e_{t,f} = g^\top \tanh\big(W\, v + U\, h_{t,f}\big),$$

où W, U et g sont des paramètres appris. Ce score est ensuite passé à travers une fonction sigmoïde pour produire un masque d'attention :

$$\alpha_{t,f} = \sigma\big(e_{t,f}\big),$$

et est appliqué au spectrogramme initial pour en extraire la source d'intérêt :

$$\widetilde{X}_{t,f} = \alpha_{t,f}\, X_{t,f}.$$

La source isolée est finalement reconstruite par transformée de Fourier inverse iSTFT$\{\widetilde{X}\}$.

L'ensemble des paramètres (ceux des BiLSTM, des matrices W, U, du vecteur g, et de la mémoire) est entraîné *end-to-end* en minimisant la perte spectrale quadratique entre le spectrogramme prédit et celui de la source cible S

$$\mathcal{L} = \sum_{t,f} \left\| S_{t,f} - \widetilde{X}_{t,f} \right\|_2^2$$

Dans le cadre de ce mémoire, le code original, initialement conçu pour Python 2.7 et Theano, a été entièrement ré-implémenté et modernisé pour être compatible avec les librairies actuelles, notamment Python 3, TensorFlow et Keras. Cette manipulation s'est révélée essentielle au vu du temps d'entraînement nécessaire du modèle initial (30h pour 4 epochs sur CPU). La nouvelle implémentation a permis l'utilisation d'un GPU, réduisant l'entraînement à quelques dizaines de minutes. De plus et afin d'éviter un sur-apprentissage observé lors des entraînements initiaux, l'architecture a été régularisée par l'ajout de couches de *dropout* après les principaux modules BiLSTM, et le taux d'apprentissage de l'optimiseur Nesterov-Adam a été ajusté à une valeur plus faible de .0001. Enfin, leur code original comprend déjà un algorithme permettant de mixer les deux audios source de façon linéaire afin d'obtenir l'audio mixe. Cependant, notre dataset comprenant déjà ces audios mixés, nous avons désactivé cette fonctionnalité.

6.2 ASAM - LSTM

Le modèle précédent a été adapté afin d'utiliser un LSTM unidirectionnel, et non plus bidirectionnel. Le seul changement consiste en l'encodage temporel du spectrogramme. Chaque trame temporelle t du spectrogramme X est désormais encodée par un LSTM unidirectionnel, dont la sortie est notée h_t. Contrairement à la version bidirectionnelle qui produisait deux vecteurs $\overrightarrow{h}_t$ et $\overleftarrow{h}_t$ ensuite sommés en h'_t, nous utilisons ici directement la sortie h_t pour chaque pas de temps.

Ce vecteur h_t est ensuite projeté linéairement en un *embedding* temps-fréquence $h_{t,f} \in \mathbb{R}^d$ comme dans la version originale. Le reste de l'architecture demeure inchangé. En particulier, le mécanisme d'attention bilinéaire repose toujours sur le calcul du score $e_{t,f}$, du masque d'attention $\alpha_{t,f}$ et du spectrogramme masqué $\widetilde{X}_{t,f}$ reconstruit ensuite par transformée de Fourier inverse.

La mémoire externe (K, V, A), les règles de mise à jour et d'inférence ainsi que la fonction de perte restent identiques.

7 Analyses statistiques

Les performances des modèles ont été évaluées à l'aide de trois métriques distinctes. Tout d'abord, le "*Signal-to-Distortion-Ratio*" (SDR) est une mesure globale de la qualité de reconstruction du signal source. Il prend en compte l'ensemble des erreurs présentes dans le signal reconstitué, à la fois les distorsions, les interférences et le bruit résiduel. Ensuite, le "*Signal-to-Interference-Ratio*" (SIR) quantifie spécifiquement la capacité du modèle à séparer le signal cible des sources interférentes. Ici, les interférences désignent les portions de signal provenant d'autres sources sonores que celle que l'on cherche à extraire. Par exemple, si le modèle conserve des portions des voix distractrices dans le signal de sortie, cela correspond à des interférences. Enfin, le "*Signal-to-Artifact-Ratio*" (SAR) évalue le niveau d'artefacts introduits par le processus de séparation. Les artefacts sont des sons ou déformations artificielles, qui ne proviennent ni de la source cible ni des autres sources, mais qui sont créés par le modèle lors du traitement du signal (un grésillement par exemple). Conjointement, ces trois mesures permettent de saisir la qualité globale de la séparation de source (Vincent et al., 2006). Exprimé en décibels (dB), un score plus élevé indique de meilleures performances.

Afin d'évaluer et de comparer les performances des deux architectures de modèle (ASAM original et ASAM modifié) à travers cinq conditions acoustiques distinctes, une analyse de variance multivariée (MANOVA) a été réalisée. Le plan factoriel était un 2 (Type de Modèle : LSTM vs BiLSTM) x 5 (Condition d'Évaluation). Cette approche a été privilégiée car les trois variables dépendantes (SDR, SIR, SAR) sont des mesures conceptuellement liées, et sont fortement corrélées ($r(sar, sdr) = .9263, r(sar, sir) = .6424, r(sir, sdr) = .8761$). L'utilisation d'une MANOVA permet un test simultané des différences de groupe sur l'ensemble de ces métriques tout en contrôlant le risque d'erreur de type I, qui serait autrement augmenté par la conduite d'analyses de variance univariées distinctes (Warne, 2014).

La validité de la MANOVA repose sur plusieurs postulats. Tout d'abord, l'indépendance des observations est assurée par l'utilisation de réseaux de neurones artificiels. En effet, une fois la phase d'entraînement terminée, les poids des modèles sont gelés, les rendant statiques. Par conséquent, chaque essai de test est un événement indépendant. Non seulement la performance dans une condition n'a aucune influence sur les autres, mais les 100 répétitions au sein d'une même condition sont également indépendantes les unes des autres. Contrairement à un participant humain, le modèle n'est sujet à aucun effet d'apprentissage, de fatigue ou de report d'une épreuve à l'autre. Chaque condition est donc traitée comme une observation indépendante. Ensuite, l'hypothèse de normalité multivariée a été testée

via un test de Mardia appliqué aux résidus du modèle. Les résultats ont révélé une violation significative de l'asymétrie ($skew = 2167.8, p < .001$) ainsi que de l'aplatissement ($kurtosis = 53.06, p < .001$). Enfin, le test M de Box, utilisé pour tester l'homogénéité des matrices de variance-covariance, s'est avéré significatif ($M = 104, p < .001$), indiquant une violation de ce postulat. Cependant, des tests de Levene indépendants pour tester l'homogénéité des variances n'ont révélé aucune différence significative pour le SDR ($F(9, 978) = 1.33, p = .215$), le SIR ($F(9, 978) = 1.65, p = .097$), ou le SAR ($F(9, 978) = 0.81, p = .606$). En conséquence de ces violations, la Trace de Pillai a été utilisée pour l'interprétation des résultats de la MANOVA, étant donné que cet indice est considéré comme le plus robuste en cas de violation des postulats (Olson, 1976).

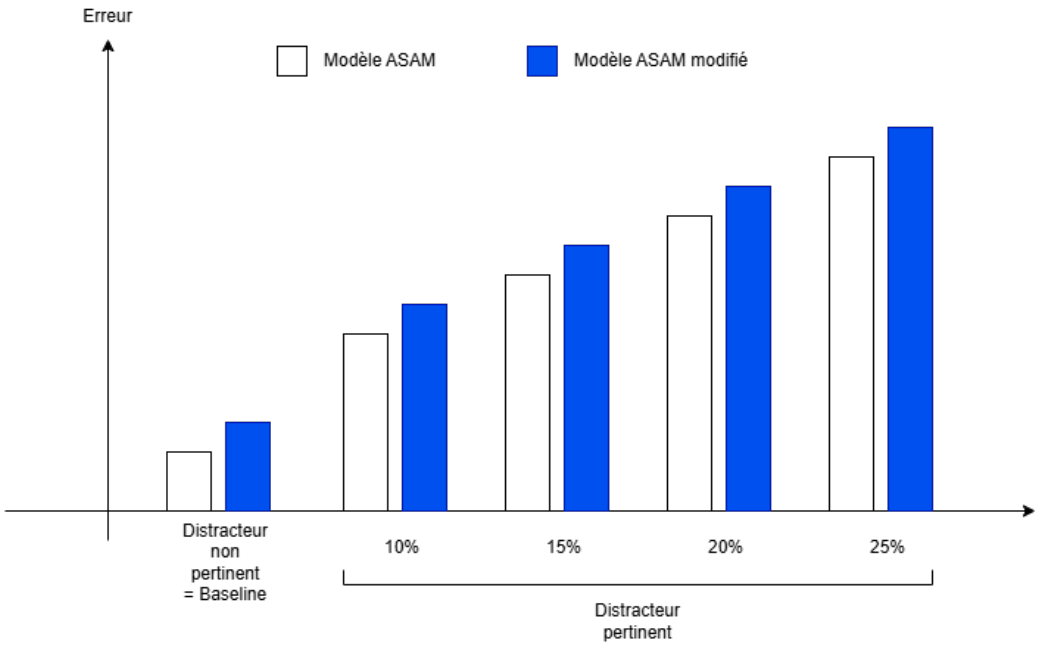


Figure 11: Résultats attendus pour la comparaison entre les modèles et les conditions d'évaluation. L'erreur correspond à la performance des modèles, mesurée via les indices de séparation de source.

Des analyses post-hoc univariées ont été réalisées sur la variable SDR, étant donné qu'il s'agit de la plus complète des trois. Une analyse de variance (ANOVA) a été réalisée afin de s'assurer que les effets observés dans la MANOVA étaient également présents en univarié. Plusieurs tests post-hoc ont ensuite été réalisés. Premièrement, afin d'évaluer l'influence des distracteurs pertinents sur les capacités des modèles, un contraste linéaire planifié a été réalisé. Ce test a opposé la condition "sans distracteur pertinent" aux quatre autres conditions d'évaluation (avec distracteur pertinent). Ensuite, afin d'identifier quelles conditions d'évaluation différaient significativement l'une de l'autre, des comparaisons multiples par paires ont été réalisées, avec un ajustement de Holm-Bonferoni pour contrôler l'erreur de type I. Les résultats attendus sont présentés dans la Figure 11.

Comme mentionné précédemment, chaque modèle sera testé dans chaque condition sur 100 échantillons distincts, ce qui fait un total de 1000 observations. Une analyse de puissance a permis de déterminer qu'avec une taille d'échantillon aussi grande, il nous sera possible de détecter des effets aussi petits que $f = .0114$ avec une puissance de 95%. La littérature sur la capture involontaire de l'attention par des stimuli signifiants met en évidence des effets moyens à grands (e.g. $\eta_p^2 = .11$ dans Kim et al., 2021). L'analyse de puissance a été réalisée via le logiciel G*Power (Faul et al., 2007, version 3.1.9.7), et pour une analyse de type '*MANOVA: Special effects and interaction*'. L'aspect "répété" des conditions n'a ici pas été pris en compte étant donné que le fait de travailler avec des modèles computationnels garanti l'indépendance des mesures. Toutes les analyses ont été réalisées en R via le logiciel RStudio (version 2025.05.1+513, Posit team, 2025) et avec l'utilisation du package emmeans (Lenth, 2025).

Le code d'analyse et le fichier de données utilisés dans ce mémoire sont disponibles sur le dépôt GitHub : https://github.com/frog-93/master_thesis.

Part III

Résultats

La Figure 12 présente les performances (en dB) des deux modèles dans les différentes conditions d'évaluation, et ce pour les trois métriques d'intérêt. Des exemples qualitatifs des spectrogrammes prédits pour chaque condition sont disponibles à l'Annexe B.

Les deux modèles ont été entraînés en utilisant une procédure d'arrêt précoce basée sur la performance sur un jeu de validation, avec une patience de 10 époques. Cela signifie que le modèle arrête automatiquement son entraînement si l'erreur en phase de validation ne s'est pas améliorée sur les 10 dernières époques.

Le modèle BiLSTM a atteint sa perte de validation minimale dès la première époque. L'entraînement s'est ensuite poursuivi jusqu'à la 11ème époque avant d'être interrompu par le mécanisme d'arrêt précoce, la performance sur le jeu de validation ne s'améliorant plus. Le modèle LSTM, quant à lui, a montré un comportement d'apprentissage légèrement plus lent. Il a atteint sa meilleure performance de validation à la 7ème époque, et l'entraînement complet a duré 17 époques.

On observe que les scores SIR moyens sont positifs pour les deux modèles dans la plupart des conditions, suggérant que les deux architectures ont appris à supprimer une partie du signal interférent. Cependant, les scores SAR et SDR sont systématiquement négatifs, indiquant que les signaux de sortie contiennent encore des distorsions et des artéfacts de traitement notables. Ces résultats suggèrent que la qualité de la séparation de source reste limitée suite à la phase d'entraînement.

Une MANOVA factorielle 2 (Type de Modèle) x 5 (Condition) a été conduite sur les métriques SDR, SIR et SAR. L'analyse a révélé un effet principal du type de modèle (Trace de Pillai $= .0435, F(3, 976) = 14.8095, p < .001$) ainsi qu'un effet principal de la condition (Trace de Pillai $= .1095, F(12, 2934) = 9.2678, p < .001$), mais pas d'effet d'interaction (Trace de Pillai $= .0019, F(12, 2934) = .1596, p = .9995$). Ces résultats indiquent que (1) l'une des deux architectures de modèle est globalement plus performante que l'autre, (2) la performance varie significativement en fonction de l'environnement acoustique, et (3) la différence entre les modèles reste stable à travers les différentes conditions.

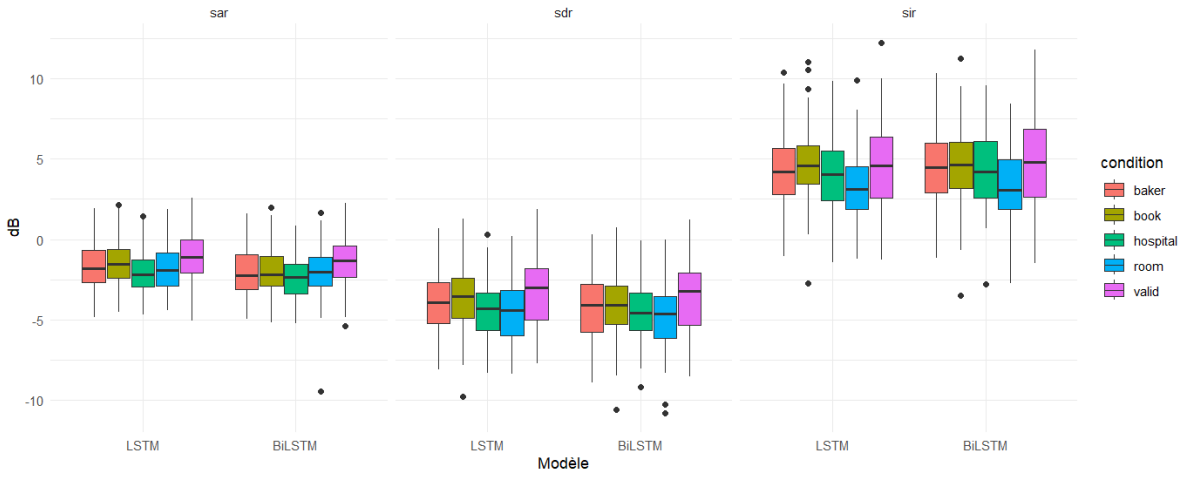


Figure 12: Performances (médiane et intervalle inter-quartile exprimés en dB) des deux modèles dans les différentes conditions d'évaluation. SAR = *Signal-to-Artifact Ratio*. SDR = *Signal-to-Distortion Ratio*. SIR = *Signal-to-Interference Ratio*. Le modèle BiLSTM correspond au modèle ASAM original. Les différentes conditions correspondent aux fréquences d'apparition suivantes dans le discours cible de la phase d'entraînement : *baker* = 25%, *book* = 20%, *hospital* = 15%, *room* = 10%, *valid* = 0% (cette dernière correspond à la condition sans distracteur pertinent).

8 Comparaison des modèles sur la tâche d'attention sélective

Les analyses suivantes ont été réalisées au niveau univarié sur la métrique SDR. Préalablement, une ANOVA 2 (Type de modèle) X 5 (Condition) a été réalisée afin de vérifier que les effets mis en évidence par la MANOVA étaient également présents en univarié. L'analyse a conclu aux mêmes effets principaux du type de modèle ($F(1, 978) = 4.825, p = .0283$) et de la condition ($F(4, 978) = 10.624, p < .001$), et n'a pas non plus révélé d'effet d'interaction ($p = .9799$).

Ainsi, le modèle LSTM a démontré des performances significativement supérieures au BiLSTM. Cependant, cette différence reste faible, avec un avantage d'environ $0.28dB$ en faveur du LSTM pour la métrique SDR.

9 Évaluation de l'influence de la fréquence d'apparition lors de l'entraînement

Le contraste planifié opposant la condition sans distracteur pertinent aux quatre autres conditions s'est révélé significatif ($t(978) = 4.861, p < .001$). Ce résultat met en évidence que les deux modèles ont eu de meilleures performances lorsque le distracteur contenait des mots non pertinents, par rapport aux conditions où le distracteur contenait des mots pertinents, c'est-à-dire appris lors de l'entraînement.

Les comparaisons multiples par paires avec correction de Tukey ont mis en évidence plusieurs différences significatives (voir Table 1), confirmant les résultats du contraste planifié. Plus précisément, la condition sans distracteur pertinent ('*valid*') a mené à des performances significativement plus élevées que les autres conditions. De manière plus inattendue, les résultats dans la condition '*book*' ont été significativement meilleurs que dans les conditions '*room*' ($p = .0003$) et '*hospital*' ($p = .0360$). Cela dit, les tailles d'effet associées à ces différences restent petites.

Table 1: Comparaisons par paires entre conditions avec correction de Tukey.

Comparaison	**Estimation**	**SE**	**df**	**t-ratio**	**p-valeur**	***d* de Cohen**
baker – book	-0.279	0.182	978	-1.531	0.5422	-0.14
baker – hospital	0.243	0.183	978	1.330	0.6728	0.12
baker – room	0.484	0.183	978	2.645	0.0632	0.24
baker – valid	-0.587	0.182	978	-3.232	0.0111*	-0.29
book – hospital	0.522	0.183	978	2.850	0.0360*	0.26
book – room	0.763	0.184	978	4.159	0.0003***	0.38
book – valid	-0.308	0.182	978	-1.693	0.4390	-0.15
hospital – room	0.242	0.184	978	1.313	0.6834	0.12
hospital – valid	-0.830	0.183	978	-4.545	0.0001***	-0.42
room – valid	-1.072	0.183	978	-5.853	<0.0001***	-0.54

Note : * $p < .05$, ** $p < .01$, *** $p < .001$

*Note 2 :*Les différentes conditions correspondent aux fréquences d'apparition suivantes dans le discours cible de la phase d'entraînement : *baker* = 25%, *book* = 20%, *hospital* = 15%, *room* = 10%, *valid* = 0%.

Part IV

Discussion

Ce mémoire visait à modéliser l'attention sélective auditive à l'aide de deux modèles computationnels distincts. Pour ce faire, nous avons entraîné, sur une tâche d'attention sélective auditive, le modèle ASAM issu de Xu et al. (2018) ainsi qu'une version modifiée avec un nouveau jeu de données que nous avons créé. Les deux modèles ont démontré des performances globalement médiocres à la tâche. Ce mémoire avait également pour objectif d'observer l'influence de différents niveaux de saillance de certains mots sur le *cocktail party effect*. Pour ce faire, nous avons regardé si des distracteurs précédemment appris comme pertinents lors de la phase d'entraînement pouvaient modifier les performances des modèles. Nos résultats ont mis en évidence que des mots vus dans le discours cible lors de la phase d'entraînement engendrent une diminution significative des performances lorsqu'ils sont présentés dans le discours distracteur, et ce, de façon non équivalente entre les différents mots.

10 Performances globales des deux architectures

L'analyse des performances révèle que les deux modèles, bien qu'ayant appris certains aspects de la tâche, n'atteignent qu'une qualité de séparation modeste, comme en témoignent les scores SDR et SAR systématiquement négatifs. L'explication de ces performances limitées peut être trouvée dans l'analyse de la phase d'entraînement. Il semblerait qu'un mécanisme de sur-apprentissage (*overfitting*) soit entré en jeu. Ce phénomène se produit lorsqu'un modèle à haute capacité, plutôt que d'extraire des règles générales, commence à mémoriser les exemples spécifiques de l'ensemble d'entraînement, perdant ainsi sa capacité à généraliser à de nouvelles données non vues (Goodfellow et al., 2016). Plusieurs indices dans nos résultats pointent vers ce diagnostic.

Le premier argument provient de l'activation rapide de l'arrêt précoce après le début de l'entraînement. Cette règle d'arrêt fait en sorte que l'entraînement s'arrête si les performances sur les données de validation ne s'améliorent pas après 10 époques. Le fait que cela se produise aussi rapidement (1 époque pour le BiLSTM, 7 époques pour le LSTM) signifie que la performance sur le jeu de données de valida-

tion a cessé de s'améliorer presque immédiatement après le début de l'entraînement. Les modèles ont donc atteint rapidement un "plafond" de généralisation, qui s'est avéré être à un niveau de performance médiocre.

Un autre argument en faveur de l'hypothèse du sur-apprentissage est que les deux modèles ont de très grandes capacités de mémorisation au vu des architectures utilisées (plus de 6 millions de paramètres chacun). Confrontés à un dataset d'entraînement de petite taille (500 exemplaires dans notre cas), les modèles ont pu atteindre un point où il était devenu mathématiquement plus simple de "tricher" en mémorisant les spectrogrammes d'entraînement plutôt que d'apprendre la tâche complexe de la séparation de sources. En comparaison, l'article original de Xu et al. (2018) a utilisé des jeux de données plus large (WSJ0 et THCHS-30 composés chacun de 4,410 exemples d'entraînement) ce qui a probablement permis de mitiger ce phénomène. En effet, la séparation de sources sonores est une tâche notoirement difficile pour les réseaux de neurones (Parande and Thomas, 2017; D. Wang and Chen, 2018) qui nécessite de nombreuses données d'entraînement. Par exemple, le modèle de référence en séparation de sources, Conv-TasNet, a été entraîné sur 20 000 audio du dataset WSJ0-2mix (Luo and Mesgarani, 2019).

11 Le sur-apprentissage et ses conséquences

Nos résultats ont démontré un effet principal du type de modèle lors de la MANOVA, mettant en évidence de meilleures performances du LSTM face au BiLSTM. Ce résultat est contre-intuitif. En effet, on s'attendrait théoriquement à ce que le modèle BiLSTM surpasse le modèle LSTM unidirectionnel. En traitant la séquence audio dans les deux sens (passé vers futur et futur vers passé), le BiLSTM dispose d'un contexte temporel plus riche pour chaque trame, ce qui se révèle crucial pour des tâches de traitement de la parole (Graves and Schmidhuber, 2005). Plusieurs hypothèses peuvent expliquer ce phénomène. Tout d'abord, on pourrait supposer un faux positif en provenance de la MANOVA. Cependant, la Trace de Pillai est l'une des statistiques les plus robustes de la MANOVA face à des violations des postulats (Olson, 1976).

Une seconde hypothèse est liée au sur-apprentissage. Le nombre de paramètres d'un modèle est directement lié à sa capacité, c'est-à-dire son habileté à s'adapter à des fonctions complexes. Un modèle avec une capacité trop élevée par rapport à la complexité et à la taille du jeu de données d'entraînement

est particulièrement susceptible de sur-apprendre en mémorisant les données plutôt qu'en généralisant (Goodfellow et al., 2016). Le modèle BiLSTM, étant plus complexe et ayant plus de paramètres que le LSTM (plus de 7 millions pour le BiLSTM, contre 6,4 millions pour le LSTM), possède une capacité supérieure. Confronté à notre jeu de données de taille limitée, il est donc possible qu'il ait sur-appris encore plus rapidement que le LSTM, et que son point d'arrêt précoce ait capturé un état de généralisation sous-optimal par rapport à son homologue plus simple. Cette interprétation est d'autant plus plausible que, bien que statistiquement significative, la différence de performance observée entre les deux modèles reste faible. Une différence aussi petite est plus probablement le résultat de différences dans les dynamiques d'apprentissage, plutôt que d'une supériorité fondamentale de l'architecture la plus simple.

12 Un cocktail party effect en trompe-l'oeil

Comme attendu, l'analyse du contraste planifié a révélé ce qui s'apparente à un *cocktail party effect*. L'introduction de distracteurs pertinents dans la phase d'évaluation a réduit les performances des deux modèles en comparaison à une condition sans distracteur pertinent. Ce résultat suggère la prise en compte du contenu sémantique à un certain niveau. Cependant, au vu de notre paradigme et des architectures utilisées, il est plus probable que la trace acoustique des mots pertinents ait été retenue/encodée dans les poids des modèles, sans réellement correspondre à l'apprentissage d'une quelconque signification des mots. Il est possible que les modèles, n'ayant pas de véritable compréhension sémantique de part leur structure, aient appris à associer les traces acoustiques et spectrales des mots cibles à une sortie "non filtrée". Lorsque ces mêmes traces acoustiques apparaissent dans le canal du distracteur, elles créent une interférence ou une ambiguïté pour le mécanisme d'attention du modèle, qui peine alors à les supprimer.

Les comparaisons multiples ont mis en évidence de meilleures performances dans la condition avec le distracteur pertinent "*book*" comparé aux autres conditions avec distracteur pertinent. Cela suggère que cette condition a été plus simple que les autres. Or, ce mot était présent à une fréquence de 20% dans le dataset d'entraînement. En suivant nos hypothèses, cette condition aurait donc dû être plus difficile que les autres. Ce résultat intriguant pourrait provenir d'un faux positif dû à l'utilisation d'une correction de Tukey dans nos comparaisons multiples. Bien que la correction de Tukey soit une procédure standard, elle est moins conservatrice que d'autres procédures comme celles de Bonferroni ou de Scheffé. Notamment, le risque de faux positifs reste prompt à l'inflation dans ce genre d'analyses,

particulièrement en présence de violations de la normalité (Seaman et al., 1991). Son utilisation dans le cadre de ce mémoire se justifie par la volonté de réaliser des analyses très larges, quitte à détecter des faux positifs. Par conséquent, la possibilité qu'un résultat marginalement significatif comme celui-ci soit un faux positif ne peut être entièrement écartée.

Cependant, il est également possible que ce résultat reflète une différence réelle. Les caractéristiques acoustiques des mots eux-mêmes pourraient jouer un rôle, leurs propriétés acoustiques intrinsèques les rendant plus ou moins perturbateurs. Cela fait référence aux concepts de masquage énergétique et informationnel (Durlach et al., 2003).

Le masquage énergétique se produit au niveau du système perceptig lorsque l'énergie acoustique d'un son submerge celle de l'autre son. Plus précisément, les bandes de temps et de fréquence se recouvrent, et il devient alors compliqué de séparer physiquement les deux sons. Il s'agit d'une interférence de bas niveau, un problème de rapport signal/bruit qui rend la cible physiquement inaudible ou indétectable (Shinn-Cunningham, 2008). Il pourrait s'agir d'un bruit sourd et très fort qui se produirait pendant un discours cible. Le bruit en question perturbe notre écoute de part ses caractéristiques purement acoustiques. Le masquage informationnel, en revanche, est un phénomène d'ordre supérieur qui intervient au niveau cognitif du traitement auditif (Durlach et al., 2003). Il apparaît même lorsque la cible est techniquement audible (pas de chevauchement physique). L'interférence provient de la confusion, de l'incertitude ou de l'incapacité du système à séparer les flux auditifs au niveau de leur contenu (Pollack, 1975). Ce type de masquage est particulièrement prononcé lorsque le distracteur partage des caractéristiques similaires avec la cible, créant une compétition attentionnelle (Kidd et al., 2008). Par exemple, de la parole masquant de la parole rentre dans cette catégorie car ils partagent le même type d'information : du langage.

Le mot "*book*" est acoustiquement simple car il est monosyllabique et composé de consonnes occlusives (/b/, /k/) qui représentent des impulsions d'énergie brèves et bien définies dans le temps (Stevens and Blumstein, 1978). En comparaison, des mots comme "*hospital*" et "*room*" possèdent des caractéristiques acoustiques intrinsèquement plus "bruyantes" et complexes. En effet, le mot "*hospital*" contient une consonne fricative /s/, un son qui, par nature, est un bruit à large bande spectrale, c'est-à-dire qu'il étale son énergie sur une large gamme de fréquences (Jongman et al., 2000). Cette "rafale de bruit" est un candidat idéal pour créer un masquage énergétique efficace, en noyant physiquement des composantes spectrales de la parole cible (Cooke, 2006). De plus, la structure polysyllabique et la durée

plus longue de "*hospital*" en font un objet de traitement plus complexe (Baddeley et al., 1975). Cette complexité augmente le masquage informationnel. Le modèle doit dépenser plus de "ressources attentionnelles" pour suivre ce flux concurrent et le distinguer de la cible, créant une confusion qui dégrade la performance (Brungart, 2001).

Ainsi, le mot "*book*", étant acoustiquement plus simple, générerait principalement un masquage énergétique de courte durée et moins étendu spectralement. À l'inverse, "*hospital*" provoquerait à la fois un masquage énergétique plus important à cause de sa consonne fricative, et un masquage informationnel accru en raison de sa plus grande complexité structurelle. Cette double source d'interférence expliquerait pourquoi il s'agit d'un distracteur plus efficace pour le modèle, menant à de moins bonnes performances, et pourquoi la condition "*book*" était comparativement plus facile.

Ces résultats ouvrent une piste de recherche intéressante sur l'influence des caractéristiques phonétiques des mots distracteurs sur l'effet *cocktail party*. Il pourrait être intéressant de tester de manière systématique cette hypothèse en créant des jeux de distracteurs contrôlés. On pourrait par exemple comparer des paires de mots ne différant que par la présence ou l'absence d'une consonne fricative, ou comparer des mots monosyllabiques et polysyllabiques ayant des profils d'amplitude et de complexité spectrale variés. Une telle expérience permettrait d'isoler plus précisément la contribution respective du masquage énergétique et informationnel dans la dégradation des performances du modèle et de quantifier l'impact de caractéristiques phonétiques spécifiques sur la séparation de sources.

13 Limites et perspectives

La limite majeure de ce travail est liée à l'adaptation du modèle ASAM. Initialement codé en Python 2 avec une version dérivée de Keras, nous avons dû l'adapter en Python 3 pour Keras et TensorFlow pour des raisons pratiques (temps d'entraînement notamment). Bien que la logique du modèle ait été respectée, il se peut que des subtilités se soient perdues dans l'adaptation. Des différences subtiles dans les implémentations des couches LSTM ou des optimiseurs entre Theano et TensorFlow pourraient avoir un impact sur les résultats. De plus, nos propres modifications (dropout, taux d'apprentissage) font de notre modèle une variante de l'ASAM original plutôt qu'une copie conforme.

Une deuxième limite provient de la taille du dataset d'entraînement utilisé. Comme mentionné

précédemment, le faible nombre d'exemples présentés lors de la phase d'entraînement a pu générer du sur-apprentissage. Une solution pour pallier à ce problème pourrait être d'utiliser des techniques d'augmentation de données (*data augmentation*, Z. Wang et al., 2025). Il s'agit de techniques permettant d'augmenter la taille d'un dataset de façon artificielle, c'est-à-dire, sans réellement acquérir de nouvelles données. Cela consiste à modifier légèrement les données de l'entraînement afin d'obtenir des versions proches, mais modifiées (par exemple, en ajoutant du bruit à l'audio initial). Cependant, toutes les techniques d'augmentation des données ne sont pas applicables dans le cadre de ce mémoire.

Par exemple, l'ajout d'un bruit de fond aléatoire pourrait interférer avec notre objectif principal qui est de mesurer l'impact spécifique de distracteurs sémantiques dans un contexte comprenant uniquement deux locuteurs. Par contre, d'autres techniques pourraient être envisagées. Notamment, de légères variations de hauteur ou d'étirement temporel pourraient constituer des données valables dans le cadre de cette expérience, tout en étant considérées comme très différentes des données originales par les modèles. Idéalement, la perspective la plus rigoureuse resterait cependant la collecte de nouvelles données en suivant le protocole de création de notre dataset, afin d'augmenter le nombre d'exemples uniques de locuteurs et de phrases.

Enfin, ce travail pourrait être étendu en remplaçant l'architecture LSTM par des modèles permettant le traitement sémantique de l'information. En effet, nos résultats ont montré une interférence basée sur la trace acoustique des mots mais, pour tester une influence sémantique, il faudrait que le modèle ait accès à la signification des mots. Cela pourrait être accompli en intégrant un enchâssement sémantique (*word embedding*) dans l'architecture. La technique la plus connue, Word2Vec (Mikolov et al., 2013), apprend à représenter les mots sous forme de vecteurs dans un espace à haute dimension où la distance entre les vecteurs reflète la similarité sémantique (par ex., les vecteurs pour 'roi' et 'reine' sont proches mais éloigné du vecteur pour 'banane').

Concrètement, on pourrait mettre en place un modèle à double voie qui traiterait l'information de deux façons distinctes, en parallèle, avant de l'intégrer. La première voie, acoustique, serait similaire au modèle LSTM actuel. Elle analyserait le mix audio afin de réaliser une première séparation brute des sources, et donc prédire la cible. En parallèle, une seconde voie, sémantique, opérerait à un niveau plus abstrait. Celle-ci utiliserait les phrases textuelles qui seraient transformées en vecteurs sémantiques par un module d'enchâssement pour en capturer la signification. Le modèle pourrait ainsi apprendre quels mots sont pertinents pour identifier la cible. Dans le cadre de notre tâche, on pourrait s'attendre à ce

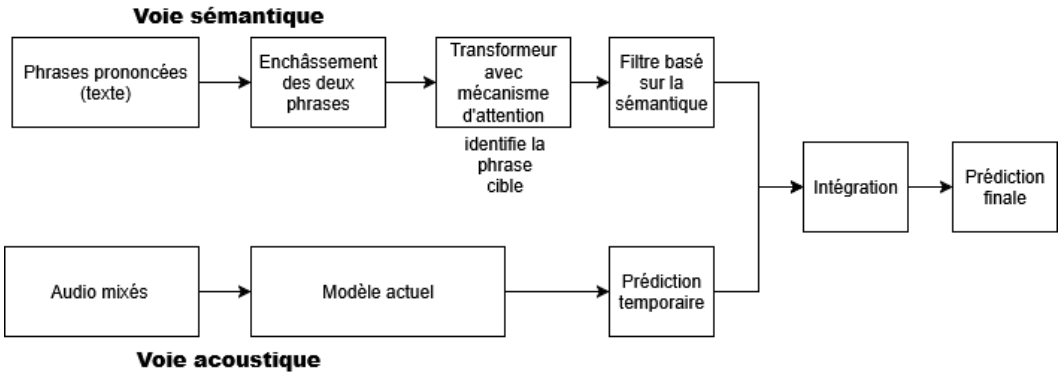


Figure 13: Schéma conceptuel de l'architecture à double voie proposée afin d'intégrer la gestion des aspects sémantiques. Le modèle combine une voie acoustique (inférieure) pour la séparation brute du signal, et une voie sémantique (supérieure) qui génère un filtre attentionnel à partir du contenu textuel. L'information des deux voies est ensuite intégrée pour produire la prédiction finale.

qu'un tel module identifie les mots pertinents tels que '*baker*' comme étant importants.

On trouverait ensuite un module d'intégration qui permettrait de prendre en compte les sorties de ces deux voies. L'information sémantique issue de la seconde voie agirait comme un filtre attentionnel de haut niveau. Elle viendrait guider et affiner le traitement de la voie acoustique en lui fournissant une attente sur le contenu à extraire. Il s'agirait d'une sorte de 'modèle prédictif'. Cela permettrait de mieux anticiper et isoler la séquence de sons correspondant à la phrase cible. La prédiction finale de la parole cible ne découlerait donc plus seulement de l'audio pur, mais de l'interaction dynamique entre le traitement du signal et la compréhension du sens. Un tel modèle dépasserait la simple séparation de sources pour simuler la manière dont la connaissance du contexte nous aide activement à suivre une conversation dans un environnement bruyant. On pourrait également faire varier le poids accordé à chaque information (acoustique, sémantique), et ainsi observer l'effet de chacune sur les prédictions du modèle. La Figure 13 illustre cette potentielle architecture.

Cependant, les techniques d'enchâssement sont également questionnables quant à leur prise en compte de la "sémantique", comme cela est largement discuté dans Defays (2023). Bien qu'elles permettent au modèle de se représenter les relations entre les mots et constituants de la phrase, rien ne garanti que cela est suffisant pour permettre une réelle compréhension du langage. Il s'agit d'un débat philosophique fondamental en intelligence artificielle, qui a été popularisé par l'expérience de la chambre chinoise (Searle, 1980). Est-ce que la manipulation du langage, même de façon très sophistiquée, suffit à produire une compréhension ? Ce débat est encore ouvert et dépasse le cadre de ce mémoire.

14 Conclusion

La modélisation en psychologie cognitive a vu apparaître de nouvelles techniques de modélisation suite à l'avènement de l'intelligence artificielle et du *deep learning*. L'utilisation de réseaux de neurones artificiels permet maintenant une modélisation précise, reproductible et falsifiable qui n'était pas possible avec les modèles boîtes-flèches. Bien que de nombreux domaines de recherche aient ainsi bénéficié de cet avancement, l'attention auditive semble être restée à l'écart. Ainsi, peu de modèles computationnels sont disponibles pour tenter d'expliquer les phénomènes de type *cocktail party effect*. Parmi eux, le modèle ASAM (Xu et al., 2018), composé d'un BiLSTM et d'un mécanisme d'attention, semble atteindre des performances prometteuses en simulant ce phénomène. Cependant, un BiLSTM a, par nature, accès à l'ensemble d'une séquence à chaque instant de traitement. Or, cela n'est pas fidèle à un mode de fonctionnement humain.

Afin de modéliser l'attention sélective auditive à l'aide d'un modèle conceptuellement plus proche du fonctionnement humain, nous avons entraîné une version modifiée du modèle ASAM sur un nouveau jeu de données. Afin d'observer l'émergence d'un *cocktail party effect*, nous avons également investigué l'influence de distracteurs précédemment appris comme pertinents lors de la phase d'entraînement sur les performances du modèle dans la phase d'évaluation. Ainsi, nous avons testé si l'influence de plusieurs mots, présentés à des fréquences différentes dans le discours cible lors de l'entraînement, pouvait perturber la phase d'évaluation lorsqu'ils étaient présents dans le discours distracteur.

Une MANOVA a mis en évidence de meilleures performances du modèle modifié comparé au modèle ASAM original. Bien que contre-intuitif, ce résultat pourrait être expliqué par un problème de sur-apprentissage lié à la complexité des modèles utilisés et à la taille limitée de notre dataset. Nous avons également observé l'émergence de ce qui semble être un *cocktail party effect*. Les mots appris comme pertinents ont diminué la performance des deux modèles lorsqu'ils étaient présents dans le discours distracteur, et certains mots sont apparus comme moins distracteurs que d'autres. La structure des modèles ne permettant pas de traitement sémantique à proprement parler, l'hypothèse privilégiée est celle d'un masquage énergétique.

Ce mémoire invite à repenser l'inclusion du traitement sémantique dans nos modèles computationnels et à investiguer le niveau de familiarité qui pourrait influencer le *cocktail party effect*.

Références

Atkinson, R. C., & Shiffrin, R. N. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, 89–195. https://doi.org/10.1016/S0079-7421(08)60422-3

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *11*(4), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *ournal of Verbal Learning and Verbal Behavior*, *14*, 575–589. https://doi.org/10.1016/s0022-5371(75)80045-4

Banich, M. T., & Compton, R. J. (2023). *Cognitive neuroscience* (5th). Cambridge University Press.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181

Bregman, A. S., & McAdams, S. (1994). Auditory scene analysis: The perceptual organisation of sound. *The Journal of Acoustical Society of America*, *95*(2), 1177–1178. https://doi.org/10.1121/1.408434

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109. https://doi.org/10.1121/1.1345696

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of Amercia*, *25*, 975–979. https://doi.org/10.1121/1.1907229

Chollet, F. (2021). *Deep learning with python, second edition*. Manning.

Conv Nets: A Modular Perspective - colah's blog. (n.d.). https://colah.github.io/posts/2014-07-Conv-Nets-Modular/

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573. https://doi.org/10.1121/1.2166600

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: From environment to theory of mind. *Neuron*, *58*(3), 306–324. https://doi.org/10.1016/j.neuron.2008.04.017

Corbetta, M., & Shulman, G. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Review Neuroscience*, *3*, 201–215. https://doi.org/10.1038/nrn755

Defays, D. (2023). *L'émergence du sens en intelligence artificielle*. Presses Universitaires de Liège.

Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, *70*(1), 80–90. https://doi.org/10.1037/h0039515

Durlach, N. I., Mason, C. R., Kidd Jr, G., Arbogast, T. L., Colburn, H. S., & Shinn-Cunningham, B. G. (2003). Note on informational masking. *The Journal of the Acoustical Society of America*, *113*(6), 2984–2987. https://doi.org/10.1121/1.1570435

Eimer, M. (2014). The neural basis of attentional control in visual search. *Trends in Cognitive Science*, *18*(10), 528–535. https://doi.org/10.1016/j.tics.2014.05.005

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1016/0364-0213(90)90002-E

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. https://doi.org/10.3758/bf03193146

Feldman, H., & Friston, K. J. (2010). Attention, uncertaincty, and free-energy. *Frontiers in Human Neuroscience*, *4*(215). https://doi.org/10.3389/fnhum.2010.00215

Frigg, R., & Hartmann, S. (2025). Models in science [https://plato.stanford.edu/entries/models-science/].

Fu, D., Weber, C., Yang, G., Kerzel, M., Nan, W., Barros, P., Wu, H., Liu, X., & Wermter, S. (2020). What can computational models learn from human selective attention? a review from an audio-visual unimodal and crossmodal perspective. *Intergrative Neuroscience*, *14*(10). https://doi.org/10.3389/fnint.2020.00010

Geva, G., Warusfel, O., Dubnov, S., Dubnov, T., Amedi, A., & Hel-Or, Y. Binaural sound source localization using a hybrid time and frequency domain model. In: In *Icassp 2024 - 2024 ieee international conference on acoustics, speech and signal processing (icassp)*. 2024. https://doi.org/10.1109/ICASSP48485.2024.10448005

Goldstein, E. B. (2015). *Cognitive psychology : Connecting mind, research and everyday experience, 4th edition*. Cengage Learning.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, *18*(5-6), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Häkkinen, S., & Rinne, T. (2018). Intrinsic, stimulus-driven and task-dependent connectivity in human auditory cortex. *Brain Struct. Funct.*, *223*, 2112–2127. https://doi.org/10.1007/s00429-018-1612-6

Har-shai Yahav, P., Sharaabi, A., & Zion Golumbic, E. (2024). The effect of voice familiarity on attention to speech in a cocktail party scenario. *Cerebral Cortex*, *34*, 1–16. https://doi.org/10.1093/cercor/bhad475

Hexgrad. (2025). Kokoro-82m (revision d8b4fc7). https://doi.org/10.57967/hf/4329

Higgins, N. C., McLaughlin, S. A., Rinne, T., & Stecker, G. C. (2017). Evidence for cue-independent spatial representation in the human auditory cortex during active listening. *Proc. Natl. Acad. Sci. U.S.A.*, *114*, e7602–e7611. https://doi.org/10.1073/pnas.1707522114

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, F., & Dan, Y. (2022). An inferior-superior colliculus circuit controls auditory cue-directed visual spatial attention. *Neuron*, *110*(1), 109–119.e3. https://doi.org/10.1016/j.neuron.2021.10.004

Huckvale, M. (2012). *Audio3d* (Version 1.83). https://www.phon.ucl.ac.uk/resource/audio3d/

Izhikevich, E. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, *14*(6), 1569–1572. https://doi.org/10.1109/TNN.2003.820440

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of english fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252–1263. https://doi.org/10.1121/1.1288413

Kalita, D. (2024, December). What is recurrent neural networks (rnn)? https://www.analyticsvidhya.com/blog/2022/03/a-brief-overview-of-recurrent-neural-networks-rnn/

Kaya, E. M., & Elhailali, M. (2016). Modelling auditory attention. *Philosophical Transactions Royal Society B*, *372*, 20160101. https://doi.org/10.1098/rstb.2016.0101

Kidd, G., Mason, C., Richards, V., Gallun, F., & Durlach, N. (2008). Informational masking. In W. Yost, A. Popper, & R. Fay (Eds.), *Auditory perception of sound sources. springer handbook of auditory research*. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-71305-2_6

Kim, A. J., Grégoire, L., & Anderson, B. A. (2021). Value-biased competition in the auditory system of the brain. *Journal of Cognitive Neuroscience*, *34*(1), 180–191. https://doi.org/10.1162/jocn_a_01785

Lakatos, P., Chen, C. M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron*, *53*(2), 279–292. https://doi.org/10.1016/j.neuron.2006.12.011

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 451–468. https://doi.org/10.1037/0096-1523.21.3.451

Lavie, N., Beck, D., & Konstantinou, N. (2014). Blinded by the load: Attention, awareness and the role of perceptual load. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *369*(1641), 20130205. https://doi.org/10.1098/rstb.2013.0205

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Lenth, R. V. (2025). *Emmeans: Estimated marginal means, aka least-squares means* [R package version 1.11.2-00002]. https://rvlenth.github.io/emmeans/

Lu, K., Xu, Y., Yin, P., Oxenham, A. J., Fritz, J. B., & Shamma, S. A. (2017). Temporal coherence structure rapidly shapes neuronal interactions. *Nature communications*, *8*, 13900. https://doi.org/10.1038/ncomms13900

Luo, Y., & Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(8), 1256–1266. https://doi.org/10.1109/TASLP.2019.2915167

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38. https://doi.org/10.1111/j.1756-8765.2008.01003.x

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. *Information fusion*, *99*, 101869. https://doi.org/10.1016/j.inffus.2023.101869

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, *11*(1), 56–60. https://doi.org/10.1080/17470215908416289

Morrison, C., Kamal, F., Campbell, K., & Taler, V. (2019). Event-related potentials associated with auditory attention capture in younger and older adults. *Neurobiology of Aging*, *77*, 20–25. https://doi.org/10.1016/j.neurobiolaging.2019.01.012

Murphy, S., Spence, C., & Dalton, P. (2017). Auditory perceptual load: A review. *Hearing Research*, *352*. https://doi.org/10.1016/j.heares.2017.02.005

Mwiti, D. (2022). *Deep learning with tensorflow and keras.*

Näätänen, R., Gaillard, A., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, *42*(4), 313–329. https://doi.org/10.1016/0001-6918(78)90006-9

Navon, D. (1989). Attentional selection: Early, late, or neither? *European Journal of Cognitive Psychology*, *1*(1), 47–68. https://doi.org/10.1080/09541448908403071

Norman, D. A. (1968). Toward a theory of memory and attention. *Psychological Review*, *75*(6), 522–536. https://doi.org/https://doi.org/10.1037/h0026699

Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, *83*(4), 579–586. https://doi.org/10.1037/0033-2909.83.4.579

Parande, P. G., & Thomas, T. (2017). A study of the cocktail party problem. *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 1–5. https://doi.org/10.1109/ICECTA.2017.8251979

Parmentier, F. (2014). The cognitive determinants of behavioral distraction by deviant auditory stimuli: A review. *Psychological Research*, *78*, 321–338. https://doi.org/10.1007/s00426-013-0534-4

Patterson, K., & Shewell, C. (1987). Speak and spell: Dissociations and word-class effects. In M. Coltheart, G. Sartori, & R. Job (Eds.), *The cognitive neuropsychology of language*. Lawrence Erlbaum Associates, Inc.

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual review of neuroscience*, *35*, 73–89. https://doi.org/10.1146/annurev-neuro-062111-150525

Pollack, I. (1975). Auditory informational masking. *J. Acoust. Soc. Am.*, *57*(1). https://doi.org/10.1121/1.1995329

Posit team. (2025). *Rstudio: Integrated development environment for r*. Posit Software, PBC. Boston, MA. http://www.posit.co/

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*, 25–42.

Posner, M., & Rothbart, M. (2007). Research on attention networks as a model for the integration of psychological science. *Annual review of psychology*, *58*, 1–23. https://doi.org/10.1146/annurev.psych.58.110405.085516

Ptak, R. (2012). The frontoparietal attention network of the human brain: Action, saliency, and a priority map of the environment. *The Neuroscientist*, *18*(5), 502–215. https://doi.org/10.1177/1073858411409051

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536. https://doi.org/10.1038/323533a0

Saalmann, Y., & Kastner, S. (2014). Neural mechanisms of spatial attention in the visual thalamus. In A. C. Nobre & S. Kastner (Eds.), *The oxford handbook of attention*. Oxford Academic. https://doi.org/10.1093/oxfordhb/9780199675111.013.013

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in neurosciences*, *32*(1), 9–18. https://doi.org/10.1016/j.tins.2008.09.012

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, *45*, 2673–2681. https://doi.org/10.1109/78.650093

Schweickert, R., & Boruff, B. (1986). Short-term memory capacity: Magic number or magic spell? *Journal of experimental psychology. Learning, memory, and cognition*, *12*(3), 419–25. https://doi.org/10.1037/0278-7393.12.3.419

Seaman, M., Levin, J., & Serlin, R. (1991). New developments in pairwise multiple comparisons some powerful and practicable procedures. *Psychological Bulletin*, *110*, 577–586. https://doi.org/10.1037/0033-2909.110.3.577

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in cognitive sciences*, *12*(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Shomstein, S., & Yantis, S. (2006). Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. *Journal of Neuroscience*, *26*(2), 435–439. https://doi.org/10.1523/JNEUROSCI.4408-05.2006

Sieroff, E. (1992). Introduction à l'attention sélective: Définitions et propriétés. *Revue de Neuropsychologie*, *2*(1), 3–27.

Singh, H. (2025, April). Attention mechanism in deep learning. https://www.analyticsvidhya.com/blog/2019/11/comprehensive-guide-attention-mechanism-deep-learning/

Spivey, M. J. (2023). Cognitive science progresses toward interactive frameworks. *Topics in cognitive science*, *15*(2), 219–254. https://doi.org/10.1111/tops.12645

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, *64*(5), 1358–1368. https://doi.org/10.1121/1.382102

Techer, F. (2016). *Impact de la colère sur l'attention, le traitement de l'information et les performances en conduite simulée* [Doctoral dissertation, Université de Nantes].

Treisman, A. M. (1964). Verbal cues, language, and meaning in selective attention. *The American Journal of Psychology*, *77*(2), 206–219. https://doi.org/10.2307/1420127

Understanding LSTM Networks – colah's blog. (n.d.). https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

Vernet, M., Quentin, R., Chanes, L., Mitsumasu, A., & Valero-Cabré, A. (2014). Frontal eye field, where art thou? anatomy, function, and non-invasive manipulation of frontal regions involved in eye movements and associated cognitive operations. *Frontiers in Integrative Neuroscience*, *8*, 66. https://doi.org/10.3389/fnint.2014.00066

Vincent, E., Gribonval, R., & Fevotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(4), 1462–1469. https://doi.org/10.1109/TSA.2005.858005

Wang, D. L., & Terman, D. (1995). 'locally excitatory globally inhibitory oscillator networks'. *IEEE Trans. Neural Networks*, *6*(1), 283–286. https://doi.org/10.1109/72.363423

Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1702–1726. https://doi.org/10.1109/TASLP.2018.2842159

Wang, Z., Wang, P., Liu, K., Wang, P., Fu, Y., Lu, C.-T., Aggarwal, C. C., Pei, J., & Zhou, Y. (2025). A comprehensive survey on data augmentation. https://arxiv.org/abs/2405.09591

Warne, R. (2014). A primer on multivariate analysis of variance (manova) for behavioral scientists. *Practical Assessment, Research, and Evaluation*, *19*(17). https://doi.org/10.7275/sm63-7h70

Wrigley, S., & Brown, G. J. (2004). A computational model of auditory selective attention. *IEEE Transactions on neural networks*, *15*(5), 1151–1163. https://doi.org/10.1109/tnn.2004.832710

Xu, J., Shi, J., Liu, G., Chen, X., & Xu, B. (2018). Modeling Attention and Memory for Auditory Selection in a Cocktail Party Environment. *Proceedings of the ... AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11879

Ylinen, A., Wikman, P., Leminen, M., & Alho, K. (2022). Task-dependent cortical activations during selective attention to audiovisual speech. *Brain Research*, *1775*, 147739. https://doi.org/10.1016/j.brainres.2021.147739

Zion Golumbic, E. M., Poeppel, D., & Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. *Brain and language*, *122*(3), 151–161. https://doi.org/10.1016/j.bandl.2011.12.010

Annexe A : Définition et interprétation d'un spectrogramme

Un spectrogramme correspond à une représentation visuelle en deux dimensions des variations du spectre d'un signal à travers le temps (Mehrish et al., 2023). Le spectrogramme est généralement obtenu via une opération mathématique appelée la transformée de Fourier à court terme (TFCT), ou *Short-Time Fourier Transform* (STFT). Cette opération vise à transformer un signal audio en graphique de fréquences. Pour ce faire, le signal est d'abord scindé en de multiples unités temporelles (une par seconde ou pas milli-seconde). Ensuite, la TFCT décompose chaque segment temporel en ses différentes fréquences constitutives et calcule l'amplitude (puissance) de chacune. Les spectres de fréquences obtenus pour chaque segment sont ensuite alignés chronologiquement pour former une image complète qui représente le signal sonore dans le temps.

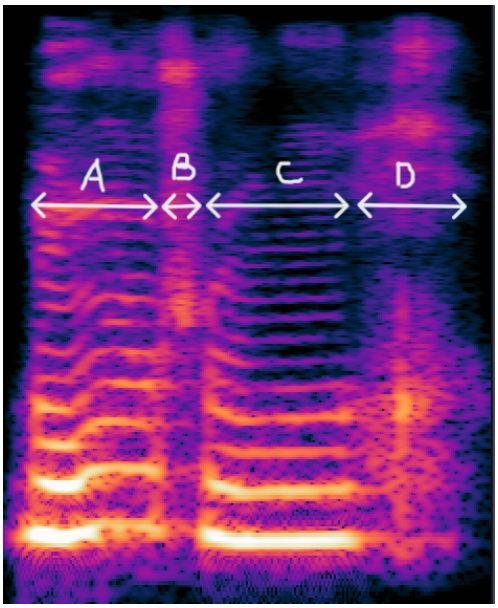


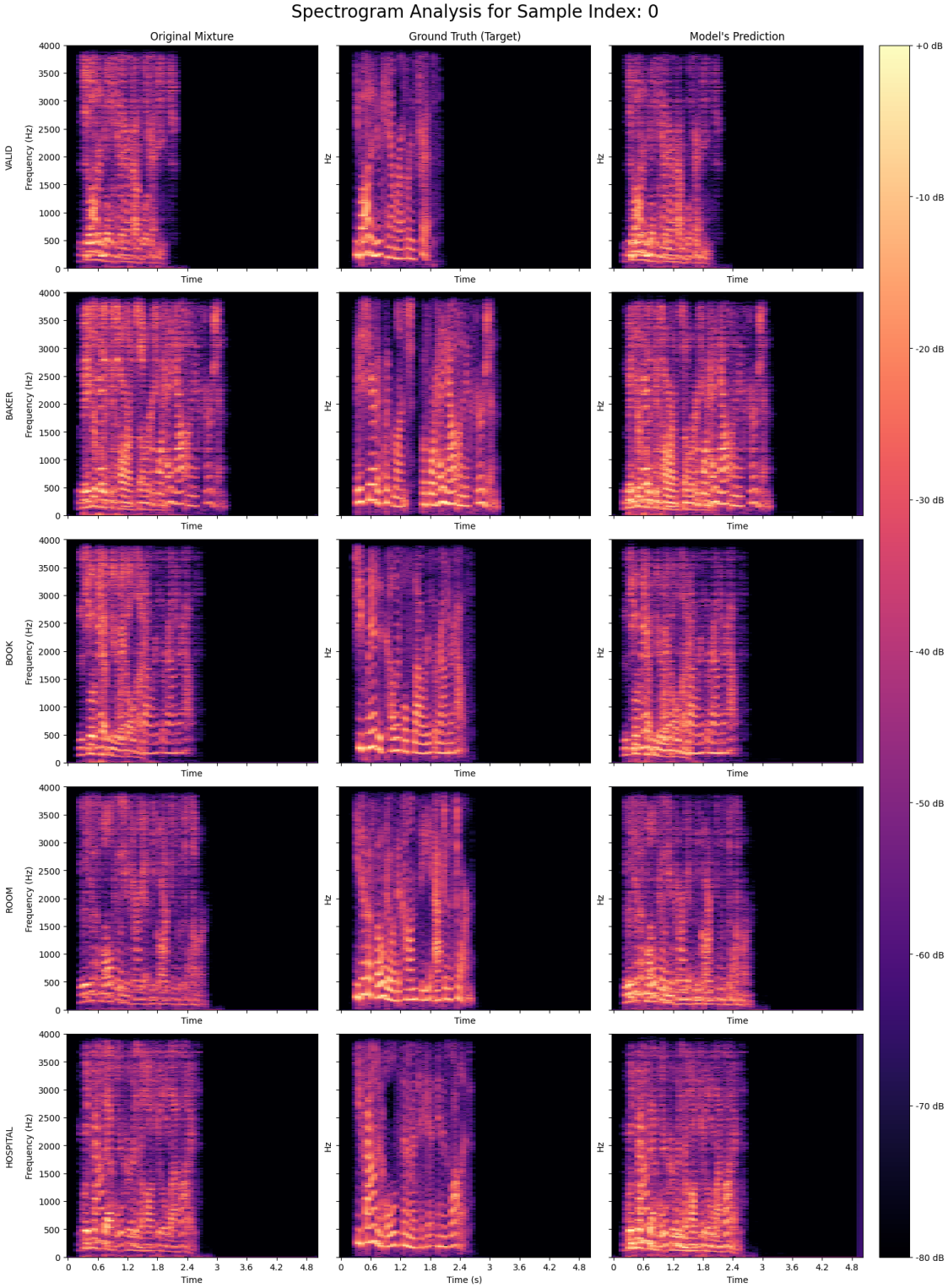
Figure 14: Spectrogramme d'un audio dans lequel le mot "Bonjour" est prononcé. (A) Syllabe « Bon », (B) consonne « j », (C) voyelle « ou », (D) consonne « r ».

L'interprétation d'un spectrogramme repose sur trois dimensions distinctes. Tout d'abord, l'évolution temporelle du signal est représentée via l'axe horizontal (abscisse). Ensuite, l'axe vertical (ordonnée) représente les fréquences mesurées en Hertz (Hz). Une fréquence basse correspond à un son grave, et une fréquence haute correspond à un son aigü. Finalement, les couleurs indiquent l'amplitude (ou la puissance) de chaque fréquence. Sur les spectrogrammes de l'Annexe A, des couleurs plus vives (0dB) correspondent à une intensité plus élevée tandis que les couleurs foncées, voire noires (-80dB), correspondent à une intensité faible. Ainsi, à chaque temps de l'audio, on peut observer quelles fréquences sont plus présentes que les autres, et ainsi déterminer si c'est plus aigu ou plus grave et retracer le signal

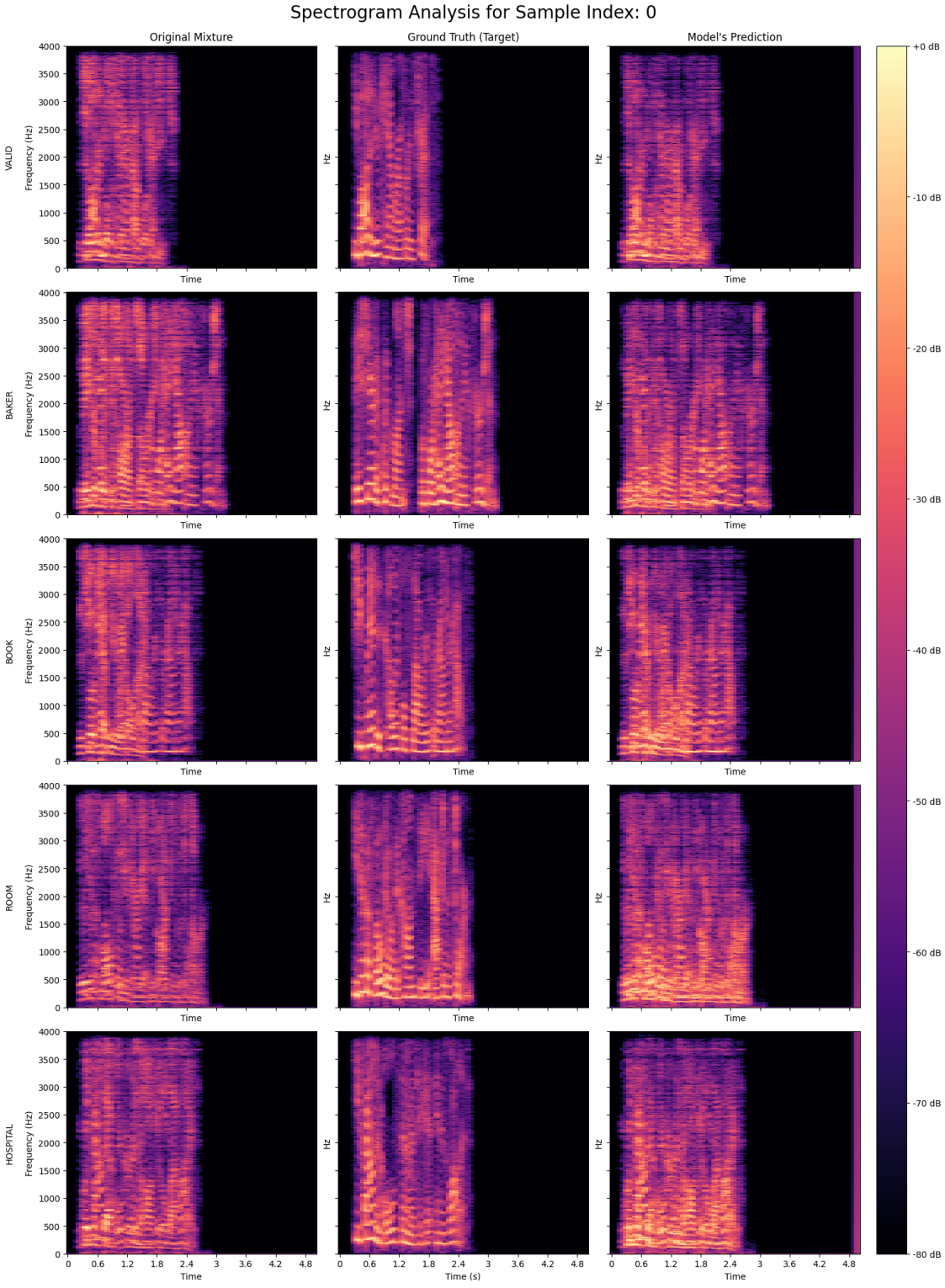
audio.

La Figure 14 illustre l'interprétation d'un spectrogramme dans lequel est prononcé le mot "Bonjour". La zone A correspond au début du mot, "Bon". Les bandes horizontales jaunes et brillantes correspondent au son "on" qui s'étend dans le temps, tandis que le "b" apparaît au tout début de l'image de façon rapide. La zone B correspond à la lettre "j" qui est plus bruyante (on voit que le tracé est plus brouillon et s'étend sur une plus grande bande de fréquences) car c'est une consonne fricative. La zone C correspond au son "ou". Comme au début, il s'agit d'un son long et, cette fois-ci, constant dans le temps. La position de ces bandes dans les fréquences a changé par rapport au son "on" car il ne s'agit pas du même son. Finalement, la zone D correspond à la lettre "r" qui est une consonne fricative entraînant beaucoup de bruit (comme le "j"). On remarque que les fréquences actives (avec une intensité plus élevée, donc des couleurs plus vives) sont plus étendues.

Annexe B : Comparaison visuelle des résultats des modèles BiLSTM et LSTM



De gauche à droite : les spectrogrammes du mix d'entrée, de la source cible et de la prédiction du modèle BiLSTM. Les lignes correspondent aux différentes conditions expérimentales.



De gauche à droite : les spectrogrammes du mix d'entrée, de la source cible et de la prédiction du modèle LSTM. Les lignes correspondent aux différentes conditions expérimentales.