

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

LE RÔLE DES INDICES VISUELS DANS LA PERCEPTION DE LA PAROLE EN CONDITION BRUITÉE :
UNE ÉTUDE PHONÉTIQUE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

MAITRISE EN LINGUISTIQUE

PAR

FRANÇOIS STEFFANN

SEPTEMBRE 2025

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.12-2023). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

Je souhaite remercier et dédier ce travail de recherche à toutes les personnes qui m'ont aidé, accompagné, soutenu, supporté (dans les deux sens du terme) et qui m'ont permis de mener à bien ce mémoire de maîtrise, et en particulier :

Ma blonde Morgane A. pour sa patience,

Ma directrice Lucie Ménard pour sa bienveillance et pour avoir cru en moi,

Mon co-directeur Dave Saint-Amour pour ses conseils avisés,

Anthony H. Poitras Loewen pour son travail et son aide dans la conception de la tâche,

Catherine P. pour ses images,

Ma famille.

AVANT-PROPOS

Ce mémoire de maitrise est l'aboutissement d'un long travail personnel et académique qui a failli ne jamais se terminer. J'ai commencé avec un projet utilisant l'électroencéphalographie pour étudier l'utilisation de la vision dans la perception de la parole, mais ma direction de recherche m'a fait réaliser que l'ampleur du projet que j'envisageais n'était pas adaptée pour un projet de maitrise. Entre-temps, l'épidémie de COVID 19 est arrivée et a forcé l'élaboration d'un nouveau projet plus réaliste et plus restreint se basant sur une observation comportementale de l'intégration audiovisuelle de la parole. L'étude d'une perturbation visuelle et de son impact sur la perception des phonèmes n'ayant jamais été faite dans ce paradigme de perturbation, un travail de recherche seulement comportementale était plus approprié pour explorer ces phénomènes. Par la suite, la réalisation de la tâche a été un défi et je n'y serais jamais parvenu sans l'aide et le soutien de ma directrice de recherche, la Dre Lucie Ménard, pour l'aspect linguistique et phonétique. Je n'aurais pas non plus réussi sans l'aide de mon co-directeur, le Dr. Dave Saint-Amour Ph. D et de son assistant de laboratoire Anthony Poitras Loewen pour l'aspect visuel et la programmation de ma tâche. La collecte, le traitement et l'analyse des données m'ont demandé de nombreuses heures de travail et ont été difficiles d'un point de vue personnel. Je luttais contre moi-même et contre la fatigue. J'ai pensé abandonner à de nombreuses reprises. Finalement, c'est en consultant un thérapeute que j'ai compris que quelque chose n'allait pas et, après un long parcours de soin, on m'a diagnostiqué un TDAH ayant entraîné un burn-out. Je ne pouvais plus lire, encore moins écrire et j'étais en permanence fatigué. La prise en charge médicale et thérapeutique m'a permis peu à peu revenir vers un état psychique compatible avec le travail académique et j'ai pu reprendre la rédaction pour enfin aujourd'hui terminer mon mémoire. J'ai tenu à partager mon parcours avec les lectrices et les lecteurs de ce mémoire, car il fait partie intégrante de mon travail et n'est pas hors du commun dans le milieu académique. Pour terminer, je tiens à souligner que sans le support moral de ma compagne, le soutien sans faille et la confiance de ma directrice de recherche Lucie Ménard, et sans l'accompagnement thérapeutique dont j'ai bénéficié, je ne serais jamais arrivé au bout de ce travail. Un grand merci à elles et à toutes les personnes qui m'ont aidé.

Bonne lecture!

TABLE DES MATIÈRES

REMERCIEMENTS	ii
AVANT-PROPOS	iii
LISTE DES FIGURES.....	vii
LISTE DES TABLEAUX	viii
LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES.....	ix
RÉSUMÉ	x
ABSTRACT	xi
INTRODUCTION	1
CHAPITRE 1 Problématique.....	3
1.1 Quand le bouche-à-oreille se pratique avec les yeux	3
1.2 Objectifs	3
CHAPITRE 2 Cadre théorique	5
2.1 Intégration audiovisuelle de la parole (IAVP)	5
2.1.1 Introduction	5
2.1.2 L'unité de parole dans le contexte de l'IAVP	5
2.2 Production et perception des phonèmes	6
2.2.1 Saillance auditive et visuelle, perception et représentation des traits phonétiques des stimuli ..	6
2.3 Recension des écrits : études comportementales de l'IAVP de 1954 à aujourd'hui	6
2.4 Pertinence de cette recherche	11
2.5 Questions de recherches :	12
CHAPITRE 3 Méthodologie.....	14
3.1 Protocole expérimental.....	14
3.1.1 Choix des stimuli	14
3.1.2 Enregistrement des stimuli	17
3.1.3 Perturbation audio.....	18
3.1.4 Perturbation vidéo	18
3.2 Conception de la tâche en ligne	19
3.3 Population	19
3.4 Passation	20
3.5 Analyse des données.....	21

3.6	Analyse statistique	21
3.6.1	Statistiques descriptives et données aberrantes.....	21
3.6.2	Analyses multiniveaux	22
3.7	Analyse secondaire :	23
3.7.1	Formules du GP des consonnes	23
3.7.2	Formules du GP des voyelles	23
3.7.3	Formules du GP des syllabes.....	24
3.7.4	Taux d'erreur et matrice de confusion	24
CHAPITRE 4	Résultats.....	25
4.1	Effets des filtres de fréquence spatiale sur le score de reconnaissance (Hypothèse 1).....	25
4.2	Effets des filtres de fréquence spatiale sur la perception des phonèmes (Hypothèse 2)	26
4.2.1	Score de reconnaissance par consonnes	26
4.2.2	Score de reconnaissance par contexte vocalique (voyelle)	27
4.2.3	Score de reconnaissance par syllabe	30
4.3	Gain perceptif.....	32
4.3.1	Gain perceptif par consonne	33
4.3.2	Gain perceptif par voyelle.....	34
4.3.3	Gain perceptif par syllabe	35
4.4	Taux d'erreur.....	36
4.4.1	Pourcentage d'erreur par consonne.....	36
4.4.2	Pourcentage d'erreur par voyelle	37
4.4.3	Pourcentage d'erreur par syllabe	38
4.5	Matrices de confusion des consonnes.....	39
4.5.1	Matrice de confusion par fréquence spatiale.....	40
4.6	Matrice de confusion des consonnes par contexte vocalique.....	43
4.6.1	/a/ pour toutes les fréquences spatiales	43
4.6.2	/a/ par filtre de fréquence spatiale.....	44
4.7	Matrice de confusion pour /i/ toutes fréquences confondues.....	45
4.7.1	Matrice de confusion pour /i/ par filtre de fréquence spatiale.....	46
4.8	Matrice de confusion pour /u/ par filtre de fréquence spatiale.....	48
4.8.1	Matrice de confusion pour /u/ par filtre de fréquence spatiale	49
CHAPITRE 5	DISCUSSION	52
5.1	Rappel des objectifs et questions de recherche	52
5.2	Questions de recherche 1	53
5.3	Question de recherche 2.....	54
5.3.1	Interaction entre les filtres de FS et les consonnes	54
5.3.2	Interaction entre les filtres de FS et les voyelles	55
5.3.3	Interaction entre les filtres de FS et les consonnes en contexte vocalique	57
5.3.4	Confusion inter-consonne et taux d'erreur	58

5.4 Résumé des principales observations	58
5.5 Limites de notre recherche	59
CONCLUSION	60
BIBLIOGRAPHIE.....	62
ANNEXE A Questionnaire des antécédents.....	65
ANNEXE B Formulaire de consentement	66

LISTE DES FIGURES

Figure 2.1 Différentes perturbations des fréquences spatiales (A sans perturbation, B perturbation à 2.7 cpf et C perturbation à 5.5 cpf), tirée de Munhall et al. (2004)	9
Figure 2.2 Floutage gaussien de l'image, tiré de Tye-Murray et al., (2016).....	10
Figure 3.1 Illustration de la corrélation entre lieu d'articulation (en bas), ouverture de la bouche et arrondissement (à gauche) et arrondissement (en haut) avec les différentes valeurs de formant acoustique, tiré de Ménard, (2015).....	15
Figure 3.2 Position des articulateurs lors de l'articulation des voyelles cibles	16
Figure 3.3 Lieux d'articulation des consonnes	17
Figure 3.4 Capture d'écran des stimuli pour les 3 niveaux de filtrage passe-bas (20, 12 et 6 cpf de gauche à droite)	18
Figure 4.1 Scores de reconnaissance moyens en fonction du filtre de fréquence spatiale et erreurs standards.	25
Figure 4.2 Scores de reconnaissance moyens par filtres de fréquence spatiale et par consonne et erreurs standards	27
Figure 4.3 Scores de reconnaissance moyens par voyelle et filtre de FS et erreurs standards	28
Figure 4.4 SDR moyens et erreurs standards par filtre de FS et syllabe séparés par contexte vocalique ..	31
Figure 4.5 Gain perceptif par filtre de FS	33
Figure 4.6 Gain perceptif moyen par consonne par rapport au filtre à 6 cpf et erreurs standards	34
Figure 4.7 Gain perceptif moyen par voyelle par rapport au filtre à 6 cpf et erreurs standards.....	35
Figure 4.8 Gain perceptif moyen par syllabe par rapport au filtre à 6 cpf et erreurs standards	36
Figure 4.9 Pourcentage d'erreur par consonne et fréquence spatiale et erreurs standards	37
Figure 4.10 Pourcentage d'erreur par voyelle et fréquence spatiale et erreurs standards.....	38
Figure 4.11 Pourcentage d'erreur par syllabe et fréquence spatiale et erreurs standards	39

LISTE DES TABLEAUX

Tableau 3.1 Profil des personnes participantes	20
Tableau 4.1 Test Post Hoc (Bonferroni) pour la variable voyelle	28
Tableau 4.2 Test Post Hoc pour la l'interaction FS*voyelle	30
Tableau 4.3 Résultats de l'analyse multiniveaux pour les syllabes	31
Tableau 4.4 Test Post-Hoc pour l'interaction syllabe-FS.....	32
Tableau 4.5 Matrice de confusion par consonne pour toutes les fréquences spatiales.....	40
Tableau 4.6 Matrice de confusion par consonne au filtre de FS 6 cpf.....	41
Tableau 4.7 Matrice de confusion par consonne au filtre de FS 12 cpf.....	41
Tableau 4.8 Matrice de confusion par consonne au filtre de FS 20 cpf.....	42
Tableau 4.9 Matrice de confusion pour le contexte vocalique /a/.....	43
Tableau 4.10 Matrice de confusion pour le contexte vocalique /a/ au filtre de FS 6.....	44
Tableau 4.11 Matrice de confusion pour le contexte vocalique /a/ au filtre de FS 12.....	44
Tableau 4.12 Matrice de confusion pour le contexte vocalique /a/ au filtre de FS 20.....	45
Tableau 4.13 Matrice de confusion pour le contexte vocalique /i/.....	45
Tableau 4.14 Matrice de confusion pour le contexte vocalique /i/ au filtre de FS 6.....	46
Tableau 4.15 Matrice de confusion pour le contexte vocalique /i/ au filtre de FS 12.....	46
Tableau 4.16 Matrice de confusion pour le contexte vocalique /i/ au filtre de FS 20.....	47
Tableau 4.17 Matrice de confusion pour le contexte vocalique /u/.....	48
Tableau 4.18 Matrice de confusion pour le contexte vocalique /u/ au filtre de FS 6.....	49
Tableau 4.19 Matrice de confusion pour le contexte vocalique /u/ au filtre de FS 12.....	49
Tableau 4.20 Matrice de confusion pour le contexte vocalique /u/ au filtre de FS 20.....	50
Tableau 4.21 Confusions inter-consonne les plus fréquentes par contexte vocalique	51

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

A : Audio

AV : Audiovisuel

Cpf : Cycle par face

dB : Décibel

FS : Fréquences spatiales

IAPV : Intégration audiovisuelle de la parole

SDR : Scores de reconnaissance

V : Visuel

RÉSUMÉ

L'étude de la perception et de l'intégration audiovisuelle de la parole est un domaine en plein essor. L'intégration audiovisuelle de la parole (IAVP) implique une utilisation concomitante par le cerveau de plusieurs flux sensoriels (p. ex. audio et visuel) pour la perception d'un stimulus langagier. L'objectif de ce mémoire est de décrire la contribution de l'information visuelle dans la perception de la parole lorsque le signal auditif est dégradé. Nous avons demandé à 22 participants-es francophones québécois-es d'identifier des vidéos de 9 syllabes ouvertes composées d'une consonne et d'une voyelle (/a/, /i/ et /u/ et /p/, /t/ et /k/) articulées par une locutrice native du français québécois. L'information visuelle des visages a été manipulée en filtrant les fréquences spatiales. Les stimuli ont été présentés en conditions audiovisuelles dans 3 conditions de filtrage visuel en présence de signal audio bruité. Lors de l'expérience, nous avons procédé à la mesure des scores de reconnaissance des syllabes comme indicateur de performance de l'IAVP. Nos données montrent que dans un contexte audio perturbé, plus la perturbation visuelle est importante et plus la reconnaissance des stimuli est difficile. De plus, la visibilité du lieu d'articulation et le contexte vocalique peuvent faciliter ou nuire à la perception des stimuli. Le contrôle des caractéristiques phonétiques et linguistiques des stimuli nous a permis de mettre en évidence les informations visuelles afférentes aux phonèmes et influençant l'IAVP. Ce travail de recherche apporte de nouvelles données sur le rôle de la vision dans la perception de la parole et caractérise phonétiquement l'information visuelle utilisée dans la perception et la représentation des unités de parole par le cerveau.

Mots clés : phonétique, intégration, audiovisuelle, perception, perturbation

ABSTRACT

The study of speech perception and audiovisual integration is a rapidly growing field. Audiovisual Integration of Speech (AVIS) refers to the brain's simultaneous use of multiple sensory streams (e.g., auditory and visual) for the perception of a speech stimulus. The purpose of this research is to describe the role of visual cues in speech perception when the auditory and visual information are altered. For our task, we asked 22 French-speaking participants (from Quebec) to identify videos of 9 open syllables composed of a consonant and a vowel (/p/, /t/, /k/ and /a/, /i/, /u/) produced by a native speaker of Quebec French. Visual disruptions were created by filtering spatial frequencies (low-pass filters). Stimuli were presented audiovisually in three visual filtering conditions alongside noisy auditory signals. During the experiment, we measured recognition scores (RS) as a performance indicator of AVIS. Our data show that in a degraded auditory context, the greater the visual disruption, the more challenging the recognition of stimuli becomes. Furthermore, the visibility of the place of articulation and the vocalic context can either facilitate or hinder the perception of stimuli. The control of the phonetic and linguistic characteristics of the stimuli allowed us to identify the visual information related to phonemes that influences AVSP. This research provides new data on the role of vision in speech perception and phonetically characterizes the visual information used in the perception and representation of speech units by the brain.

Keywords: phonetics, audiovisual, integration, perception, perturbation

INTRODUCTION

La perception est, d'une manière globale, le phénomène qui traduit les sensations du monde extérieur par l'intermédiaire des sens en un ensemble d'influx nerveux électriques traités par le cerveau pour créer notre représentation du monde. Le plus souvent, nos différents sens, comme la vision, l'ouïe ou le toucher, fonctionnent ensemble pour interpréter les stimuli du monde extérieur. C'est ce que nous appelons l'intégration multisensorielle (Calvert et Thesen, 2004 ; Massaro et Cohen, 1995 ; Möttönen et Sams, 2008). Lors de l'intégration multisensorielle des stimuli de la parole, le cerveau met en relation les informations perçues par les différents sens afin d'améliorer le traitement des stimuli (Callan *et al.*, 2004 ; Calvert et Thesen, 2004 ; Macleod et Summerfield, 1987 ; Robert-Ribes *et al.*, 1998). Pour la présente recherche, nous avons travaillé sur un type d'intégration multisensorielle : l'intégration audiovisuelle de la parole (IAVP). L'IAVP s'illustre notamment lorsque nous écoutons une personne parler dans un environnement bruyant. Dans cette situation, notre cerveau utilise l'information auditive et visuelle afin de mieux percevoir les sons. Cet exemple illustre un traitement cérébral de plus haut niveau de la parole : la vision entre en jeu pour aider l'audition et ainsi mieux traiter la parole perçue. L'enjeu pour les scientifiques est de savoir comment se manifeste cette intégration multisensorielle d'un point de vue comportemental et fonctionnel. Dans ce travail, nous étudions le traitement cérébral de la parole perçue auditivement et visuellement lorsque le son et la vision sont dégradés. Au début du processus de recherche, nous souhaitions travailler en laboratoire de phonétique avec les personnes participantes et leur présenter notre test de perception, mais l'arrivée du COVID a rendu cela impossible. Nous avons dû mettre au point un protocole expérimental permettant de récolter nos données à distance. Ce revirement de situation n'a pas eu que des côtés négatifs puisqu'il nous a permis de mettre au point un protocole efficace et relativement léger pouvant être réalisé chez soi par les personnes participantes. Notre protocole de recherche nous a donc permis d'obtenir des données comportementales sur l'intégration audiovisuelle de la parole en milieu bruyant. Pour ce faire, nous avons enregistré une locutrice native du français québécois prononçant des syllabes cibles. Nous avons ensuite créé de courtes séquences vidéo pour lesquelles nous avons bruité l'audio et filtré les fréquences spatiales des images. Puis, nous avons créé notre tâche de perception sur le logiciel Psychopy (Open Science Tools Ltd., 2022b) et nous l'avons mis en ligne sur la plateforme Pavlovia (Open Science Tools Ltd., 2022a). Les personnes participantes devaient simplement cliquer sur un lien qui leur était envoyé et commencer la tâche. Nous avons ainsi pu obtenir les données sur lesquelles nous avons travaillé et nos résultats montrent que l'utilisation de la vision dans la perception de la parole est un processus complexe qui varie en fonction des phonèmes et de leur contexte phonétique.

Nos résultats sont les premiers de ce genre dans la littérature scientifique et ils ouvrent la voie vers des investigations plus poussées afin de mieux caractériser l'intégration multimodale de la parole.

CHAPITRE 1 Problématique

1.1 Quand le bouche-à-oreille se pratique avec les yeux

Dans ce travail de recherche, nous étudions la perception et l'IAVP dans un paradigme de double perturbation : **le son et la vidéo sont perturbés**. Nous avons choisi ce type de tâche, car elle permet de forcer l'utilisation de la vision avec une perturbation audio fixe et d'étudier la réaction à une variation de l'information visuelle en manipulant les fréquences spatiales par filtrage. On sait déjà qu'une perturbation du retour auditif, visuel et même articuloire (pathologique ou artificielle), lors de l'acquisition ou lors de la perception de la parole, entraîne une modification de la représentation mentale des phonèmes et influence leur production temporairement ou de manière permanente dans les cas pathologiques (Trudeau-Fisette, 2015). Bien que l'on comprenne de mieux en mieux l'intégration audiovisuelle de la parole, peu de travaux en linguistique se sont intéressés à la perception visuelle des phonèmes (Campbell et Massaro, 1997). Nous avons donc réalisé une mesure comportementale de l'impact de différents niveaux de perturbation visuelle sur l'IAVP en condition audio bruitée fixe. Notre travail est à la croisée de la phonétique, de la psychologie de la perception (vision) et de l'audiologie. Les connaissances sur l'intégration visuelle des traits phonétiques de la parole sont limitées. Notre tâche de perception permet d'étudier, grâce à une double perturbation audio et visuelle, le rôle de la vision dans la représentation multimodale des percepts de parole. Ceci n'a encore jamais été fait à notre connaissance et nos résultats permettent de mieux comprendre l'utilisation de la vision dans la compréhension, la représentation et la production de la parole. Notre recherche s'appuie sur la théorie d'une perception de la parole multimodale telle que décrite par (Schwartz *et al.*, 2012) dans son modèle appelé PACT (Perception for Action Control Theory). Nous reviendrons plus loin sur cette approche.

1.2 Objectifs

Nous voulons étudier l'impact d'une perturbation visuelle sur la perception audiovisuelle des traits phonétiques que sont le lieu d'articulation et l'aperture. Les travaux sur l'IAVP montrent l'impact négatif d'une perturbation visuelle sur les scores de perception en condition doublement perturbée audio et visuelle (Dickinson et Taylor, 2011 ; Legault *et al.*, 2010). Ces travaux, du fait de leur objectif de recherche, ne contrôlent pas avec précision certains paramètres influençant l'IAVP comme les effets de sens et de fréquences des mots utilisés dans les stimuli ainsi que les fréquences spatiales des stimuli visuels (Connine *et al.*, 1993 ; Kutas et Hillyard, 1980 ; Liben-Nowell *et al.*, 2019 ; Osterhout et Holcomb, 1992). Nous

proposons donc un protocole expérimental basé sur les travaux en phonétique et en psychologie de la perception pour mieux contrôler ces paramètres. En perturbant l'information visuelle en présence d'un signal audio bruité, nous pouvons étudier l'IAVP au niveau du phonème et de la syllabe en analysant la résistance des traits articulatoires et acoustiques à la perturbation visuelle. Il devient ainsi possible de décomposer et d'étudier spécifiquement ces paramètres articulatoires et acoustiques et leur utilisation dans le processus d'IAVP.

CHAPITRE 2 Cadre théorique

2.1 Intégration audiovisuelle de la parole (IAVP)

2.1.1 Introduction

Le décodage et la représentation mentale des phonèmes (unités acoustiques) sont multimodaux (Schwartz *et al.*, 2012). L'association de ceux-ci avec les percepts langagiers permet de comprendre les syllabes, les mots et enfin les phrases. En condition normale, le système auditif humain (de l'oreille externe au cortex auditif) est suffisant pour comprendre la parole ; en témoigne notre capacité à entendre et comprendre un interlocuteur de notre langue sans le voir. Cependant, un signal auditif appauvri ou perturbé impacte ces processus. De la même façon, un déficit sensoriel acquis ou inné comme la surdité ou la cécité aura une influence sur l'acquisition de la parole et sa production (Ménard, 2015). Ainsi, quand le signal audio est perturbé (bruit ambiant dans un restaurant par exemple), la vision devient nécessaire à la perception. Les informations acoustiques et visuelles sont alors intégrées par le cerveau pour mieux décoder le message (Benoit *et al.*, 1994). En résumé, l'IAVP combine la vision et l'audition afin d'améliorer le décodage des percepts de parole.

2.1.2 L'unité de parole dans le contexte de l'IAVP

La théorie de la perception pour le contrôle de l'action ou PACT (Schwartz *et al.*, 2012) décrit ces unités de parole à travers leurs aspects neurologiques, articulatoires et acoustiques et définit une unité de parole complexe, impliquant à la fois une représentation interne motrice du phonème et sa réalisation perceptible. On parle dès lors d'**unité perceptuo-motrice de parole ou percept**. L'unité de parole est donc un percept multimodal complexe intégrant les représentations articulatoires, somesthésiques, acoustiques et visuelles (Schwartz *et al.*, 2012). Cette définition est la plus adéquate, car elle permet une approche complète neurologique et physique de l'unité de parole, à contrepied d'une définition seulement acoustique ou articulatoire du phonème. La PACT définit donc l'unité de parole comme représentation mentale complexe multimodale dont la production tient compte des contraintes de perception. Étudier la perception renseigne donc sur la production et inversement.

2.2 Production et perception des phonèmes

2.2.1 Saillance auditive et visuelle, perception et représentation des traits phonétiques des stimuli

Tous les phonèmes ne sont pas affectés de la même manière par une dégradation sensorielle (Ménard, 2015). En modalité audiovisuelle, dans un contexte de perturbation audio seule, Robert-Ribes *et al.* (1998) observent que la perturbation audio n'affecte pas les voyelles de la même façon. Les traits phonétiques de ces dernières ont une résistance spécifique à la modalité de perturbation. On parle de saillance articulatoire des paramètres acoustiques et phonétiques. Dans le cas de la perception visuelle de la parole, la saillance articulatoire des phonèmes est liée à l'utilisation plus ou moins importante des lèvres et de la mâchoire lors de leur articulation. Les auteurs remarquent que les saillances visuelle et auditive de /a, i, y/ sont complémentaires. Ainsi, en modalité audio, le /a/ sera mieux perçu que le /i/, qui sera mieux perçu que le /y/. En modalité visuelle, on obtiendra les résultats inverses (Robert-Ribes *et al.*, 1998). Plusieurs questions se posent : Comment sont impactés les procédés d'IAVP par une perturbation visuelle ? Quelle est la saillance visuelle des traits articulatoires des phonèmes et quels traits articulatoires ou acoustiques résistent le mieux à cette perturbation ? C'est ce que nous étudions dans ce travail de recherche.

2.3 Recension des écrits : études comportementales de l'IAVP de 1954 à aujourd'hui

Sumby et Pollack (1954) sont parmi les premiers à avoir étudié l'implication de la vision dans la perception de la parole. Dans leur tâche, un locuteur natif de l'anglais est entraîné à lire une liste de mots choisis parmi 256 mots bisyllabiques. Le signal audio est perturbé en temps réel. La difficulté de perception est augmentée en baissant le niveau de la voix et en gardant la perturbation au même niveau. Les participants-es lisent la liste de mots avant l'expérience. Une moitié des participants-es voit la personne qui prononce les mots de la liste, l'autre moitié ne la voit pas. On observe une différence significative entre les scores de reconnaissance des deux modalités. Les personnes ayant accès aux deux modalités (AV) ont de meilleurs scores de reconnaissance même quand le signal audio est très perturbé et les scores de réussite ne descendent pas en-dessous de 40%. Pour la modalité audio seule, on observe des courbes sigmoïdales qui finissent en dessous de 20% de reconnaissance pour le niveau de perturbation maximum. Ce travail de recherche est l'un des premiers à montrer l'importance de la vision dans la perception de la parole et est encore cité dans les travaux récents. Bien qu'indéniablement pertinent, le protocole expérimental est limité par l'utilisation de mots (effet de sens ou de fréquence que nous traiterons à la fin de cette section) et par la production en direct des stimuli par un acteur (variation entre chaque stimulus et d'un participant à l'autre), ceci étant lié aux contraintes technologiques de l'époque.

L'étude de Binnie *et al.* (1974) traite de la saillance sensorielle des traits phonétiques des consonnes de l'anglais pour différents niveaux de bruit audio. L'utilisation de syllabes CV avec la voyelle /a/ permet, contrairement à l'étude précédente, de contrôler les effets de sens et fréquences liés à l'utilisation de mots comme stimuli et permet d'étudier avec précision l'impact des traits phonétiques de chaque consonne employée. On observe dans cette étude que le lieu d'articulation est le trait phonétique le plus sensible à la perturbation audio. Par ailleurs, dans la modalité audiovisuelle, on observe une amélioration des scores de reconnaissance quand les phonèmes diffèrent dans leur lieu d'articulation. Plus la perturbation audio est forte et plus la compensation de la vision est importante (écart de score de reconnaissance entre les conditions). Dans la condition visuelle seule, les personnes participantes distinguent presque sans erreur 5 lieux d'articulation. Ces phonèmes visuels sont appelés visèmes. Le lieu d'articulation 'antérieur' est le visème qui résiste le mieux à la perturbation visuelle¹.

L'étude de Benoit *et al.* (1994) met en évidence le rôle de la vision dans la perception des phonèmes du français et confirme les résultats de Sumby et Pollack (1954). Plus spécifiquement, les chercheurs-es étudient l'impact des contextes vocalique [a, i, y] sur la perception des consonnes [b, v, z, ʀ, ʒ, l] et inversement, dans de faux mots respectant la phonologie du français. Le choix des voyelles est justifié par les contrastes forts les opposant (hauteur et arrondissement). On observe dans ce travail que le contexte vocalique le plus facile à percevoir en modalité audio seule est [a], suivi de [i], à son tour suivi de [y]. En modalité AV, on obtient la hiérarchie de perception suivante : [y] > [a] > [i] ([y] est assimilable au [u]). L'arrondissement qui est visible devient le trait phonétique le mieux perçu en contexte AV (Benoit *et al.*, 1994). La méthodologie utilisée par ces auteurs a inspiré la création de notre tâche dans le choix des voyelles et est en partie reprise dans l'étude suivante.

La tâche de Robert-Ribes *et al.* (1998) étudie l'impact d'une perturbation audio sur la perception A et AV de voyelles orales du français et fournit des résultats identiques à ceux de Benoît *et al.* (1994). Les voyelles utilisées dans la tâche de perception permettent de tester l'impact de la perturbation audio sur les traits phonétiques des stimuli. De plus, l'utilisation des voyelles plutôt que des mots ou des phrases suppriment les éventuels effets de sens et fréquence que ces derniers peuvent induire. Les participants-es entendent et/ou voient les voyelles articulées et doivent appuyer sur les touches d'un clavier pour choisir la voyelle entendue. Les stimuli sont perturbés à différents niveaux de bruit blanc gaussien. Comme pour

¹ Les traits phonétiques impactés par la perturbation et leur résistance à celle-ci nous indiquent quels phonèmes employer dans notre tâche, nous y reviendrons quand nous traiterons de notre protocole expérimental.

l'expérience de Sumby et Pollack (1954), plus le niveau de bruit augmente, plus les scores de reconnaissance diminuent. La condition audiovisuelle correspond toujours à de meilleurs scores de reconnaissance que la condition audio seule. Les stimuli en condition visuelle seule permettent d'obtenir un score de reconnaissance de 45%. Même sans l'audition, la vision permet donc d'extraire un certain nombre d'informations sur les phonèmes, les traits les mieux perçus visuellement étant : arrondissement > aperture > antériorité. En condition audio seule (A), les traits les mieux perçus dans le bruit sont : aperture > antériorité > arrondissement. Enfin, l'étude confirme que la vision améliore toujours les scores de reconnaissance. En condition audiovisuelle (AV), les traits les mieux perçus dans le bruit sont : arrondissement > aperture > antériorité. L'ordre n'est pas le même qu'en condition A et tous les traits phonétiques n'apportent pas le même avantage perceptif. En conclusion, cette étude démontre qu'il y a une synergie perceptive entre l'audition et la vision améliorant la perception globale notamment pour l'arrondissement.

Les deux études précédentes nous ont mené à nous interroger sur l'impact d'une perturbation visuelle sur la perception AV de la parole. Si une perturbation audio change la facilité de perception d'un phonème, qu'en est-il d'une perturbation visuelle? Campbell et Massaro (1997) traitent de l'impact d'une perturbation visuelle sur la capacité à lire la parole sur les lèvres, une condition similaire à la condition visuelle de Robert-Ribes *et al.* (1998). Cette expérience étudie l'impact d'une perturbation visuelle sur la transmission des informations visuelles de la parole à l'aide de visages de synthèse dont l'image est filtrée en termes de fréquence spatiale (*spatial quantization*). Elle montre que plus les fréquences spatiales (en cpf, *cycle per face*) disponibles sont basses (plus gros détails visuels), plus la reconnaissance des syllabes visuelles est difficile. De plus, certains visèmes sont plus résistants que d'autres à la dégradation visuelle. Cette étude montre qu'une perturbation équivalente à un filtrage des fréquences spatiales entraîne une diminution des scores de reconnaissance des visèmes. Elle montre aussi qu'une grande partie des informations visuelles se situe dans les fréquences spatiales plus basses. En effet, il y a plus de perte d'information, c'est-à-dire de baisse des scores de reconnaissance, entre 16 et 4 cpf qu'entre 323 et 16 cpf².

² Ces informations sont cruciales pour la mise au point de notre protocole expérimentale, car elles nous permettront de choisir quelles tranches/spectres de fréquences spatiales sont les plus pertinentes pour la mise au point de notre perturbation visuelle.

Munhall *et al.* (2004) ont réalisé une étude dans des conditions qui se rapprochent de celles de notre travail de recherche. Les auteurs-es étudient l'effet d'une perturbation visuelle lorsque le signal auditif est aussi perturbé. Les participant-es doivent identifier des mots-clefs contenus dans des phrases. Les phrases sont prononcées par un locuteur de la langue et enregistrées en vidéo. Les images sont ensuite perturbées par ordinateur et une perturbation audio est ajoutée afin de rendre plus difficile la tâche et d'éviter les effets de plafond. La Figure 2.1 illustre les différents niveaux de perturbation employés. Nous voyons par exemple que le filtrage des fréquences hautes entraîne une perte de l'information visuelle fine (B). Deux types de perturbations visuelles sont employées, un filtre par bande de fréquences spatiales dans les deux premières tâches et un filtre passe-bas dans la dernière tâche.



Figure 2.1 Différentes perturbations des fréquences spatiales (A sans perturbation, B perturbation à 2.7 cpf et C perturbation à 5.5 cpf), tirée de Munhall et al. (2004)

Cette méthode met en évidence la présence d'informations perceptives différentes pour chaque bande de fréquence spatiale. Cela suggère que l'information présente dans les fréquences basses n'est pas redondante avec celle présente dans les fréquences plus hautes, d'où la baisse des scores dans les fréquences plus hautes (perte de l'information des fréquences plus basses). Ainsi, en additionnant les fréquences à l'aide d'un filtre passe-bas, on obtient une courbe croissante avec l'augmentation des fréquences spatiales, comme dans l'expérience de Campbell et Massaro (1997). Les résultats obtenus ici permettent de comprendre la répartition de l'information dans le spectre des fréquences spatiales utilisé pour la perception AV de la parole. Dans le même ordre d'idée, mais en audiologie, les travaux de recherche de Dickinson et Taylor, (2011) et Legault *et al.*, (2010) étudient l'effet d'une perturbation visuelle sur la perception de la parole en condition audio bruitée. Les chercheurs-ses utilisent une

perturbation audio dans toutes les conditions (bruit rose ou blanc) et une perturbation optique (une lentille convexe ou un feuillet d'occlusion) simulant différentes pertes d'acuité visuelle. Les chercheurs-es comparent ces perturbations à la condition audio seule et à la condition audiovisuelle sans perturbation. Les personnes participantes doivent reconnaître trois mots cibles dans les phrases entendues. Elles ont pris connaissance des phrases auparavant. Les résultats montrent que peu importe la perturbation, les conditions audiovisuelles correspondent toujours à des scores de reconnaissances plus élevés ($p < 0.001$) que pour la condition audio seule. De plus, la perturbation visuelle la plus forte obtient des scores de reconnaissance moins élevés ($p < 0.001$) que la condition audiovisuelle sans perturbation. Les résultats de ces études sont en accord avec les résultats d'études mentionnées ici (Campbell et Massaro, 1997 ; Munhall *et al.*, 2004) qui observent une diminution des scores de reconnaissance avec le filtrage de l'information visuelle. Cependant, l'utilisation de perturbations optiques (floutage), bien qu'assimilables à un filtrage des fréquences spatiales hautes, ne permet pas de contrôler les paramètres de l'information visuelle avec précision. Enfin, l'utilisation de phrases et de mots ne permet pas de contrôler d'éventuels effets de sens ou de fréquences que peuvent induire ces premiers. Nous reviendrons sur ces points dans la section traitant de notre protocole expérimental.

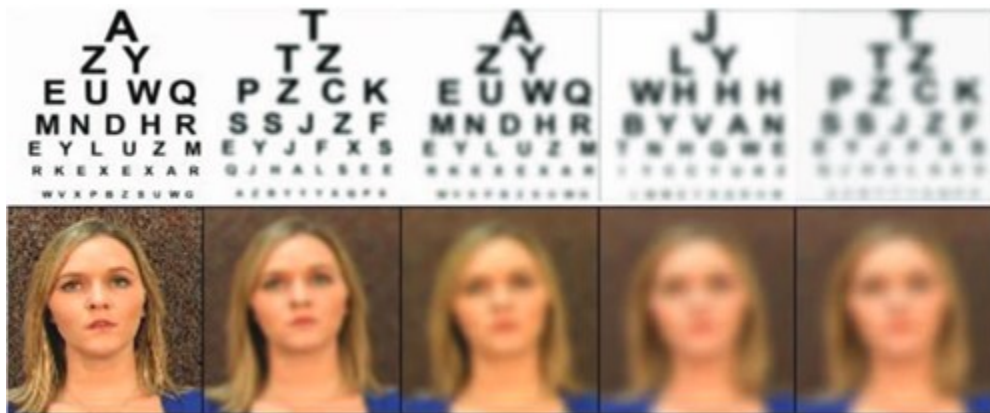


Figure 2.2 Floutage gaussien de l'image, tiré de Tye-Murray *et al.*, (2016)

Une étude sur les capacités de lecture labiale et de reconnaissance audiovisuelle au cours de la vie utilise une méthode similaire aux deux études précédentes (Tye-Murray *et al.*, 2016). Dans ce cas, la perturbation visuelle est réalisée par ordinateur, (floutage gaussien³) (Figure 2.2). On obtient ici aussi une diminution des scores de reconnaissance quand la perturbation augmente. On notera que les stimuli utilisés sont aussi

³ Le floutage Gaussien atténue les fréquences spatiales hautes et s'apparente à un filtre passe-bas (Young, 2020)

des mots-clefs à reconnaître dans des phrases, ce qui limite la portée des résultats obtenus si on considère les effets de sens et fréquences qu'ils peuvent induire. En effet, l'utilisation, dans certains des travaux mentionnés, de mots ou de phrases comme stimuli peut entraîner des effets de sens, de fréquence ou encore d'analyse syntaxique (Connine *et al.*, 1993 ; Kutas et Hillyard, 1980 ; Liben-Nowell *et al.*, 2019). Ces éléments, contrairement aux éléments phonétiques simples comme les phonèmes ou les syllabes, induisent un processus de traitement cognitif de plus haut niveau, lequel arrive de facto après les traitements de plus bas niveau phonétiques et phonologiques et allonge le temps de traitement (Connine *et al.*, 1993 ; Osterhout et Holcomb, 1992). Pour étudier le traitement phonétique lié à l'IAVP et son impact sur les temps de réponse, il faut donc le séparer des traitements de plus haut niveau en utilisant des stimuli de 'bas niveau' comme des syllabes ou des phonèmes n'induisant pas les traitements de plus haut niveau (sens, fréquence, syntaxe ou toute autre stratégie d'interprétation de la parole). Nous prenons ces éléments en compte dans la construction de nos stimuli.

Pour conclure cette section, nous voyons que la modalité audiovisuelle, quand le signal audio est perturbé, entraîne toujours une amélioration des scores de reconnaissance par rapport aux modalités audio et visuelle seules. Quand on ajoute une perturbation visuelle en condition AV ou V, les scores de reconnaissance diminuent systématiquement. Nous proposons donc d'étudier, du point de vue phonétique, l'effet d'une perturbation visuelle contrôlée sur l'IAVP, en nous focalisant sur les phonèmes et leurs traits et non sur des mots inclus dans des phrases.

2.4 Pertinence de cette recherche

Considérant l'état des connaissances sur l'IAVP et notamment le concept de percept produit et perçu proposé par la PACT (Schwartz *et al.*, 2012), nous proposons d'approcher l'IAVP avec le regard de la linguistique et de la phonétique. Pour comprendre l'impact d'une perturbation visuelle sur l'intégration de la parole et pour étudier le traitement phonétique lié à l'IAVP, il faut se focaliser sur les traitements phonétiques dits de 'bas niveau' en utilisant des stimuli de 'bas niveau' comme des syllabes. L'utilisation de syllabes permet en outre d'évaluer la saillance perceptive de combinaisons de traits phonétiques, les suites de consonnes et voyelles différentes impliquant une dynamique visuelle différente. Les études en phonétique qui utilisent le paradigme de perturbation audio utilisent cette même méthodologie (Benoît Christian *et al.*, 1994 ; Binnie *et al.*, 1974 ; Robert-Ribes *et al.*, 1998 ; Sumbly et Pollack, 1954). Notre apport à la recherche consiste à étendre cette méthode expérimentale aux perturbations visuelles. Cela permet de faire percevoir les phonèmes dans un contexte plus isolé et d'observer, d'un point de vue

comportemental, la perception audiovisuelle. Nous espérons compléter les données obtenues par les travaux de recherches précédents qui mettent en évidence la complémentarité de l'audition et de la vision dans le décodage de la parole et la hiérarchie perceptive des traits phonétiques dans les deux modalités (Benoît Christian *et al.*, 1994 ; Robert-Ribes *et al.*, 1998). Nous testons ainsi la résistance ou la susceptibilité des traits phonétiques à la perturbation visuelle. Cette approche nous permet de mieux comprendre la saillance sensorielle des traits phonétiques, c'est-à-dire, la saillance auditive et la saillance visuelle des phonèmes lors de la perception de la parole. En résumé, notre travail de recherche permet de mieux comprendre l'intégration audiovisuelle lors de la perception de la parole.

2.5 Questions de recherches :

Q.1 Quel est l'impact d'une perturbation visuelle sur la perception de la parole en condition audio bruitée/perturbée ? (Binnie *et al.*, 1974 ; Campbell et Massaro, 1997 ; Dickinson et Taylor, 2011 ; Legault *et al.*, 2010 ; Munhall *et al.*, 2004 ; Robert-Ribes *et al.*, 1998 ; Schwartz *et al.*, 2012 ; Sumbly et Pollack, 1954 ; Tye-Murray *et al.*, 2016)

Q.2 Y-a-t-il des traits phonétiques qui résistent mieux à la perturbation visuelle que d'autres : lieu d'articulation, type de phonème (voyelles, consonnes) ou encore mode d'articulation des voyelles (arrondissement)? (Binnie *et al.*, 1974 ; Robert-Ribes *et al.*, 1998 ; Schwartz *et al.*, 2012 ; Sumbly et Pollack, 1954)

Pour répondre à nos questions de recherche, nous proposons un protocole expérimental qui utilisera des syllabes plutôt que des mots, et une perturbation visuelle contrôlée grâce à l'utilisation de filtres de fréquences spatiales passe-bas (Campbell et Massaro, 1997 ; Munhall *et al.*, 2004). Nous utiliserons aussi une perturbation audio fixe par du bruit rose qui permettra d'éviter les effets plafonds en augmentant la difficulté de perception auditive (Campbell et Massaro, 1997 ; Munhall *et al.*, 2004 ; Robert-Ribes *et al.*, 1998 ; Sumbly et Pollack, 1954). Nous testerons la perception de la saillance visuelle des consonnes /p/, /t/ et /k/ dans trois contextes vocaliques /Ca/, /Ci/ et /Cu/ de locutrices et locuteurs du français québécois (Benoit *et al.*, 1994 ; Binnie *et al.*, 1974 ; Robert-Ribes *et al.*, 1998).

Considérant nos questions de recherche, sur la base des travaux présentés dans les sections précédentes, nos hypothèses sont les suivantes :

- **H.1** *Plus le filtre de FS sera bas (et donc plus la perturbation sera importante), plus les scores de reconnaissance seront faibles.*
- **H.2** *La perturbation aura des effets différents selon le domaine et le type de phonème sur lequel elle est appliquée (syllabe vs phonème, et voyelle vs consonne). La saillance articulatoire des traits phonétiques influencera les scores de reconnaissance (Binnie et al., 1974 ; Macleod et Summerfield, 1987 ; Robert-Ribes et al., 1998)*

CHAPITRE 3 Méthodologie

3.1 Protocole expérimental

3.1.1 Choix des stimuli

Nous avons réalisé une tâche de perception de la parole en conditions audio et vidéo perturbées. Les personnes réalisant la tâche devront discriminer les consonnes dans 3 contextes vocaliques (3 syllabes ouvertes de 2 phonèmes) :

- Contexte vocalique 1 : /pa/, /ta/ et /ka/
- Contexte vocalique 2 : /pi/, /ti/ et /ki/
- Contexte vocalique 3 : /pu/, /tu/ et /ku/

Le choix des phonèmes de nos stimuli a été déterminé par plusieurs facteurs phonétiques et articulatoires. Pour commencer, nous avons choisi les trois voyelles cardinales [a, i, u], car elles sont les plus courantes dans les langues du monde, elles sont orales et offrent 3 dimensions de contraste acoustique et articulatoire (aperture, antériorité et arrondissement) (Maddieson, 1984). Pour explorer la perturbation visuelle, il nous a semblé pertinent d'utiliser les voyelles avec les plus grandes distances acoustiques et articulatoires entre elles, afin de maximiser les contrastes auditifs (Maddieson, 1984). Cela évite d'éventuelles confusions observées lors de la perception de voyelles articulées proches l'une de l'autre (Robert-Ribes *et al.*, 1998). La Figure 3.1 illustre ce fait, on y voit la corrélation entre les paramètres articulatoires et les valeurs de formant F1, F2 et F3. Deux voyelles comme le [u] et le [o] partageront plus de traits articulatoires que le [a] et le [u], car elles sont toutes deux postérieures et arrondies. Cela pourra entraîner plus de confusion lors de leur perception dans une tâche (Robert-Ribes *et al.*, 1998) par rapport à 2 voyelles qui s'opposent par leurs lieux d'articulation (l'une est antérieure et l'autre est postérieure, par exemple). En utilisant une distance similaire entre les voyelles, nous limitons ce type de confusions, ce qui facilitera l'interprétation de nos résultats.

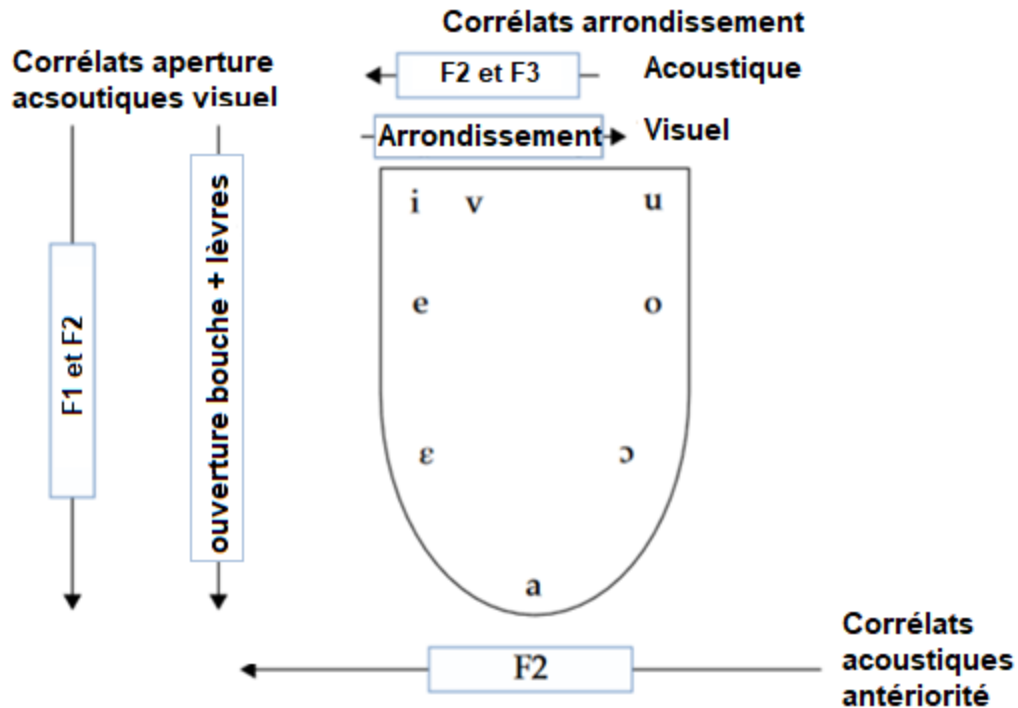


Figure 3.1 Illustration de la corrélation entre lieu d'articulation (en bas), ouverture de la bouche et arrondissement (à gauche) et arrondissement (en haut) avec les différentes valeurs de formant acoustiques, tiré de Ménard, (2015)

De plus, ces voyelles étant présentes dans la quasi-totalité des langues, elles rendent possible la comparaison de nos résultats avec la phonétique d'autres langues et leur donnent une portée plus universelle. Enfin, les effets coarticulatoires audibles et visibles de ces voyelles sont bien étudiés en français (Benguerel et Cowan, 1974 ; Benoit *et al.*, 1994 ; Robert-Ribes *et al.*, 1998) et nous permettent de formuler des hypothèses précises sur les effets de notre perturbation. Nous voyons dans la Figure 3.2 les positions de la langue lors de l'articulation de ces trois voyelles et leur position relative dans un trapèze vocalique. Nous voyons comme dans la Figure 3.1 que la distance articulaire est maximale (aperture et antériorité) pour les trois voyelles représentées, et il faut ajouter à cela une troisième dimension liée à l'arrondissement pour le [u] (symbolisé par un cercle sur les lèvres). Ces 3 dimensions articulaires se manifestent parfois visuellement (l'aperture par le biais de l'ouverture de la bouche et arrondissement des lèvres) et toujours acoustiquement en faisant varier les valeurs des formants F1, F2 et F3 notamment (la Figure 3.2 illustre ces corrélations).

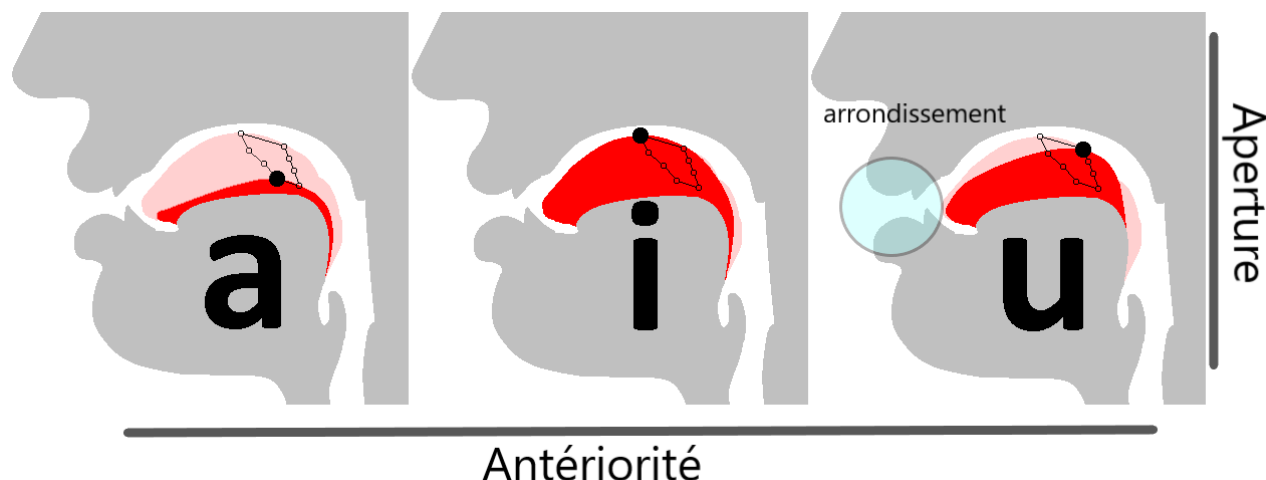


Figure 3.2 Position des articulateurs lors de l'articulation des voyelles cibles

Quand l'information acoustique est détériorée, la personne participante doit utiliser les informations visuelles disponibles. Ainsi, chaque contexte vocalique viendra influencer par coarticulation les informations acoustiques et visuelles liées aux consonnes (Benoit *et al.*, 1994 ; Robert-Ribes *et al.*, 1998). De plus, les traits articulatoires sont perçus différemment lorsqu'ils sont perturbés acoustiquement. L'étude de Robert-Ribes *et al.* 1998) montre la robustesse des traits à la perturbation en fonction de la modalité de perception :

- Modalité audio seule : aperture>antériorité>arrondissement
- Modalité visuelle seule : arrondissement>aperture>antériorité

Rappelons que les stimuli en modalité audiovisuelle obtiennent toujours de meilleurs scores de reconnaissance que les stimuli en modalité auditive. Cependant, les scores de perception en modalité AV n'est pas la somme simple de ses deux modalités isolées. Il faut, selon les auteurs-es, prendre en compte la spécificité de chaque contexte vocalique, car la saillance articulatoire peut être plus visuelle (arrondissement) ou plus acoustique (aperture) selon les traits articulatoires. Le cerveau utilisera, selon le contexte de perception, l'information la plus pertinente disponible pour traiter l'information perçue.

Pour les consonnes, les contraintes articulatoires sont différentes de celles des voyelles. Nous avons choisi des consonnes s'articulant oralement, plosives et non-voisées. Ces trois paramètres sont suffisants pour

créer un contraste visuel en changeant le lieu d'articulation. Comme pour les voyelles, des consonnes articulées proches et ayant des points d'articulation en commun seraient plus difficiles à discerner. En effet, la différence entre [p] / [m] ou [t] / [d] est quasi-nulle visuellement, car ces paires de consonnes sont articulées au même endroit. Nous avons donc choisi [p, t, k]. Comme nous le voyons dans la Figure 3.3, ces 3 consonnes orales, plosives et non-voisées s'articulent en obstruant complètement puis en relâchant brusquement le passage de l'air dans la cavité buccale au niveau des lèvres pour le [p], de l'arrière des dents et de l'apex de la langue pour le [t] et du voile du palais et du dos de la langue pour le [k]. Nous avons choisi des consonnes orales et non nasales, car la nasalité entraîne des effets de coarticulation complexes qui influencent le contexte vocalique et rendent plus difficile la perception (Malécot et Metz, 2009). Pour terminer, nous avons choisi des plosives non-voisées car la pression intraorale lors de leur articulation est plus élevée que les voisées (Arkebauer Herbert J. *et al.*, 1967) ce qui maximise l'effet de relâchement de l'air et rend la consonne visuellement plus saillante (gonflement des lèvres et joues lors de l'augmentation de la pression intra-orale avant le relâchement de la pression) (Shipp, 1973). En regardant la Figure 3.3, on peut déjà supposer, grâce aux lieux d'articulation, la facilité relative de perception visuelle de la consonne [p].

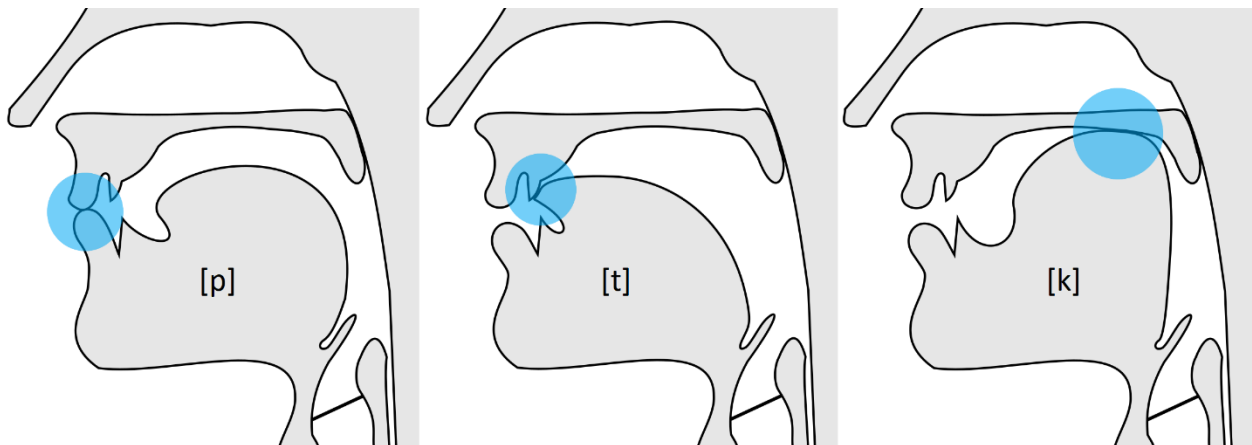


Figure 3.3 Lieux d'articulation des consonnes

3.1.2 Enregistrement des stimuli

Les syllabes ont été articulées clairement par une locutrice du français québécois. Nous avons réalisé un enregistrement synchrone audiovisuel et chaque stimulus a été répété une dizaine de fois et un seul exemplaire a été sélectionné sur la base de la clarté de la production. Pour l'enregistrement, nous avons utilisé un enregistreur Zoom™ H1N et une caméra Canon™ 550D avec un objectif 18/55mm. Les clips audio

et vidéo de 9 syllabes ont ensuite été extraits à l'aide du logiciel d'édition vidéo Shotcut (Dennedy, 2021). Le son a alors été séparé des vidéos pour le traitement audio.

3.1.3 Perturbation audio

Le son a été traité à l'aide d'un script dans le programme d'analyse phonétique Praat (Boersma et Weenink, 2021), afin d'y ajouter le bruit rose de la perturbation audio. Pour établir le niveau de bruit rose nécessaire à la perturbation, nous nous sommes appuyés sur l'étude de Robert-Ribes *et al.*, (1998). Nous avons ensuite réalisé des tests afin de vérifier que la perturbation était suffisamment forte. Nous voulions éviter un effet plafond à 100% de réussite si la perturbation était trop faible. Nous avons établi la perturbation à un niveau de bruit rose un peu plus fort à -18 dB de ratio bruit-signal (RBS) pour obtenir 60% de réussite en modalité audio seule.

3.1.4 Perturbation vidéo

La perturbation vidéo a été réalisée avec l'aide du laboratoire de psychologie de la perception de Dave Saint-Amour. Trois filtres (filtrage gaussien) de fréquences spatiales passe-bas à 6 cycles par face (cpf), 12 cpf et 20 cpf ont été appliqués aux extraits vidéo dans le logiciel Matlab (The MathWorks Inc., 2021). Nous avons défini ces trois valeurs en nous basant sur les recherches ayant déjà utilisé ce type de perturbation de Campbell et Massaro, (1997) et Munhall *et al.*, (2004). La Figure 3.4 présente des exemples de stimuli que nous avons obtenus après filtrage de la vidéo. Les images étaient en couleurs.



Figure 3.4 Capture d'écran des stimuli pour les 3 niveaux de filtrage passe-bas (20, 12 et 6 cpf de gauche à droite)

Nous avons ensuite synchronisé les extraits vidéo avec l'audio perturbés. Une fois nos 9 stimuli prêts, nous les avons intégrés à notre tâche de perception.

3.2 Conception de la tâche en ligne

Considérant les contraintes de la pandémie de COVID-19, nous avons dû concevoir un test pouvant se réaliser en ligne et à distance de manière indépendante par les personnes participantes. Le matériel informatique (taille de l'écran, écouteur et OS) étant différent pour chaque personne participante, nous testerons ces paramètres comme des variables aléatoires dans notre modèle statistique et nous reviendrons sur ce sujet dans notre discussion.

Nous avons conçu la tâche à l'aide du logiciel Psychopy (Open Science Tools Ltd., 2022b) et nous l'avons transférée en ligne sur la plateforme Pavlovia (Open Science Tools Ltd., 2022a). Pour que les ordinateurs de toutes les personnes participantes affichent des stimuli de la même taille malgré les dimensions de l'écran variables, nous avons ajouté un script de calibrage d'écran conçu par l'assistant du laboratoire de Psychologie de la perception (Anthony Hosein Poitras Loewen). Il fallait, pour calibrer le test, utiliser un objet de taille standardisée que tout le monde possède : une carte bancaire. Avant la passation, l'image d'une carte bancaire s'affichait sur l'écran, et à l'aide des flèches du clavier, il fallait ajuster sa taille afin que l'image corresponde à la taille de notre carte bancaire réelle que l'on apposait sur l'écran. Cela permettait d'afficher un stimulus de la même taille pour toutes les personnes participantes. Chaque stimulus était répété 20 fois à 3 niveaux de perturbation pour les 9 syllabes. Il y avait donc 540 stimuli par passation. Les stimuli duraient en moyenne 1059 ms avec un écart-type de 94,6 ms et l'attaque de la consonne était en moyenne à 540 ms avec un écart-type de 57,9 ms.

3.3 Population

Pour recruter les personnes participantes, nous sommes passés par les associations universitaires de l'UQAM et par nos réseaux personnels. Les personnes intéressées devaient remplir un questionnaire préalable afin d'indiquer s'ils-elles avaient une déficience auditive ou visuelle corrigée. Nous n'acceptons que les personnes n'ayant pas de trouble de l'audition ou de la vision ou ayant un trouble de la vision corrigé. Nous récupérons différentes informations pour le traitement des données : sexe, âge, niveau d'étude et occupation. Nous récupérons aussi les informations liées au matériel utilisé (type d'écouteurs,

types d'ordinateur et taille de l'écran). Pour récompenser les participants.es, un tirage au sort a été réalisé et deux personnes ont reçu un prix de 50\$. La tâche a été effectuée par 22 personnes âgées de 21 à 36 ans, incluant 15 femmes et 7 hommes. Parmi ces 22 personnes, 8 étaient étudiants-es en linguistique et avaient donc une connaissance poussée de la phonétique et de la phonologie du français québécois. De plus, sur ces 22 personnes, 19 étaient droitières et 18 avaient un trouble de la vision corrigé par des lunettes ou des lentilles. Nous reviendrons sur les effets potentiels de ces paramètres et leur traitement comme variables aléatoires dans la partie 3.5 Analyses des données.

Tableau 3.1 Profil des personnes participantes

Participant·es	Femmes	Droitières·ers	Âge moyen	Min- max	Étudiant·es	Linguistes	Troubles de la vision corrigés
22	15	19	28,6	21- 36ans	15	8	18

3.4 Passation

Notre tâche était constituée de 3 sessions d'environ 10 minutes qui se suivaient sans pause, mais avec un écran indiquant la fin d'une session et lors duquel la personne participante devait appuyer sur espace pour passer à la session suivante. Une série de syllabes était testée dans chaque session en suivant l'ordre [a], [i] puis [u]. Les personnes participantes devaient discriminer entre les trois consonnes, la voyelle étant la même pour chaque session de test. La passation se faisait sur ordinateur personnel. Les personnes participantes avaient pour consignes de passer la tâche dans un environnement calme, où elles ne seraient pas dérangées, en utilisant des écouteurs (casque ou écouteur intra-auriculaire) et sans prendre de pause. La passation durait entre 25 et 30 minutes. Pour lancer la tâche, il fallait cliquer sur le lien préalablement transmis. La tâche commençait par le calibrage de l'écran à l'aide d'une carte de banque au format standard. Une fois l'écran calibré, la tâche commençait par une phase d'entraînement avec de faux stimuli présentés 10 fois. Une fois cette phase terminée un écran indiquait que la tâche commencerait quand le·a participant·e appuierait sur la barre d'espacement. Les stimuli étaient présentés dans un ordre aléatoire

fixe pour tous les participants-es. Une fois la tâche terminée, la fenêtre se fermait et les résultats étaient transmis sur le serveur sécurisé de la plateforme Pavlovia (Open Science Tools Ltd., 2022a).

3.5 Analyse des données

Dans les travaux mentionnés dans notre revue de littérature, le score de reconnaissance est la seule variable dépendante employée pour mesurer l'effet de la perturbation sur l'intégration de la parole. Il permet de mesurer la réussite d'une personne participante à une tâche et nous indique d'éventuels effets plafond si la tâche est trop facile. Cette variable dépendante est donc essentielle à la compréhension et à l'interprétation de l'effet des variables indépendantes comme le lieu d'articulation ou le filtre de fréquences spatiales appliqué. Nous n'avons pas utilisé le temps de réponse, car nous ne pouvions pas avoir de contrôle précis de celui-ci, le test ne pouvant pas être passé sous notre supervision.

Nos variables indépendantes (VI) sont les suivantes :

- Le filtre de fréquences spatiales
- Le contexte vocalique
- La consonne (lieu d'articulation)

Les autres effets éventuels (participants, matériel, etc.) sont analysés en tant que variables aléatoires dans notre modèle statistique. Après le traitement statistique, nous analysons les SDR grâce à des matrices de confusion. Cela nous permet de mieux comprendre le profil d'erreur pour les syllabes, les consonnes et les voyelles des stimuli. Nous mesurons également le gain perceptif des filtres de fréquence spatiale 12 cpf et 20 cpf par rapport au filtre 6 cpf pour les syllabes, les consonne et les voyelles.

3.6 Analyse statistique

3.6.1 Statistiques descriptives et données aberrantes

Une fois nos données obtenues, nous avons 11880 entrées obtenues au test et aucune donnée manquante pour la variable dépendante *bonne réponse*. Nous avons utilisé les temps de réponses pour

éliminer les données aberrantes. Nous avons une asymétrie positive liée au fait qu'il y a un plus grand nombre de temps de réponse courts que de temps de réponse longs, ce qui est normal pour ce type de variable. L'indice de courbure de Pearson, qui doit être inférieur à 1 pour une distribution normale, était de 2,12. Afin d'éliminer les valeurs aberrantes, nous avons analysé l'espace interquartile IQR et utilisé la formule suivante de Tukey pour déterminer l'intervalle de données pertinent à partir du premier et du troisième quartile (Q1 et Q3 dans la formule) :

- $[Q1 - 1.5 \times IQR ; Q3 + 1.5 \times IQR]$

Cette formule permet d'obtenir les données situées à 1,5 espace interquartile du milieu de la distribution. Notre limite inférieure était négative nous avons donc pris 0 ms et notre limite supérieure était 874 ms. Il restait donc 9661 bonnes réponses et 1390 réponses incorrectes, soit 11051 items au total ou 93,02% des réponses obtenues avant traitement. Après traitement des données, la courbe présentait une distribution normale avec un indice d'asymétrie de 0,504.

3.6.2 Analyses multiniveaux

Nous avons donc réalisé des analyses multiniveaux ou modèles linéaires à effet mixte. Ce type d'analyse permet de prendre en compte différents niveaux de regroupement des données, comme *la personne participante* qui est analysée comme variable aléatoire. Si ces variables aléatoires ont un effet sur la distribution des données, celui-ci est pris en compte dans l'analyse des autres variables. Pour simplifier l'interprétation des résultats, nous avons réalisé deux analyses séparées, une pour les syllabes et une les consonnes et voyelles. Cela est nécessaire, car aux traits phonémiques des composants de la syllabe s'ajoutent les effets phonologiques liés à l'articulation et ce tout n'équivaut pas à la simple somme des traits phonémiques. Ainsi, il est plus facile d'interpréter l'effet des phonèmes séparément ou de la syllabe comme un tout, plutôt que traiter toutes les informations dans la même analyse. Nous avons utilisé la variable dépendante *score de reconnaissance* (SDR) qui correspond au pourcentage de reconnaissance correcte pour chaque contexte vocalique de l'item présenté. Dans la première analyse, nous avons étudié l'effet des VI *filtre de FS* et *syllabe* sur les SDR. Cela nous permet de mesurer l'effet des filtres de FS sur le stimulus en tant qu'unité perçue. Dans la deuxième analyse nous avons étudié l'effet des VI : *filtre de FS*, *consonnes* et *contexte vocalique* sur les SDR. Ces VI nous permettent de tester l'effet des filtres de FS en décomposant notre stimulus en deux phonèmes (consonne et voyelle) et nous permet de comparer cet effet à celui des filtres de FS sur les syllabes. Concernant les variables aléatoires, l'analyse statistique nous montre que seule la variable aléatoire *participant* a un impact significatif sur la distribution des données.

Les variables aléatoires liées au sexe, à la main dominante, à la taille de l'écran, au niveau d'étude, à la correction visuelle et au matériel audio employé n'ont pas d'effet significatif sur la distribution des données et nous ne les avons retirées de notre analyse. Voici les formules employées dans l'analyse multiniveaux, aussi appelée modèle linéaire à effet mixte (*Linear mixed model fit by REML*), du logiciel Jamovi, (2023) :

- **Modèle consonne & voyelle** : $Réponse \sim 1 + FS + consonne + voyelle + FS:consonne + FS:voyelle + consonne:voyelle + FS:consonne:voyelle + (1 | participant)$
- **Modèle syllabe** : $Réponse \sim 1 + FS + syllable + FS:syllable + (1 | participant)$

3.7 Analyse secondaire :

Afin d'approfondir l'analyse et l'interprétation de nos résultats, nous avons calculé le gain perceptif, le taux d'erreur et nous avons réalisé les matrices de confusion pour la tâche.

3.7.1 Formules du GP des consonnes

Pour calculer les gains perceptifs par consonne, nous avons soustrait les SDR par consonne en pourcent du filtre à 6 cpf au SDR en pourcent des consonnes pour les filtres de FS 12 cpf et 20 cpf. Les formules étaient les suivantes :

- $GP_{\text{filtre 12 cpf des consonnes}} = SDR_{\text{filtre 12cpf des consonnes}} - SDR_{\text{filtre 6cpf des consonnes}}$
- $GP_{\text{filtre 20 cpf des consonnes}} = SDR_{\text{filtre 20cpf des consonnes}} - SDR_{\text{filtre 6cpf des consonnes}}$

3.7.2 Formules du GP des voyelles

Pour calculer les gains perceptifs par voyelle, nous avons soustrait les SDR par voyelle en pourcent du filtre à 6 cpf au SDR en pourcent des voyelles pour les filtres de FS 12 cpf et 20 cpf. Les formules étaient les suivantes :

- $GP_{\text{filtre 12 cpf des voyelles}} = SDR_{\text{filtre 12cpf des voyelles}} - SDR_{\text{filtre 6cpf des voyelles}}$
- $GP_{\text{filtre 20 cpf des voyelles}} = SDR_{\text{filtre 20cpf des voyelles}} - SDR_{\text{filtre 6cpf des voyelles}}$

3.7.3 Formules du GP des syllabes

Pour calculer les gains perceptifs par syllabes, nous avons soustrait les SDR par syllabe en pourcent du filtre à 6 cpf au SDR en pourcent des syllabes des filtres de FS 12 cpf et 20 cpf. Les formules étaient les suivantes :

- $GP_{\text{filtre 12 cpf des syllabes}} = SDR_{\text{filtre 12cpf des syllabes}} - SDR_{\text{filtre 6cpf des syllabes}}$
- $GP_{\text{filtre 20 cpf des syllabes}} = SDR_{\text{filtre 20cpf des syllabes}} - SDR_{\text{filtre 6cpf des syllabes}}$

3.7.4 Taux d'erreur et matrice de confusion

Pour compléter l'analyse des données, nous avons déterminé le taux d'erreur par consonnes, voyelles et syllabes pour chaque filtre de FS. Nous avons aussi mesuré les pourcentages de confusion inter-consonne pour les 3 valeurs de filtre de FS. Cela nous permet de savoir par quelle autre consonne une consonne est confondue et ainsi de mieux comprendre les erreurs des personnes participantes.

CHAPITRE 4 Résultats

Dans ce mémoire, nous nous sommes intéressés à la perception audiovisuelle de syllabes en français par le biais d'une approche impliquant la dégradation du signal audio et du signal visuel à différents degrés. Les stimuli impliquaient des séquences syllabiques ouvertes formées de consonnes variant selon le lieu d'articulation et des voyelles cardinales. Dans ce chapitre, nous présentons une analyse détaillée des scores de reconnaissance, reliée à nos hypothèses de départ.

4.1 Effets des filtres de fréquence spatiale sur le score de reconnaissance (Hypothèse 1)

Rappel : H.1 *Plus la valeur des filtres de FS sera faible (et donc plus la perturbation sera importante), plus les scores de reconnaissance seront faibles.*

Cette section présente les scores de reconnaissance (SDR) moyens obtenus pour chaque valeur de fréquence spatiale (FS) des filtres passe-bas. Ces résultats permettent d'évaluer comment la perception des participants-es est affectée par les filtres de FS appliqués sur les vidéos de la tâche. Ils rendent compte de la difficulté relative de perception des syllabes articulées. Ainsi, plus la valeur du filtre de FS est basse et plus l'information visuelle est grossière (floue, voir Figure 4.1). Les FS basses correspondent aux formes générales d'une image et les FS hautes aux informations plus détaillées. Rappelons que le signal audio est dégradé de façon uniforme pour tous les stimuli. Cette première figure nous montre les SDR correspondants à chaque filtre de FS.

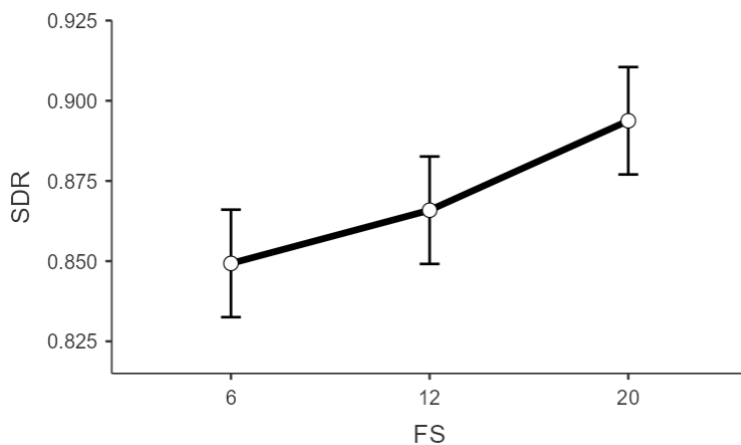


Figure 4.1 Scores de reconnaissance moyens en fonction du filtre de fréquence spatiale et erreurs standards.

On observe, dans la figure 4.1, l'effet des trois fréquences spatiales sur les SDR. Les tests statistiques confirment l'effet significatif de la FS sur les scores de reconnaissance ($F(2,11003)=19,75$; $p<0.001$). Parmi les trois valeurs de filtre, plus le filtre de FS est bas (moins d'information visuelle, image plus « floue ») plus le SDR est faible. Les tests post-hoc (Bonferroni) sur la variable *filtre de FS* nous indiquent que les différences de score de reconnaissance sont statistiquement significatives entre 6 cpf et 20 cpf ($t=-6,22$; $p<0.001$) et entre 12 et 20 ($t=-3.91$; $p<0.001$), mais pas entre les filtres 6 cpf et 12 cpf. Il est important de noter qu'à la fréquence 20 cpf, le score moyen est de 89,7%, ce qui suggère que la modalité visuelle est véritablement fonctionnelle dans le processus d'identification perceptive, puisque le signal audio est bruité.

4.2 Effets des filtres de fréquence spatiale sur la perception des phonèmes (Hypothèse 2)

- **Rappel : H. 2** *La perturbation aura des effets différents selon le domaine et le type de phonème sur lequel elle est appliquée (syllabe vs phonème, et voyelle vs consonne). La saillance articulatoire des traits phonétiques influencera les scores de reconnaissance* (Binnie et al., 1974 ; Macleod et Summerfield, 1987 ; Robert-Ribes et al., 1998)

4.2.1 Score de reconnaissance par consonnes

La seconde hypothèse de travail sera testée par différents types d'analyse. Dans cette section, nous évaluons l'impact des consonnes sur les scores. La Figure 4.2 présente l'effet des filtres de fréquences spatiales par consonne. Les SDR augmentent significativement avec les valeurs de filtres de FS pour toutes les consonnes. L'interaction *consonnes*FS* n'est pas significative. On remarque, en plus de l'effet des filtres de FS, un effet statistiquement significatif de la consonne sur les SDR ($p<0.001$, $F(2, 11024)=154,42$). On observe que les syllabes articulées avec /p/ entraînent un SDR moyen plus élevé par rapport aux syllabes produites avec les deux autres consonnes /t/ et /k/. Les tests post-hocs confirment que les différences de SDR inter-consonnes sont significatives pour le /p/ et le /k/ (Post hoc pour l'interaction, k-p : $t=-16,45$; k-t : $t=-3,01$; p-t : $t= 13,48$, $p<0.001$). La saillance visuelle des consonnes a donc un impact sur les SDR. Ces résultats confirment le volet consonne de l'hypothèse H2.

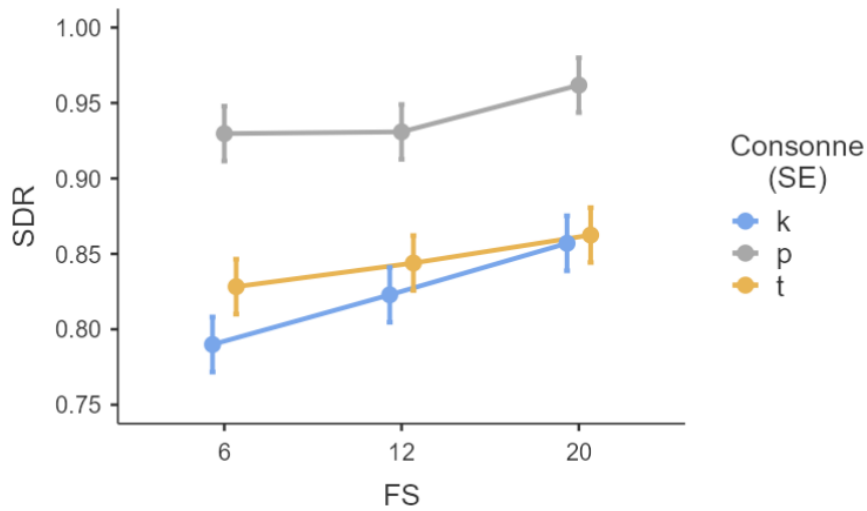


Figure 4.2 Scores de reconnaissance moyens par filtres de fréquence spatiale et par consonne et erreurs standards

4.2.2 Score de reconnaissance par contexte vocalique (voyelle)

Lorsqu'on observe graphiquement l'effet des filtres de fréquence spatiale par contexte vocalique (Figure 4.3), on remarque que l'effet des filtres de FS varie en fonction de la voyelle. La voyelle /a/ a les plus hauts scores moyens, suivie de /i/ et enfin de /u/. De plus, on voit que pour le /u/ et le /a/, le filtre à 20 cpf semble obtenir de meilleurs SDR moyens que les deux autres filtres.

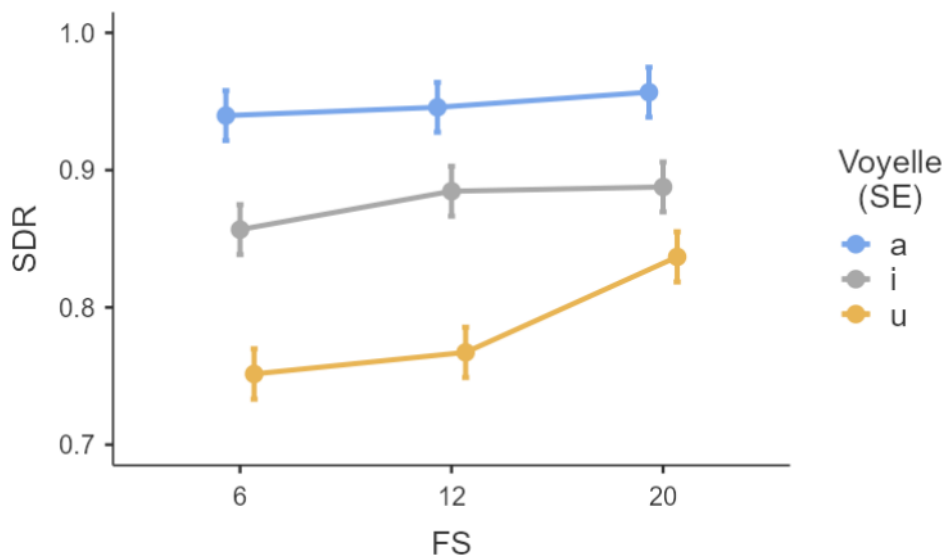


Figure 4.3 Scores de reconnaissance moyens par voyelle et filtre de FS et erreurs standards

L'analyse statistique montre des effets significatifs des variables *voyelle* ($F(2, 11004) = 255,59, p < 0.001$) et *filtre de FS* ($F(2, 11003) = 19,75, p < 0.001$) comme effets principaux et en interaction *FS*voyelle* ($F(2, 11003) = 5,88, p < 0.001$). Les tests Post-Hoc indiquent des différences significatives de score de reconnaissance pour toutes les voyelles (Tableau 4.1).

Tableau 4.1 Test Post Hoc (Bonferroni) pour la variable voyelle

Voyelle		t	ddl	p _{bonferroni}
a	i	10.1	11004	< .001
a	u	22.6	11004	< .001
i	u	12.7	11005	< .001

Dans l'interaction *voyelle*FS*, pour /u/, les scores de reconnaissance entre les filtres 6 cpf et 20 cpf et entre les filtres 12 cpf et 20 cpf sont significativement différents (

Tableau 4.2). Cela confirme ce qui est observé graphiquement. La voyelle /u/ est la seule voyelle (parmi les 3 voyelles testées ici) avec un trait articulaire d'arrondissement saillant visuellement et cela se répercute sur les SDR par rapport aux filtres de fréquences spatiales plus bas. Pour celle-ci, il y a plus de détails visuels et l'augmentation des SDR est plus importante par rapport aux deux autres voyelles. Ces données viennent confirmer statistiquement et graphiquement l'aspect vocalique de l'hypothèse 2. La saillance visuelle des voyelles combinée aux valeurs de filtre de FS se répercute sur les SDR.

Tableau 4.2 Test Post Hoc pour la l'interaction *FS*voyelle*

FS	Voyelle	FS	Voyelle	t	ddl	p _{bonferroni}
6	u	12	u	-1.240	11003	1
12	u	20	u	-5.524	11003	< .001
6	u	20	u	-6.728	11003	< .001

4.2.3 Score de reconnaissance par syllabe

Afin de brosser un portrait de l'effet de la dynamique articulatoire sur les valeurs de SDR dans les différents contextes de perturbation, nous présentons, dans la Figure 4.4, les scores moyens de SDR pour chaque syllabe dans les trois conditions (contexte vocalique) de la tâche. Il est à noter que les unités syllabiques permettent de distinguer les stimuli, non pas selon les valeurs statiques de leurs traits phonétiques (comme dans les sections précédentes), mais plutôt selon l'ampleur de la dynamique visuelle. Une syllabe comme [pa], par exemple, implique le passage d'une consonne bilabiale impliquant la compression des lèvres, à une voyelle ouverte, impliquant l'ouverture maximale des lèvres. En revanche, une syllabe comme [ku] n'implique que très peu de déplacements labiaux entre la consonne et la voyelle, l'arrondissement des lèvres de la voyelle [u] pouvant être maximalement anticipé dès le début de la consonne. On remarque la différence marquée pour le contexte [u], combinés aux 3 consonnes, par rapport aux deux autres contextes vocaliques ([a] et [i]).

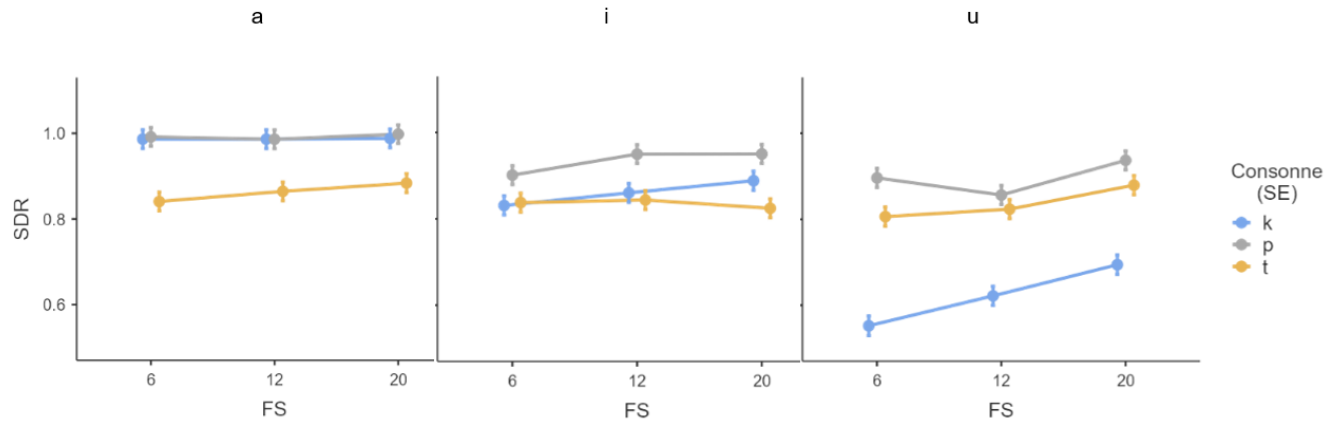


Figure 4.4 SDR moyens et erreurs standards par filtre de FS et syllabe séparés par contexte vocalique

Lorsqu'on regarde l'effet combiné des filtres de fréquences spatiales par syllabes sur les scores de reconnaissance (Figure 4.4), on remarque que la plupart des syllabes ont des SDR qui augmentent avec la valeur du filtre de FS (excepté /pu/ et /ti/). Le filtre 6 cpf a les SDR les plus faibles, ce qui est attendu. Dans cette condition, la syllabe /pu/ est l'item le plus saillant visuellement (combinaison de l'arrondissement et du point d'articulation labiale), comparé aux /ku/ et /tu/ il obtient logiquement les meilleurs SDR. Cela vient appuyer notre hypothèse 2 sur la saillance des traits visuels. Les syllabes /pa/ et /ka/ ont des scores supérieurs (environ 98% de réussite alors que les autres scores plafonnent à 88%) avec des écarts faibles (moins de 1%) entre filtres de fréquence spatiale. La syllabe /ku/ se trouvent nettement en dessous du reste, entre 54% et 68% alors que la moyenne est à 85,9%. Enfin, les syllabes contenant le /p/ obtiennent toujours les meilleurs SDR. Les résultats détaillés de l'analyse statistique sont présentés aux tableaux 4.3 et 4.4.

Tableau 4.3 Résultats de l'analyse multiniveaux pour les syllabes

	F	Num ddl	Den ddl	p
FS	19.76	2	11003	< .001
Syllabe	151.62	8	11832	< .001

	F	Num ddl	Den ddl	p
FS * Syllabe	3.05	16	11003	< .001

Tableau 4.4 Test Post-Hoc pour l'interaction syllabe-FS

FS	Syllabe	FS	Syllabe	Difference	SE	t	df	p_{bonferroni}
6	ku	20	ku	-0,14210	0,0226	-6.290	11003	< ,001

4.3 Gain perceptif

Afin d'analyser plus en détail l'effet des filtres de fréquence spatiale sur les SDR, nous avons calculé les gains perceptifs (GP) par syllabe, consonne et contexte vocalique. Ces données nous permettent d'apprécier l'effet positif ou négatif des différentes variables indépendantes sur les SDR. Nous avons obtenu les graphiques suivants (Figure 4.5) :

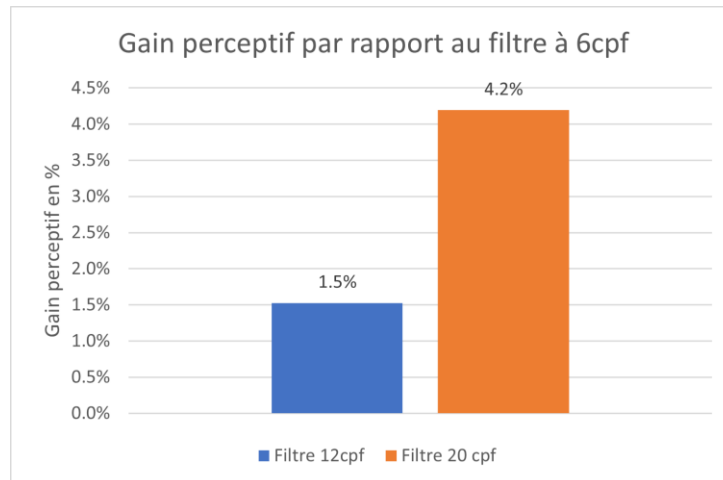


Figure 4.5 Gain perceptif par filtre de FS

Dans la Figure 4.5, nous observons le gain perceptif pour tous les stimuli confondus. Nous voyons ici le gain perceptif des filtres de FS 12 et 20 par rapport au filtre à 6 cpf. Cette figure complète la Figure 4.1 qui décrit les SDR par filtre de FS et illustre l'amélioration des SDR avec l'augmentation de la valeur des filtres. Voyons maintenant plus en détail l'effet des différents paramètres articulatoires sur le GP.

4.3.1 Gain perceptif par consonne

La Figure 4.6 présente les valeurs moyennes de GP par consonne et par FS. Le GP augmente de manière statistiquement significative avec la valeur du filtre de fréquence spatiale par rapport au filtre à 6 cpf pour toutes les consonnes (Kruskal-Wallis : $X^2 = 7,45$ $p < 0,05$, $ddl = 2$, $\epsilon^2 = 0,931$). Le GP varie selon la consonne et augmente le plus par rapport au filtre à 6 cpf pour $[p] < [t] < [k]$. Le GP par consonne est toujours supérieur pour le filtre à 20 cpf par rapport au GP du filtre à 12 cpf pour toutes les consonnes. De façon générale, ces résultats confirment que l'information visuelle accessible sur le visage de l'interlocuteur contribue à des scores de reconnaissance plus élevés. Les résultats selon la nature des phonèmes seront présentés dans les sections suivantes.

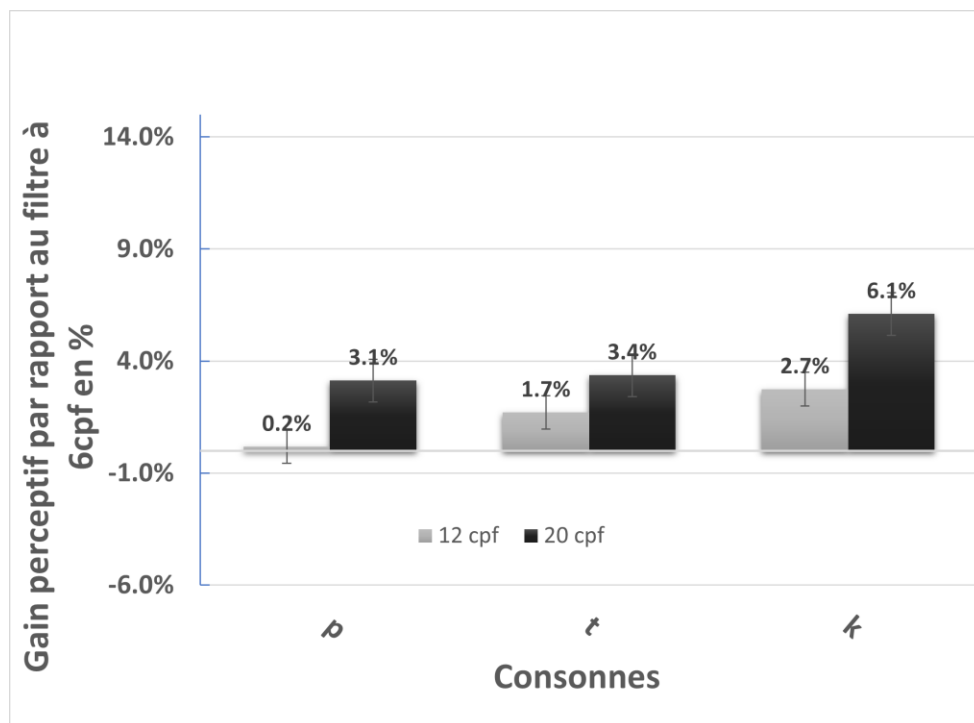


Figure 4.6 Gain perceptif moyen par consonne par rapport au filtre à 6 cpf et erreurs standards

4.3.2 Gain perceptif par voyelle

La Figure 4.7 présente les valeurs moyennes de GP en fonction de la FS et de la voyelle. Dans cette figure, le GP augmente avec la valeur du filtre de fréquence spatiale pour toutes les voyelles par rapport au filtre à 6 cpf. Le GP varie selon la voyelle et augmente de $[a] < [i] < [u]$ pour le filtre à 20 cpf. Le GP par voyelle est toujours supérieur pour le filtre à 20 cpf par rapport au GP du filtre à 12 cpf pour toutes les voyelles. À 20 cpf, le gain perceptif est plus important pour le $[u]$ que pour les deux autres voyelles. Il y a un effet statistiquement significatif du filtre de FS sur le gain par rapport au filtre à 6 cpf (Kruskal-Wallis : $\chi^2 = 6,71,45$ $p < 0,05$, ddl = 2, $\epsilon^2 = 0,839$). La voyelle arrondie $/u/$ est donc celle qui bénéficie le plus d'un signal visuel minimalement dégradé, suivie de $/i/$ et de $/a/$.

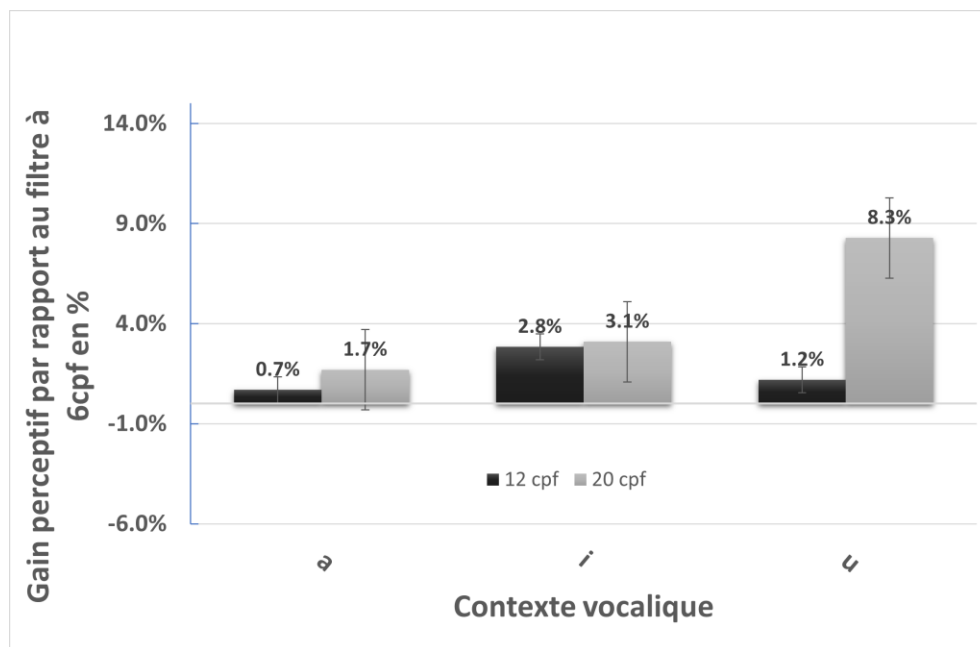


Figure 4.7 Gain perceptif moyen par voyelle par rapport au filtre à 6 cpf et erreurs standards

4.3.3 Gain perceptif par syllabe

Afin d'apprécier la dynamique articulatoire impliquée dans l'enchaînement de consonnes et de voyelles, le Figure 4.8 présente la variation des valeurs de GP par syllabes. Nous voyons dans cette figure que le gain perceptif (GP) varie d'un stimulus à l'autre et que la syllabe [ku] a les GP les plus importants, 7,3% et 14,3%, respectivement pour les filtres 12 et 20 cpf par rapport au filtre à 6 cpf. On observe aussi que [pa, pu] ont des GP négatifs pour le filtre à 12 cpf et que [ti] a un GP négatif pour le filtre à 20 cpf. Pour [ta, pi, ki, tu, ku] le GP augmente entre le filtre 12 cpf et 20 cpf. Il n'y a quasiment pas de gain perceptif pour [pa, ka]. De manière générale, on observe un gain statistiquement significatif entre 6 et 12 cpf et entre 6 et 20 cpf pour les syllabes ($X^2 = 18,4$ $p < 0,001$, $ddl = 2$, $\epsilon^2 = 0,708$). Les séquences qui bénéficient donc d'une meilleure qualité visuelle sont donc celles qui impliquent la voyelle /u/. Étonnamment, malgré le fait que les consonnes /t/ et /k/ ne sont que minimalement distinctes visuellement, lorsqu'elles sont suivies de la même voyelle /u/, elles sont différemment affectées par la dégradation visuelle (cf. gains perceptifs différents pour /tu/ et /ku/ dans la figure 4.8). Il en est de même pour chacune des autres voyelles (nous y reviendrons dans la discussion).

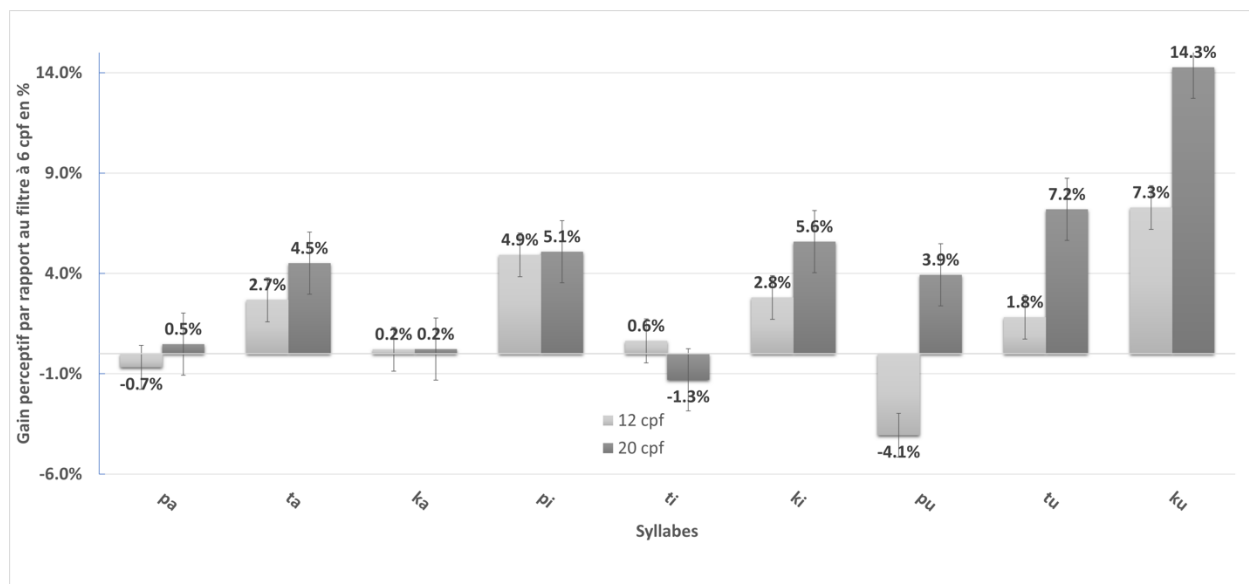


Figure 4.8 Gain perceptif moyen par syllabe par rapport au filtre à 6 cpf et erreurs standards

4.4 Taux d'erreur

4.4.1 Pourcentage d'erreur par consonne

La Figure 4.9 illustre les pourcentages d'erreur de chaque consonne pour les trois valeurs de filtre de FS. Nous remarquons que les pourcentages d'erreur diminuent avec l'augmentation des valeurs de filtre de FS pour les trois consonnes. Le [k] et le [t] qui sont des consonnes dont le contraste est non-visible, obtiennent des pourcentages d'erreur 2 à 3 fois plus hauts par rapport au [p] qui est visible. Le [k] obtient des pourcentages d'erreur légèrement plus hauts que le [t], sauf pour le filtre à 20 cpf où les pourcentages d'erreur ne diffèrent que de 0,2% entre les deux consonnes. Enfin, on remarque que le taux d'erreur pour le [t] et le [k] diminue de façon constante entre les trois filtres, alors que le [p] décroît le plus entre les filtres à 12 et 20 cpf.

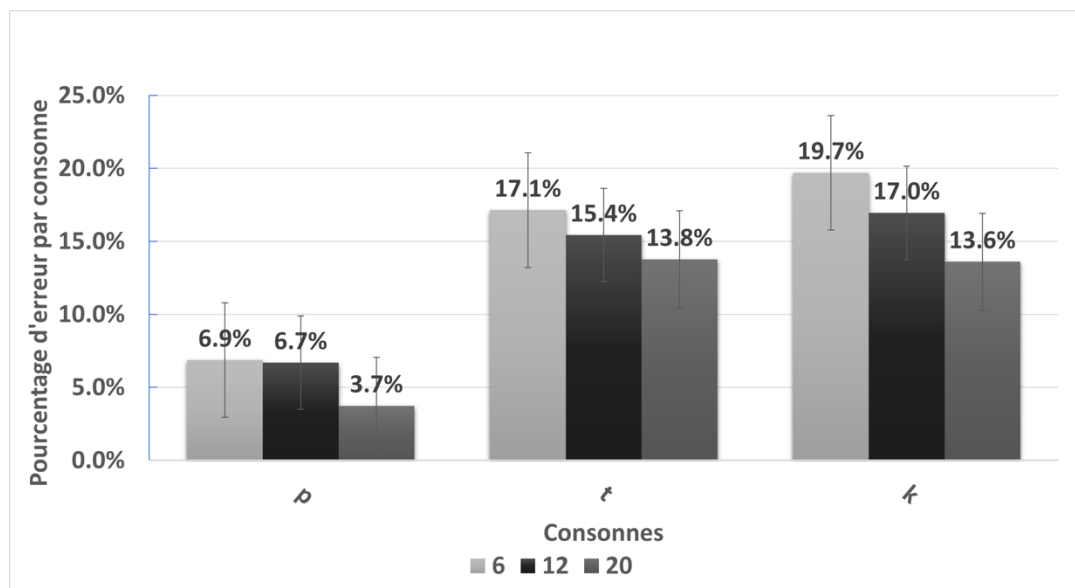


Figure 4.9 Pourcentage d'erreur par consonne et fréquence spatiale et erreurs standards

4.4.2 Pourcentage d'erreur par voyelle

La Figure 4.10 illustre les pourcentages d'erreur de chaque contexte vocalique pour les trois valeurs de filtre de FS. Nous remarquons que les pourcentages d'erreur diminuent avec l'augmentation des valeurs de filtre de FS pour les trois voyelles. Cet effet est statistiquement significatif ($X^2 = 18,4$ $p < 0,05$, $ddl = 2$, $\epsilon^2 = 0,9$). Le [u], qui est la seule voyelle arrondie et donc avec un trait articulatoire visible, obtient les plus hauts taux d'erreur entre 24,2% à 15,9% selon le filtre. La voyelle [i] se situe entre 14,2% et 11,1% d'erreur et le [a] a des taux d'erreur entre 5,9% et 4,2%. Le [u] est la voyelle qui voit son taux d'erreur diminuer le plus entre les filtres de FS 12 cpf et 20 cpf. Le [i] est la voyelle dont le taux d'erreur diminue le plus entre 6 et 12 cpf. Enfin, le pourcentage associé à [a] semble diminuer de manière constante entre les 3 filtres de FS.

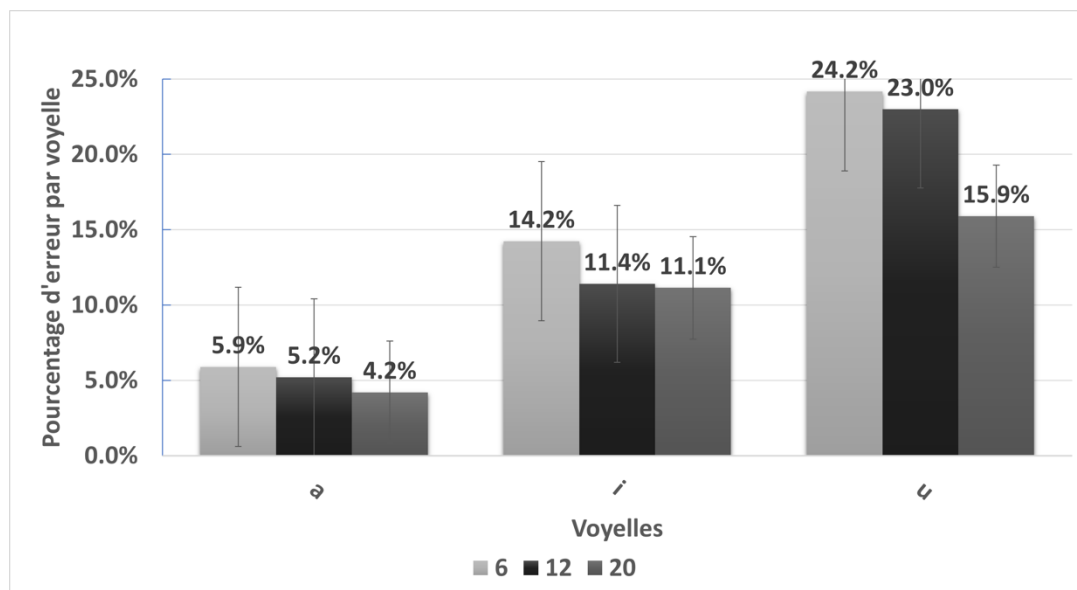


Figure 4.10 Pourcentage d'erreur par voyelle et fréquence spatiale et erreurs standards

4.4.3 Pourcentage d'erreur par syllabe

La Figure 4.11 donne les pourcentages d'erreur pour chaque syllabe pour les trois filtres de FS. Nous voyons ici qu'il y a une grande disparité entre les pourcentages d'erreur obtenus pour chaque syllabe. De plus, les pourcentages d'erreur ne sont pas toujours décroissants quand la valeur du filtre de FS augmente. Nous y reviendrons dans la discussion. Pour [pa, ka] les pourcentages d'erreur restent inférieurs à 1,5% pour toutes les valeurs de filtre. Pour [ta, pi, ki, tu, ku], le pourcentage d'erreur diminue quand la valeur du filtre augmente. La syllabe [pi] obtient un pourcentage d'erreur plus élevé pour le filtre à 6 cpf, alors que les deux autres filtres obtiennent des pourcentages d'erreur plus faibles à 4,9% et 4,8%. Pour [ti], le filtre à 12 cpf obtient le plus faible pourcentage d'erreur à 15,6%, par rapport aux filtres à 12 cpf (16,2%) et 20 cpf (17,5%). La syllabe [pu] a le plus haut pourcentage d'erreur pour le filtre intermédiaire à 12 cpf, puis pour le filtre à 6 cpf avec 10,2% et enfin à 20 cpf avec 6,3%. Pour terminer, l'item [ku] obtient des pourcentages d'erreur nettement plus élevés que les autres syllabes avec un pourcentage d'erreur minimale de 30,5% pour le filtre à 20 cpf (supérieurs aux pourcentages d'erreur des autres syllabes). Le filtre à 12 cpf obtient 37,5% d'erreur et le filtre à 6 cpf obtient 44,8% d'erreur. Il y a un effet statistiquement significatif de la FS sur les taux d'erreur des syllabes ($X^2 = 8,99$ $p < 0,05$, $ddl = 2$, $\epsilon^2 = 0,346$).

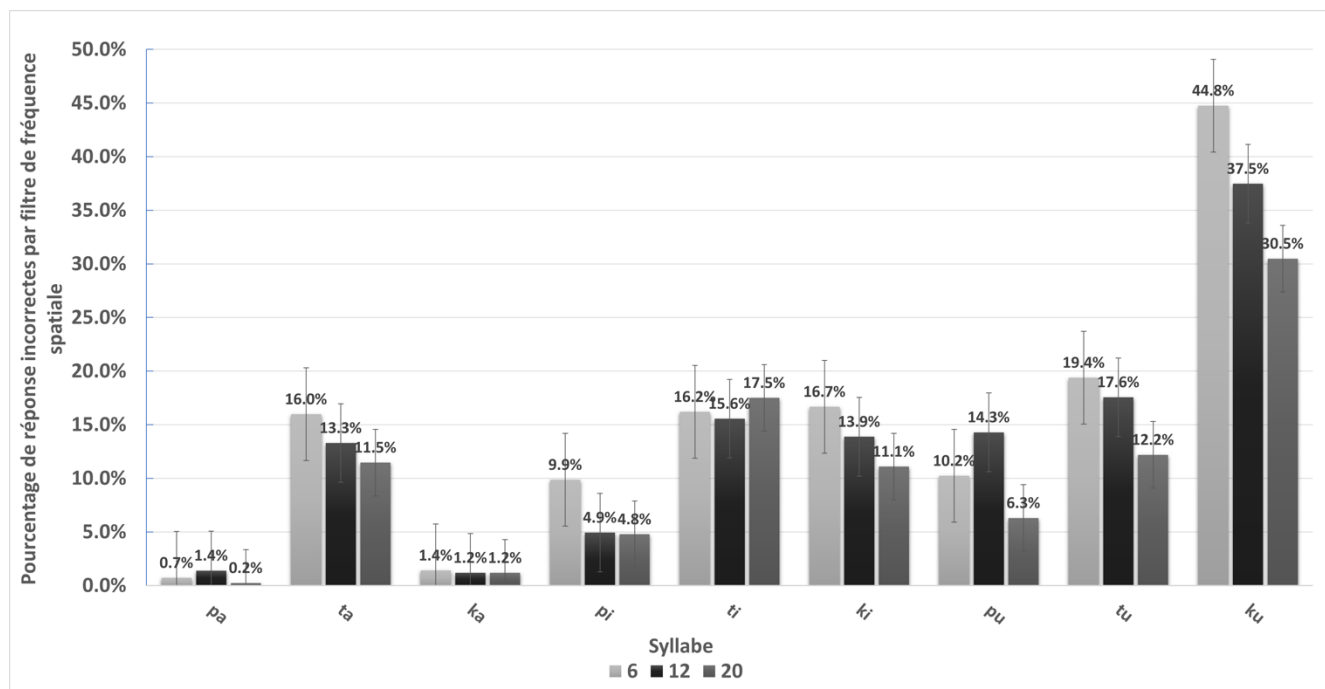


Figure 4.11 Pourcentage d'erreur par syllabe et fréquence spatiale et erreurs standards

4.5 Matrices de confusion des consonnes

Dans la section suivante, nous présentons les matrices de confusions obtenues pour les consonnes. Ces tableaux nous indiquent les profils d'erreurs obtenus lors de la réalisation de la tâche et nous permettent de comprendre le type de confusion existant entre les différents traits phonétiques des stimuli. Ainsi, nous pouvons savoir si le /t/ se confond plus avec le /p/ ou le /k/ par exemple et si la perturbation visuelle a un impact sur ces erreurs. Les phonèmes testés sont écrits horizontalement et les phonèmes donnés en réponses sont écrits verticalement. Par exemple dans le Tableau 4.5, les personnes participantes ont donné la réponse /t/ lorsque le stimulus contenait /k/ dans 17,1% des cas.

Tableau 4.5 Matrice de confusion par consonne pour toutes les fréquences spatiales

	Stimuli		
	p	t	k
p	93,0%	1,7%	1,4%
t	3,7%	83,4%	17,1%
k	3,3%	14,9%	81,5%

Le Tableau 4.5 est la matrice de confusion par consonne pour tous les filtres de FS. On voit ici que les personnes participantes ont répondu /t/ pour /k/ dans 17,1% des cas et /k/ pour /t/ dans 14,9% des cas. On comprend qu'il y a dans les conditions de l'expérience une confusion entre ces deux phonèmes. On voit aussi que le /p/ est moins confondu qu'avec les autres phonèmes (moins de 4% d'erreur). Les deux phonèmes /k/ et /t/ ne sont pas visibles du fait de leur point d'articulation respectif (vélaire et dental) et sont en conséquence plus souvent intervertis. Pour terminer, le /k/ est moins visible que le /t/ et obtient des scores de confusion avec /t/ plus élevé que l'inverse.

4.5.1 Matrice de confusion par fréquence spatiale

Dans cette section, nous regardons le même profil d'erreur, mais en filtrant les matrices de confusion par fréquence spatiale. La valeur du filtre est écrite en haut à gauche du tableau. Dans le Tableau 4.6, on retrouve le même profil d'erreur (confusion entre /k/ et /t/), mais avec des taux d'erreur un peu plus élevés. C'est ce que nous attendions, car c'est la plus forte perturbation visuelle (image la plus floue).

Tableau 4.6 Matrice de confusion par consonne au filtre de FS 6 cpf

FS 6	Stimuli		
	p	t	k
p	91,7%	3,3%	2,9%
t	4,8%	81,3%	19,2%
k	3,5%	15,5%	78,0%

Tableau 4.7 Matrice de confusion par consonne au filtre de FS 12 cpf

FS 12	Stimuli		
	p	t	k
p	92,2%	1,1%	0,8%
t	3,8%	83,7%	17,3%
k	4,0%	15,2%	81,9%

Pour, le filtre à 12 cpf (Tableau 4.7), on retrouve aussi le même profil d'erreur (confusion /k/ et /t/), avec des taux d'erreur un peu plus faibles que pour le filtre à 6 cpf.

Tableau 4.8 Matrice de confusion par consonne au filtre de FS 20 cpf

FS 20	Stimuli		
	p	t	k
p	95,0%	0,8%	0,5%
t	2,6%	85,1%	14,9%
k	2,4%	14,2%	84,5%

Enfin, pour la perturbation à 20 cpf (Tableau 4.8), filtre laissant passer le plus d'information visuelle, on obtient à nouveau le même profil d'erreur et des taux d'erreur un peu plus faibles que pour le filtre de FS 12 cpf.

4.6 Matrice de confusion des consonnes par contexte vocalique

Dans la section suivante, nous présentons les matrices de confusions obtenues pour les consonnes en contexte vocalique.

4.6.1 /a/ pour toutes les fréquences spatiales

Tableau 4.9 Matrice de confusion pour le contexte vocalique /a/

	Stimuli		
	pa	ta	ka
pa	98,6%	1,1%	0,1%
ta	0,9%	85,9%	1,6%
ka	0,5%	13,0%	98,3%

Dans le contexte vocalique [a] (Tableau 4.9), on observe la dynamique des taux erreurs est différentes de celles observées pour toutes les voyelles confondues dans la première matrice (voir Tableau 4.5). Les scores de reconnaissance (SDR) globaux pour /pa/ et /ka/ sont nettement plus haut que pour /ta/. Le /ta/ a un SDR plus faible à 82,7% et il est surtout confondu avec /ka/. Il semble que le contraste /p/ - /k/ dans ce contexte vocalique soit renforcé et ils sont très peu confondus l'un avec l'autre.

4.6.2 /a/ par filtre de fréquence spatiale

Tableau 4.10 Matrice de confusion pour le contexte vocalique /a/ au filtre de FS 6

FS 6	Stimuli		
	pa	ta	ka
pa	98,2%	2,5%	0,2%
ta	1,4%	82,7%	1,8%
ka	0,5%	14,8%	98,0%

Tableau 4.11 Matrice de confusion pour le contexte vocalique /a/ au filtre de FS 12

FS 12	Stimuli		
	pa	ta	ka
pa	98,6%	0,5%	0,0%
ta	0,9%	87,0%	1,8%
ka	0,5%	12,5%	98,2%

Tableau 4.12 Matrice de confusion pour le contexte vocalique /a/ au filtre de FS 20

FS 20	Stimuli		
	pa	ta	ka
pa	99,1%	0,5%	0,0%
ta	0,5%	88,0%	1,1%
ka	0,5%	11,6%	98,9%

L'examen des matrices de confusion par filtre de fréquence spatiale (Tableau 4.10, Tableau 4.11 et Tableau 4.12) pour le contexte vocalique /a/ nous permet de retrouver la diminution graduelle des pourcentages d'erreur avec l'augmentation des valeurs de filtre de FS. Par contre, seule la confusion du /ta/ par le /ka/ est prépondérante, la confusion inverse n'est presque pas observée.

4.7 Matrice de confusion pour /i/ toutes fréquences confondues

Tableau 4.13 Matrice de confusion pour le contexte vocalique /i/

	Stimuli		
	pi	ti	ki
pi	92,6%	1,1%	1,6%
ti	5,7%	83,5%	13,9%
ki	1,7%	15,5%	84,5%

Pour le /i/ (Tableau 4.13), on retrouve principalement la confusion de /t/ par /k/ (15,5%) et de /k/ par /t/ (13,9%) que l'on obtenait dans la matrice de confusion globale (Tableau 4.5). On retrouve aussi la confusion /ti/ pour /ki/ (5,7%). Contrairement au /a/, le /i/ s'inscrit dans la tendance générale observée pour les consonnes avec une plus grande confusion pour les consonnes articulées en arrière de la bouche et non visibles (/t/ et /k/) et une légère confusion entre les consonnes articulées dans la zone antérieure du palais (/p/ et /t/).

4.7.1 Matrice de confusion pour /i/ par filtre de fréquence spatiale

Tableau 4.14 Matrice de confusion pour le contexte vocalique /i/ au filtre de FS 6

FS 6		Stimuli	
réponses	pi	ti	ki
pi	89,3%	0,9%	3,0%
ti	8,4%	83,9%	15,2%
ki	2,3%	15,2%	81,8%

Tableau 4.15 Matrice de confusion pour le contexte vocalique /i/ au filtre de FS 12

FS 12		Stimuli	
réponses	pi	ti	ki
pi	94,5%	1,6%	0,9%
ti	3,6%	84,3%	13,6%

ki	1,8%	14,1%	85,5%
-----------	-------------	--------------	--------------

Tableau 4.16 Matrice de confusion pour le contexte vocalique /i/ au filtre de FS 20

FS 20	Stimuli		
réponses	pi	ti	ki
pi	93,9%	0,7%	0,9%
ti	5,0%	82,3%	13,0%
ki	1,1%	17,0%	86,1%

Comme attendu, les résultats pour /i/ (Tableau 4.14, Tableau 4.15 et Tableau 4.16) par filtre de FS montrent des effets similaires pour toutes les FS confondues avec une diminution des pourcentages de confusion quand la valeur du filtre de FS passe de 6 à 12. Cependant, on observe une augmentation du pourcentage de confusion de /pi/ par /ti/ et de /ti/ par /ki/ pour le filtre 20, par rapport au filtre 12. La confusion de /ti/ par /ki/ (17%) y est même plus importante que pour le filtre 6. Il semble que pour le contexte vocalique /i/ le filtre de fréquence spatiale 20 augmente la confusion pour certaines consonnes.

4.8 Matrice de confusion pour /u/ par filtre de fréquence spatiale

Tableau 4.17 Matrice de confusion pour le contexte vocalique /u/

réponses	Stimuli		
	pu	tu	ku
pu	87,7%	3,0%	2,5%
tu	4,5%	80,7%	35,9%
ku	7,7%	16,4%	61,6%

Dans le contexte vocalique /u/ (Tableau 4.17), on observe des scores de réussite 2 fois plus faibles pour la consonne /k/. La syllabe /ku/ est confondue avec la syllabe /tu/ dans 35,9% des cas. Il semble que ce contexte vocalique rende la perception des consonnes dont le point d'articulation est non visible plus difficile. On observe aussi que le /pu/ est confondu avec le /ku/ dans 7,7% des cas et avec le /tu/ dans 4,5% des cas.

4.8.1 Matrice de confusion pour /u/ par filtre de fréquence spatiale

Tableau 4.18 Matrice de confusion pour le contexte vocalique /u/ au filtre de FS 6

FS 6		Stimuli		
réponses	pu	tu	ku	
pu	87,7%	6,4%	5,5%	
tu	4,5%	77,3%	40,5%	
ku	7,7%	16,4%	54,1%	

Tableau 4.19 Matrice de confusion pour le contexte vocalique /u/ au filtre de FS 12

FS 12		Stimuli		
réponses	pu	tu	ku	
pu	83,4%	1,4%	1,4%	
tu	6,8%	79,8%	36,6%	
ku	9,8%	18,9%	62,0%	

Tableau 4.20 Matrice de confusion pour le contexte vocalique /u/ au filtre de FS 20

FS 20 réponses	Stimuli		
	pu	tu	ku
pu	92,0%	1,1%	0,7%
tu	2,3%	85,0%	30,7%
ku	5,7%	13,9%	68,6%

Enfin, les résultats pour /u/ par fréquence spatiale (*Tableau 4.18 ; Tableau 4.19 ; Tableau 4.20*) montrent que la confusion du /ku/ avec le /tu/ est la plus fréquente et qu'elle diminue quand les filtres de FS augmentent. Pour la confusion du /tu/ avec le /ku/, la valeur la plus forte se trouve au filtre de FS 12 (18,9%). On retrouve cette tendance pour la confusion du /pu/ avec les deux autres consonnes /k/ et /t/, avec respectivement des pourcentages de confusion maximums au filtre de FS 12 à 6,8% pour /tu/ et 9,8% pour /ku/. Les taux d'erreur et les confusions sont plus importants pour le /u/ que pour les autres contextes vocaliques et la visibilité du point d'articulation pour /p/ par rapport à /t/ et /k/ semblent moins bien perçue. Ceci est probablement lié à l'arrondissement des lèvres dans l'articulation du /u/ (voyelle labiale) qui ajoute par coarticulation une certaine labialité même aux consonnes articulées postérieures.

Tableau 4.21 Confusions inter-consonnes les plus fréquentes par contexte vocalique

	Contexte /Ca/			Contexte /Ci/			Contexte /Cu/		
	FS6	FS12	FS20	FS6	FS12	FS20	FS6	FS12	FS20
p	t	t-k	t-k	t	t	t	k	k	k
t	k	k	k	k	k	k	k	k	k
k	t	t	t	t	t	t	t	t	t

Si on regarde les confusions les plus fréquentes entre les consonnes par contexte vocalique, un patron émerge. Les consonnes ne se confondent pas les unes avec les autres de manière arbitraire. Pour le /a/, pour le filtre le plus bas (les autres filtres n'ayant presque pas d'erreur), le /p/ est confondu avec le /t/, le /t/ avec le /k/ et le /k/ avec le /t/. Pour le /i/, le /p/ est seulement confondu avec le /t/, mais le /t/ est seulement confondu avec le /k/ et inversement. Enfin, pour le /u/, le /p/ est seulement confondu avec le /k/, mais le /k/ est seulement confondu avec le /t/ et inversement.

CHAPITRE 5 DISCUSSION

5.1 Rappel des objectifs et questions de recherche

Le but de ce travail de recherche était d'explorer l'intégration audiovisuelle de la parole quand l'information visuelle et le son sont dégradés (perturbés). Afin de nous assurer que la vision était bien impliquée dans la perception de la parole et pour éviter les effets plafonds, nous avons perturbé le son avec du bruit rose à -18 dB SNR. Ce protocole expérimental nous a permis d'explorer et de décrire l'effet de la perturbation visuelle sur le décodage des unités de parole (syllabes, consonnes et voyelles). Pour mesurer les performances à la tâche des personnes participantes, nous avons mesuré les scores de reconnaissance (SDR), analysé les profils d'erreur et étudié l'amélioration ou la dégradation des SDR apportée par nos 3 niveaux de perturbation visuelle (gain perceptif). Nos résultats montrent que la perturbation visuelle entraîne une baisse statistiquement significative des SDR liée au niveau de perturbation. En somme, plus le filtre de fréquence spatiale (FS) est bas et plus les SDR sont faibles, ce qui confirme les résultats d'études antérieures. De plus, la nature des syllabes, des consonnes et des voyelles module cette diminution des SDR en contexte bruité. En effet, nous avons relevé des effets d'interaction statistiquement significatifs entre les variables *syllabe* et *filtre de FS* et entre les variables *filtre de FS* et *voyelle*. Nos données démontrent l'importance du contrôle des paramètres phonétiques des stimuli lorsqu'on étudie l'intégration audiovisuelle de la parole (IAVP), car les phonèmes sont les premiers éléments influençant la distribution des résultats et il serait très difficile d'analyser leur impact s'ils étaient intégrés dans un stimulus plus complexe comme une phrase.

Pour rappel, nos questions et hypothèses de recherche étaient les suivantes :

Q.1 Quel est l'impact d'une perturbation visuelle sur la perception de la parole en condition audio bruitée/perturbée ? (Binnie *et al.*, 1974 ; Campbell et Massaro, 1997 ; Dickinson et Taylor, 2011 ; Legault *et al.*, 2010 ; Munhall *et al.*, 2004 ; Robert-Ribes *et al.*, 1998 ; Schwartz *et al.*, 2012 ; Sumby et Pollack, 1954 ; Tye-Murray *et al.*, 2016)

- **H.1** Plus la valeur des filtres de FS sera faible (et donc plus la perturbation sera importante), plus les scores de reconnaissance seront faibles.

Q.2 Y-a-t-il des traits phonétiques qui résistent mieux à la perturbation visuelle que d'autres : lieu d'articulation, type de phonème (voyelles, consonnes) ou encore mode d'articulation des voyelles (arrondissement)? (Binnie *et al.*, 1974 ; Robert-Ribes *et al.*, 1998 ; Schwartz *et al.*, 2012 ; Sumbly et Pollack, 1954)

- **H.2** *La perturbation aura des effets différents selon le domaine et le type de phonème sur lequel elle est appliquée (syllabe vs phonème, et voyelle vs consonne). La saillance articulatoire des traits phonétiques influencera les scores de reconnaissance* (Binnie *et al.*, 1974 ; Macleod et Summerfield, 1987 ; Robert-Ribes *et al.*, 1998)

5.2 Questions de recherche 1

Nos résultats montrent que plus la valeur du filtre de FS est basse et plus les scores de reconnaissance sont faibles. Les scores de reconnaissance sont présentés dans le chapitre 25 Revenons maintenant sur l'hypothèse que nous avons posée :

- **H.1** *Plus la valeur des filtres de FS sera faible, plus la perturbation sera importante et plus les scores de reconnaissance seront faibles.*

Cette hypothèse est confirmée, si l'on regarde les résultats d'un seul tenant. Généralement, plus la valeur du filtre de FS est faible, et donc, plus la perturbation visuelle est forte (plus l'image est « floue »), plus les scores de reconnaissance sont faibles. En cela, nos résultats confirment les tendances générales notées dans les études antérieures. Nous observons les scores de reconnaissance moyens les plus hauts pour le filtre à 20 cpf, puis à 12 cpf et enfin à 6 cpf. L'effet est confirmé statistiquement entre les filtres de FS 6 cpf et 20 cpf et 12 cpf ET 20 cpf. Plus le stimulus vidéo est net, plus les personnes participantes réussiront à décoder des informations visuelles leur permettant de reconstituer la syllabe entendue. En condition visuelle perturbée, le degré de perturbation influence les SDR et donc la facilité de perception de la parole. De plus, le gain perceptif pour le filtre de FS 20 est plus important que pour le filtre de FS 12 (Figure 4.5). Ces résultats montrent que l'information visuelle est primordiale pour la compréhension de la parole quand l'audio est bruité. Plus l'information visuelle est grossière (filtre de FS le plus bas) et plus les SDR sont faibles. On peut donc s'attendre à ce qu'une perte d'information visuelle (trouble de la vision, faible luminosité ou encore port d'un masque par exemple), dans un milieu où le son est aussi perturbé, ait un impact sur la capacité de décodage de la parole d'une personne. Ces résultats corroborent ceux de la

littérature que nous avons consultée sur la difficulté relative de perception d'un stimulus de parole en condition visuelle perturbée (Callan *et al.*, 2004 ; Dickinson et Taylor, 2011 ; Legault *et al.*, 2010 ; Massaro et Cohen, 1995).

L'interprétation de nos résultats dans la perspective de cette question de recherche est générale. Nous parlons d'image plus nette ou plus floue, mais quand on regarde l'interaction entre les VI (phonème ou syllabe) et les filtres de FS, d'autres tendances se dégagent. En effet, le filtrage des fréquences spatiales supprime l'information de manière contrôlée et les résultats ne s'améliorent pas toujours quand il y a plus d'informations visuelles.

5.3 Question de recherche 2

Q.2 Y-a-t-il des traits phonétiques qui résistent mieux à la perturbation visuelle que d'autres : lieu d'articulation, type de phonème (voyelles, consonnes) ou encore mode d'articulation des voyelles (arrondissement)? (Binnie *et al.*, 1974 ; Robert-Ribes *et al.*, 1998 ; Schwartz *et al.*, 2012 ; Sumby et Pollack, 1954)

- **H.2** *La perturbation aura des effets différents selon le domaine et le type de phonème sur lequel elle est appliquée (syllabe vs phonème, et voyelle vs consonne). La saillance articulatoire des traits phonétiques influencera les scores de reconnaissance* (Binnie *et al.*, 1974 ; Macleod et Summerfield, 1987 ; Robert-Ribes *et al.*, 1998)

L'analyse graphique et statistique de nos résultats ainsi que l'analyse des matrices de confusion confirment cette hypothèse. En effet, lorsqu'on regarde les résultats par consonne, par contexte vocalique et par syllabe, on observe des effets spécifiques de chaque variable sur les SDR.

5.3.1 Interaction entre les filtres de FS et les consonnes

La perturbation visuelle entraîne une baisse des SDR moyens et une augmentation avec la valeur du filtre de FS qui varie selon la consonne. Les effets sont corroborés graphiquement et statistiquement (Tableau 4.6, Tableau 4.7 et Tableau 4.8), mais l'interaction entre les variables *filtre de FS* et *consonne* n'est pas statistiquement significative. La consonne /p/ (bilabiale), qui est plus visible, obtient les meilleurs SDR, suivi de la consonne /t/ (alvéolaire) et finalement de la consonne /k/ (palatale), qui est le moins visible (Figure 4.2, Figure 4.6 et Figure 4.9). La matrice de confusion des consonnes, tous filtres de FS confondus,

appuie aussi ces résultats (Tableau 4.5). Moins le point d'articulation de la consonne est visible et plus il y a de confusion. De plus, peu importe le contexte vocalique, la confusion pour la consonne /k/ se fait surtout avec la consonne /t/ et inversement (voir Tableau 4.21). En somme, les deux consonnes les moins visibles sont surtout confondues entre elles. Pour les matrices de confusion par filtre de FS pour les consonnes, plus le filtre de FS est bas (moins d'information visuelle disponible) et plus la confusion entre les consonnes est fréquente (Tableau 4.6, Tableau 4.7 et Tableau 4.8). L'analyse du gain perceptif (GP) par consonne vient aussi appuyer notre analyse. Les filtres de FS 12 cpf et 20 cpf améliorent la perception des consonnes par rapport au filtre à 6 cpf. On remarque que le GP est plus important pour les filtres à 12 et 20 cpf quand la consonne est moins visible (/t/ et /k/) et que le GP est plus important pour le filtre à 20 cpf pour toutes les consonnes. De plus, le filtre de FS à 12 cpf n'apporte presque pas de GP pour le /p/ (0,2%) contrairement au filtre à 20 cpf (3,1%). La consonne /p/ étant la consonne la plus visible, on peut se demander si les fréquences plus élevées du filtre à 20 cpf, qui contiennent des informations visuelles plus fines, sont spécifiques à sa perception et à la perception du mouvement des lèvres, par opposition aux mouvements plus grossiers de la mâchoire qui seraient liés à des FS plus basses (Campbell et Massaro, 1997). Le GP de la consonne /k/ est le plus important pour les deux filtres de FS alors qu'elle est la moins visible. Cela peut être lié au fait qu'elle est moins susceptible aux effets plafonds, car elle a les SDR par FS les plus faibles et que l'information visuelle apportée par les filtres de FS plus élevés permet de mieux discriminer les deux autres consonnes et ipso facto de réduire le taux d'erreur du /k/. Ces données posent la question de la perception directe ou indirecte des consonnes. Dans un contexte donné, l'information perçue est-elle aussi importante que l'absence d'information ? Par exemple, pour déduire le /k/, le cerveau exploiterait l'absence d'information visuelle et éliminerait les probabilités de /p/ et /t/. Voyons maintenant comment se comportent les SDR dans différents contextes vocaliques.

5.3.2 Interaction entre les filtres de FS et les voyelles

L'effet des filtres de FS et de la voyelle sur les SDR est corroboré graphiquement et statistiquement. De plus, l'interaction entre les variables *filtre de FS* et *voyelle* est statistiquement significative. Cependant, l'interprétation de cet effet est plus complexe que pour les consonnes. Pour mieux comprendre cela, il faut revenir sur l'articulation des phonèmes. Les voyelles sont articulées par la modulation de la résonance des cavités orales et labiale grâce à la langue, la mandibule et les lèvres. En comparaison, on articule les consonnes par constriction ou blocage du flux d'air à l'aide des articulateurs fixes et mobiles et de la langue. Il n'y a donc pas, pour les voyelles, de point d'articulation facilement identifiable comme pour les consonnes. Seules la labialité et l'ouverture relative de la bouche sont perceptibles visuellement. De plus,

il est bien connu en phonétique que la production des voyelles peut varier d'une personne à l'autre et que les données acoustiques pures ne suffisent pas à les caractériser (Maddieson, 1984 ; Trudeau-Fisette, 2015). Par conséquent, il faut aussi prendre en compte la position relative des voyelles (caractérisée par les valeurs des formants F1, F2 et F3) les unes par rapport aux autres dans le trapèze vocalique (Figure 3.1). C'est pour cette raison que nous avons choisi 3 voyelles séparées par 3 paramètres de contraste : l'aperture de la bouche, l'antériorité de la langue et l'arrondissement des lèvres. En regardant les résultats obtenus par contexte vocalique (Figure 4.3), on observe un écart de SDR moyen plus important que pour les consonnes avec une amplitude d'environ 25 points de SDR. Comme pour les consonnes, les voyelles se distinguent toutes de manière statistiquement significative les unes des autres. Le /u/ qui est la seule voyelle avec un trait articulatoire visible (l'arrondissement des lèvres) et qui est articulé en arrière de la bouche avec les articulateurs 'fermés' est la voyelle qui obtient les SDR moyens les plus faibles. Il semble donc logique qu'il soit le plus impacté par la perturbation visuelle. Cependant, contrairement à la consonne /p/, le /u/ bien que visible partiellement ne facilite pas la perception. Le /i/ qui est articulé avec les articulateurs en position haute et en avant obtient la deuxième place, puis le /a/ qui est ouvert et postérieur obtient les meilleurs SDR. L'aperture semble être le paramètre le plus facilement perçu, suivi de l'antériorité. Pour comprendre l'impact de l'arrondissement, il faut regarder le gain par voyelle (Figure 4.7). Nous voyons que pour le /u/, le filtre à 20 cpf a le GP le plus important (8,3%) par rapport au filtre à 12 cpf (1,2%) et par rapport aux autres voyelles. Les fréquences hautes qui portent les détails plus fins sont donc nécessaires à sa perception et les informations visuelles qui sont liées à cette voyelle ne sont perçues que quand les FS sont suffisamment hautes pour permettre la discrimination. La labialité du /u/ doit donc être liée au FS plus hautes. On retrouve une tendance similaire pour la consonne bilabiale /p/ qui a un GP faible entre les filtres 6 cpf et 12 cpf et qui voit son GP augmenté de 0,2% à 3,1% entre les filtres à 12 cpf et 20 cpf. Pour les voyelles /a/ et /i/ le filtre à 20 cpf n'apporte pas beaucoup d'informations (GP $\leq$ 1% entre les filtres à 12 et 20 cpf). Il semble que les FS plus basses (inférieures à 6 cpf et 12 cpf) portent suffisamment d'informations relatives au /a/ et au /i/ pour faciliter leur perception. En effet, les informations contenues dans les fréquences spatiales plus basses seraient plutôt structurelles (Campbell et Massaro, 1997) et on peut penser qu'elles portent les informations liées à la position de la mâchoire. Si l'on regarde les données de Robert-Ribes *et al.* (1998), les traits vocaliques ont une hiérarchie de perception qui varie avec la modalité de perception :

Modalité audio seule : aperture > antériorité > arrondissement

Modalité visuelle seule : arrondissement > aperture > antériorité

Nos données sont concordantes avec ces hiérarchies de perception. En effet, si l'arrondissement est le trait le mieux perçu visuellement, il sera donc le plus sensible à une perturbation visuelle. À cela s'ajoute sa difficulté de perception acoustique. En résumé, le /u/ qui est une voyelle complexe acoustiquement et avec des traits visuels fins est donc plus dépendante que le /i/ et le /a/ aux informations visuelles fines pour sa perception et est plus affecté par le filtrage passe-bas des FS des deux premiers filtres. Le /a/ semble moins dépendre de la vision et son GP ne dépasse pas les 1,7% avec le filtre de FS le plus haut. On sait par ailleurs qu'il résiste le mieux à la perturbation acoustique et les FS basses sont suffisantes pour percevoir son aperture (Campbell et Massaro, 1997 ; Robert-Ribes *et al.*, 1998). Pour terminer, le /i/ a une différence de GP plus importante entre les filtres 6 cpf et 12 cpf qu'entre les filtres 12 cpf et 20 cpf. On peut en déduire que les FS inférieures à 12 cpf contiennent l'information visuelle nécessaire à sa perception. Pour conclure, l'analyse des données relatives aux voyelles a mis en évidence deux groupes, la voyelle /u/ plus visuelle et nécessitant des FS plus hautes pour sa perception et les voyelles /a/ et /i/ qui sont moins sensibles à la perturbation audio et qui dépendent des FS plus basses pour leur perception visuelle.

5.3.3 Interaction entre les filtres de FS et les consonnes en contexte vocalique

Pour clore cette section, analysons maintenant les patrons de confusions entre les consonnes (Tableau 4.21). L'interaction de la consonne et de la voyelle dans la syllabe rend l'interprétation des résultats plus complexe. Rappelons-nous des principales tendances observées pour les consonnes et pour les voyelles. Le /p/ facilitait la perception et obtenait les plus hauts SDR pour tous les filtres de FS. Le /k/ était plus difficilement reconnu, mais avait le plus haut GP et le /t/ se situait entre les deux pour toutes ces mesures. Concernant les voyelles, le /a/ était le mieux identifié suivi du /i/, et le /u/ se détachait avec des SDR plus faibles que les deux autres voyelles. Le /u/ bénéficiait du plus haut GP, mais seulement à partir du filtre de FS 20 cpf. Le /i/ profitait d'un GP dès le filtre à 12 cpf et ne progressait pas beaucoup avec le filtre à 20 cpf. Enfin, le GP de /a/ n'était que peu impacté par les deux plus hauts filtres de FS. Considérant ces données, nous voyons que les syllabes /pV/ obtiennent les meilleurs SDR. En revanche, elles ont les GP les plus faibles voire un GP négatif (-4,1% pour /pu/ au filtre à 12 cpf) sauf pour le contexte /i/ où le filtre à 12 cpf apporte le plus important GP. Les syllabes /tV/ ont un SDR quasi nul ou négatif pour le contexte /i/, mais croissant pour les deux autres voyelles avec des GP plus importants au filtre à 20 cpf. Les syllabes en /kV/ ont les SDR les plus hauts en contexte /a/ avec très peu de GP pour les filtres 12 cpf et 20 cpf. Le contexte

/i/ voit son GP augmenté de manière constante pour les filtres de FS 12 cpf et 20 cpf. Enfin, le contexte vocalique /u/ bénéficie des SDR les plus faibles, mais des GP les plus importants. Le GP est selon nous l'indice qui donne le plus d'information sur l'intégration des données visuelles de la parole. Si le contexte /pV/ semble améliorer les SDR, il limite le GP voire détériore la perception de la syllabe. Cela se voit dans le contexte /u/ où le filtre à 12 cpf rend la syllabe plus difficile à percevoir. La syllabe /pi/ en revanche bénéficie surtout des informations en dessous de 12 cpf, tout comme la voyelle /i/ prise séparément. Il semble que la perception des syllabes puisse être plus influencée par la consonne ou par la voyelle selon les cas. Pour terminer, le contexte /u/ qui est le plus influencé par la perturbation visuelle rend la perception du /k/ encore plus difficile avec des SDR moyens plus faibles. Cependant, tout comme quand ils sont pris séparément, ces deux phonèmes induisent le GP le plus important. Ces données mettent en évidence l'influence des phonèmes l'un sur l'autre et sur la perception de la syllabe. Les contextes vocaliques influenceront dans une plus grande mesure les consonnes les moins visibles et le /u/ qui porte un trait phonétique visible aura l'effet le plus fort sur la perception de la syllabe. L'influence des consonnes sera liée à l'antériorité de leur point d'articulation et la consonne /p/ qui, comme le /u/ est labiale, aura l'influence la plus notoire sur la perception des syllabes. On voit dans la syllabe /pu/ une diminution de -4,1% GP pour le filtre de FS 12 cpf, puis une augmentation de 3,9%. De plus, les SDR moyens sont parmi les plus hauts pour le /pu/ toutes syllabes confondues. Cela illustre l'influence forte des deux phonèmes labiaux constituant la syllabe qui bénéficie d'un GP avec les FS les plus hautes qui portent les détails fins.

5.3.4 Confusion inter-consonne et taux d'erreur

Les confusions inter-consonnes sont influencées par le contexte vocalique. Dans le contexte /u/, le /p/ est surtout confondu avec le /k/, dans les contextes /i/ et /a/, il est plutôt confondu avec le /t/. Le /t/ est surtout confondu avec le /k/ dans tous les contextes vocaliques. Enfin, le /k/ est surtout confondu avec le /t/ dans tous les contextes vocaliques. On voit ici que le contexte vocalique peut influencer le type d'erreur. Le /i/ qui est articulé proche du /t/ entraînera la perception d'un /t/ au lieu d'un /p/ et le /u/ qui est labial entraînera la perception d'un /k/ au lieu d'un /p/. Encore une fois, la voyelle /u/ et la consonne /p/ labiales influencent la perception de l'autre phonème de la syllabe.

5.4 Résumé des principales observations

Nous avons observé l'influence de la perturbation sur les phonèmes labiaux et il en ressort que le /u/ et le /p/ bénéficient le plus de l'augmentation de la valeur du filtre de FS à 20 cpf. Ceci se reflète dans les taux d'erreur et les GP que nous avons calculés. La perception des autres phonèmes ne portant pas de trait

phonétique visible est elle aussi influencée par la perturbation visuelle, mais de manière moins évidente. La perception du /i/ semble plus bénéficier du passage au filtre à 12 cpf qu'à 20 cpf. Et le /a/ est légèrement mieux perçu avec chaque augmentation de filtre de FS. Le /t/ est le /k/ semblent eux plus bénéficier du filtre à 12 cpf. Concernant les syllabes, leur perception et leur sensibilité à la perturbation sont liées aux phonèmes les constituant, avec des effets plus évidents de /p/ et de /u/. Dans les autres cas, les matrices de confusions montrent que la saillance articulatoire des traits phonétiques influence la perception, si le stimulus porte une voyelle ouverte il sera mieux perçu qu'une voyelle fermée et si le point d'articulation de la consonne est plus antérieur, elle sera un peu mieux perçue.

5.5 Limites de notre recherche

Lors de la réalisation de ce travail de recherche, nous avons rencontré différents obstacles pratiques et méthodologiques. Tout d'abord, nous n'avons pas pu faire passer la tâche dans un environnement contrôlé et le fait que chaque personne participante utilisait un matériel différent pour la passation peut avoir influencé nos données. Pour contrôler cela, nous avons standardisé le test et fait en sorte que l'aspect visuel soit le même pour toutes les personnes participantes. Nous avons aussi vérifié statistiquement, grâce aux variables aléatoires, l'influence du matériel et les variations inter-individus. Deuxièmement, la longueur de la tâche nous a limités dans le nombre de filtres de fréquence spatiale. Il serait intéressant de filtrer l'image de différentes façons, un filtre passe-haut par exemple, qui supprime d'abord les fréquences spatiales basses et de voir l'impact que cela pourrait avoir sur la perception des voyelles. Le fait que nous soyons les premiers à avoir étudié l'effet d'une perturbation visuelle sur la perception des phonèmes nous a limités, car il y avait peu de littérature sur laquelle s'appuyer et nous ne pouvions que comparer nos résultats à des travaux se rapprochant du nôtre. Cela limite la portée de nos interprétations et ouvre donc la voie à des travaux plus aboutis sur le sujet.

CONCLUSION

L'objectif de ce mémoire était d'explorer les effets d'une perturbation visuelle sur la perception de la parole quand le son est dégradé. Ce paradigme de double perturbation forçait l'utilisation de la vision avec une perturbation audio fixe et a permis de mesurer l'effet de nos trois niveaux de perturbation visuelle sur la perception des 9 syllabes de notre corpus. N'ayant que peu de littérature sur laquelle nous baser pour construire notre tâche, nous avons créé une tâche de perception originale réalisable à distance (contexte COVID). Nous avons choisi les phonèmes les plus pertinents en nous basant sur leur fréquence dans les langues du monde et les contrastes phonétiques qu'ils avaient les uns avec les autres. Pour notre perturbation visuelle, nous nous sommes limités à 3 niveaux de perturbation en nous basant sur des travaux utilisant le filtrage des fréquences spatiales dans des tâches de perception du visage et de la parole. Nos choix étaient faits dans l'objectif de créer un travail de recherche donnant un point de départ aux futurs travaux. Par conséquent, nos questions de recherche étaient simples. Pour rappel :

Q.1 Quel est l'impact d'une perturbation visuelle sur la perception de la parole en condition audio bruitée/perturbée ? (Binnie *et al.*, 1974 ; Campbell et Massaro, 1997 ; Dickinson et Taylor, 2011 ; Legault *et al.*, 2010 ; Munhall *et al.*, 2004 ; Robert-Ribes *et al.*, 1998 ; Schwartz *et al.*, 2012 ; Sumby et Pollack, 1954 ; Tye-Murray *et al.*, 2016)

Q.2 Y-a-t-il des traits phonétiques qui résistent mieux à la perturbation visuelle que d'autres : lieu d'articulation, type de phonème (voyelles, consonnes) ou encore mode d'articulation des voyelles (arrondissement)? (Binnie *et al.*, 1974 ; Robert-Ribes *et al.*, 1998 ; Schwartz *et al.*, 2012 ; Sumby et Pollack, 1954)

Les données que nous avons obtenues valident les hypothèses que nous avons posées. Le filtrage passe-bas des fréquences spatiales rend plus difficile la perception des syllabes et plus le filtre est bas (image moins nette) et plus la perception est difficile. Nous avons aussi observé que les traits phonétiques des phonèmes réagissent différemment à la perturbation. Nous avons constaté que les phonèmes portant le trait articulatoire visible de labialité sont plus sensibles à la perturbation visuelle et obtiennent un gain perceptif plus important. Le /p/ facilite la perception des phonèmes alors que le /u/ la rend plus difficile en entraînant des confusions entre le /p/ et le /k/. Les voyelles /a/ et /i/ facilitent la perception et on peut penser que les données visuelles présentes aux filtres les plus bas sont suffisantes pour aider à les

percevoir. En général, les voyelles influencent la perception des consonnes et entraînent des profils d'erreurs spécifiques. Nous avons conclu de ces résultats que les phonèmes contenant un trait articulatoire visible bénéficient plus des fréquences spatiales hautes qui portent des informations visuelles fines comme les petits mouvements des lèvres et qu'ils sont plus impactés par une perturbation visuelle. Ces résultats montrent que la vision est essentielle à la perception de la parole dans la continuité de la littérature sur le sujet. On peut se demander si ces conclusions s'étendent aux autres phonèmes portant un trait articulatoire visible. De plus, il serait intéressant de voir comment sont confondues les voyelles entre elles en utilisant un contexte consonantique plutôt que vocalique. Nos résultats vont aussi dans le sens des travaux sur l'acquisition de la parole qui montrent l'importance de la vision dans ce processus (Kuhl *et al.*, 2005). Ils montrent aussi que dans des cas pathologiques de perte d'acuité visuelle, il ne faut pas sous-estimer l'impact sur la perception des phonèmes. Pour terminer, il serait intéressant de voir ce qu'il en est dans des langues autres que le français. Comment réagiraient des traits phonétiques saillants visuellement s'opposant selon des traits phonétiques qui diffèrent du français ? En conclusion, nos données offrent des perspectives d'investigation pour les travaux futurs sur l'intégration audiovisuelle de la parole.

BIBLIOGRAPHIE

- Arkebauer Herbert J., Hixon Thomas J. et Hardy James C. (1967). Peak intraoral air pressures during speech. *Journal of Speech and Hearing Research*, 10(2), 196-208.
<https://doi.org/10.1044/jshr.1002.196>
- Benguerel, A.-P. et Cowan, H. A. (1974). Coarticulation of upper lip protrusion in french. *Phonetica*, 30(1), 41-55.
- Benoit, C., Mohamadi, T. et Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of french. *Journal of Speech, Language, and Hearing Research*, 37(5), 1195-1203.
- Benoît Christian, Mohamadi Tayeb et Kandel Sonia. (1994). Effects of phonetic context on audio-visual intelligibility of french. *Journal of Speech, Language, and Hearing Research*, 37(5), 1195-1203.
<https://doi.org/10.1044/jshr.3705.1195>
- Binnie, C., Montgomery, A. et Jackson, P. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research*, 17(4), 619-630.
<https://doi.org/10.1044/jshr.1704.619>
- Boersma, P. et Weenink, D. (2021). *Praat: doing phonetics by computer* (version 6.1.43).
<http://www.praat.org/>
- Callan, D., Jones, J., Munhall, K., Kroos, C., Callan, A. et Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience*, 16, 805-816.
<https://doi.org/10.1162/089892904970771>
- Calvert et Thesen. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *Journal of Physiology*, 98(1), 191-205.
<https://doi.org/10.1016/j.jphysparis.2004.03.018>
- Campbell, C. et Massaro, D. (1997). Perception of visible speech: influence of spatial quantization. *Perception*, 26, 627-644. <https://doi.org/10.1068/p260627>
- Connine, C., Titone, D. et Wang, J. (1993). Auditory word recognition: extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 19, 81-94. <https://doi.org/10.1037/0278-7393.19.1.81>
- Dennedy, D. (2021). *Shotcut* (version 21.05.18). Meltytech LLC.
- Dickinson, C. M. et Taylor, J. (2011). The effect of simulated visual impairment on speech-reading ability. *Ophthalmic and Physiological Optics*, 31(3), 249-257. <https://doi.org/10.1111/j.1475-1313.2010.00810.x>
- Jamovi. (2023). *The jamovi project* (version 2.2.5). <https://www.jamovi.org/>

- Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T. et Pruitt, J. (2005). Early speech perception and later language development: implications for the « critical period ». *Language Learning and Development*, 1(3-4), 237-264. <https://doi.org/10.1080/15475441.2005.9671948>
- Kutas, M. et Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, 207(4427), 203 LP - 205. <https://doi.org/10.1126/science.7350657>
- Legault, I., Gagné, J.-P., Rhoualem, W. et Gosselin, P. (2010). The effects of blurred vision on auditory-visual speech perception in younger and older adults. *International Journal of Audiology*, 49, 904-911. <https://doi.org/10.3109/14992027.2010.509112>
- Liben-Nowell, D., Strand, J., Sharp, A., Wexler, T. et Woods, K. (2019). The danger of testing by selecting controlled subsets, with applications to spoken-word recognition. *Journal of Cognition*, 2, 2. <https://doi.org/10.5334/joc.51>
- Macleod, A. et Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, 21(2), 131-141. <https://doi.org/10.3109/03005368709077786>
- Maddieson, I. (1984). *Patterns of sounds*. Cambridge University Press.
- Malécot, A. et Metz, G. (2009). Progressive nasal assimilation in french. *Phonetica*, 26(4), 193-209. <https://doi.org/10.1159/000259410>
- Massaro, D. W. et Cohen, M. M. (1995). Perceiving talking faces. *Current Directions in Psychological Science*, 4(4), 104-109. <https://doi.org/10.1111/1467-8721.ep10772401>
- Ménard, L. (2015). Multimodal speech production. *The Handbook of Speech Production* (p. 200-221). <https://doi.org/10.1002/9781118584156.ch10>
- Möttönen, R. et Sams, M. (2008). Audiovisual interaction. *Handbook of signal processing in acoustics*. D. Havelock, S. Kuwano et M. Vorländer (dir.), (p. 731-745). Springer New York. https://doi.org/10.1007/978-0-387-30441-0_37
- Munhall, K. G., Kroos, C., Jozan, G. et Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception & Psychophysics*, 66(4), 574-583. <https://doi.org/10.3758/BF03194902>
- Open Science Tools Ltd. pavlovia.org. <https://pavlovia.org/> 2022.
- Open Science Tools Ltd. Psychopy.org. Copyright 2002-2018, Open Science ToolsLtd. ; Copyright 2002-2018, Jonathan Peirce. <https://www.psychopy.org/> 2022.
- Osterhout, L. et Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6), 785-806. [https://doi.org/10.1016/0749-596X\(92\)90039-Z](https://doi.org/10.1016/0749-596X(92)90039-Z)
- Robert-Ribes, J., Schwartz, J.-L., Lallouache, T. et Escudier, P. (1998). Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise.

- The Journal of the Acoustical Society of America*, 103(6), 3677-3689.
<https://doi.org/10.1121/1.423069>
- Schwartz, J.-L., Basirat, A., Ménard, L. et Sato, M. (2012). The perception-for-action-control theory (PACT): a perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, 25(5), 336-354. <https://doi.org/10.1016/j.jneuroling.2009.12.004>
- Shipp, T. (1973). Intraoral air pressure and lip occlusion in midvocalic stop consonant production. *Journal of Phonetics*, 1(2), 167-170.
- Sumby, W. H. et Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2), 212-215. <https://doi.org/10.1121/1.1907309>
- The MathWorks Inc. (2021). *MATLAB*. The MathWorks Inc. <https://www.mathworks.com>
- Trudeau-Fisette, P. (2015). *L'apport de la vision dans la perception et la production des voyelles: une étude comparative entre locuteurs adultes adultes voyants et non-voyants* [Mémoire, Université du Québec à Montréal]. <https://archipel.uqam.ca/8685/1/M14131.pdf>
- Tye-Murray, N., Spehar, B., Myerson, J., Hale, S. et Sommers, M. (2016). Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. *Psychology and Aging*, 31(4), 380.
- Young, A. T. Blurring and unblurring. <https://aty.sdsu.edu/bibliog/latex/scan/blur.html> 2020.

ANNEXE A Questionnaire des antécédents :

Université du Québec à Montréal

Automne 2021

Titre : **Étude en linguistique sur la perception audio et visuelle**

Questionnaire antécédents

Nom :

Courriel :

Numéro de téléphone (facultatif) :

Âge :

Sexe :

Avez-vous un trouble de l'audition ou du langage connu? Oui – Non (effacez la mention inutile)

Avez-vous un trouble de la vision? Oui – Non

Si oui, lequel : (si vous ne vous rappelez plus, passez à la question suivante)

Ce trouble de la vision est-il corrigé par des lunettes ou lentilles de contact ? Oui – Non

Quelle.s est/sont votre/vos langue.s maternelle.s? (Ex. français québécois)

Matériel :

Type d'écouteurs utilisés : casque d'écoute ou petits écouteurs auriculaires (effacez la mention inutile)
avec ou sans fil. (effacez la mention inutile)

Taille approximative de l'écran (en pouce) :

Navigateur utilisé :

Merci beaucoup!

ANNEXE B Formulaire de consentement



FORMULAIRE DE CONSENTEMENT

Titre du projet de recherche

Impact d'une perturbation visuelle en condition audio bruitée sur l'intégration audiovisuelle (une étude linguistique sur l'impact de perturbations audios et visuelles sur la perception de syllabes)

Étudiant-chercheur

François Steffann Ba., étudiant à la maîtrise en linguistique, tél. : 514-880-6405, courriel : steffann.francois@courrier.uqam.ca

Direction de recherche

Lucie Ménard Ph. D., Département de linguistique, tél. : 514-987-3000 poste 3515 courriel : menard.lucie@uqam.ca

Dave Saint-Amour Ph. D., Département de Psychologie, tél. : 514-987-3000 poste 7698 courriel : saint-amour.dave@uqam.ca

Préambule

Nous vous demandons de participer à un projet de recherche qui implique la réalisation d'une tâche de reconnaissance de syllabe du français. Il vous sera demandé de visionner une série d'extraits audios et vidéos d'une personne articulant des syllabes du français. Le son ou la vidéo pourront être perturbés (bruit perturbant le son et/ou floutage perturbant la vidéo).

Avant d'accepter de participer à ce projet de recherche, veuillez prendre le temps de comprendre et de considérer attentivement les renseignements qui suivent.

Ce formulaire de consentement vous explique le but de cette étude, les procédures, les avantages, les risques et inconvénients, de même que les personnes avec qui communiquer au besoin.

Le présent formulaire de consentement peut contenir des mots que vous ne comprenez pas. Nous vous invitons à poser toutes les questions que vous jugerez utiles.

Description du projet et de ses objectifs :

Ce projet est une étude en linguistique phonétique qui étudie la perception de la parole chez l'humain et l'impact d'une perturbation audio (bruit) et visuelle (floutage) sur celle-ci. Le projet se déroule sur un peu plus d'un an dans le cadre de mon mémoire de maîtrise et implique la passation du test mentionné par une vingtaine de participant.e.s. Nous travaillerons avec des adultes entre 18 et 35 ans n'ayant pas de problème d'audition et n'ayant pas de problème de vision non corrigés par des lunettes. Notre but étant d'observer le comportement typique de perception audio et visuelle dans une population ne présentant pas de trouble connu. Nous voulons ainsi voir comment une perturbation visuelle peut influencer la perception de la parole, quand les conditions d'écoutes sont aussi perturbées. Nous pensons que la perception sera plus difficile quand les conditions sont perturbées.

Nature et durée de votre participation

La tâche sera réalisée sur votre ordinateur personnel et nécessitera l'utilisation d'écouteurs auriculaires ou d'un casque d'écoute. Le test durera **entre 20 et 30 minutes**, et pendant celui-ci il vous sera demandé de visionner une série d'extraits audios et vidéos d'une personne articulant des syllabes du français. Comme nous l'avons dit, le son ou la vidéo pourront être perturbés (bruit perturbant le son et/floutage perturbant la vidéo). À la suite de chaque visionnement, nous vous demanderons d'identifier la syllabe entendue en choisissant à l'aide de votre clavier une des réponses offertes sur votre écran.

Lieu : La passation se fera en ligne sur votre ordinateur

Avant la passation, je vous transmettrai par courriel les instructions détaillées et je répondrai à toutes vos questions sur le formulaire de consentement et le questionnaire sur les antécédents. Dans le courriel d'instructions, je vous expliquerai alors la tâche que vous devrez accomplir, puis vous devrez accomplir le test seul.e sur le logiciel de passation via le lien fourni. Après le test, vous pourrez me joindre par courriel en cas de problème ou pour toute question.

Avantages liés à la participation

Vous pourrez participer au tirage au sort qui permettra à deux des participant.e.s de gagner 50\$. Vous aurez contribué à l'avancement de la science et aurez aidé la recherche à mieux comprendre l'utilisation de la vision dans la compréhension de la parole.

Risques liés à la participation

En principe, aucun risque ne sont liés à la participation à cette recherche.

Confidentialité

Lors de la passation, il vous sera attribué un code aléatoire permettant d'identifier les données du test. Toutes vos données nominales (Nom, courriel, numéro de téléphone... permettant de vous identifier personnellement) seront séparées des données du test. Vos données nominales seront associées à votre code aléatoire dans un fichier à part encrypté et protégé par mot de passe auquel seul moi (l'étudiant chercheur) aurai accès. Ce fichier sera détruit ainsi que toutes vos informations nominales une fois le projet terminé.

Toutes les données du test seront anonymisées grâce au code aléatoire et stockées dans un fichier sur un ordinateur protégé par mot de passe. Seuls moi et les chercheur.e.s travaillant sur le projet auront accès à ces données, et ces données ne permettront pas de vous identifier. Les données présentées dans les travaux de recherche le seront de manière agrégée (moyenne ou courbe par exemple) et ne permettront pas l'identification. Si des données individuelles devaient être présentées dans un travail de recherche ou dans une publication scientifique, elles seront identifiées par le code aléatoire et ne permettront pas de vous identifier.

Une fois le projet terminé, nous garderons les données anonymisées dans nos archives et détruirons (tel que mentionné) les données nominales permettant l'identification. Nous ne prévoyons pas de détruire les données collectées lors de l'expérience et qui ne permettent pas de vous identifier (données non-nominales et anonymisées).

Utilisation secondaire des données

Une fois le projet de recherche terminé, il se peut que les données, une fois anonymisées soient utilisées pour un projet de recherche ultérieure par moi ou un membre de la communauté scientifique. Elles ne permettront cependant pas de vous identifier. Acceptez-vous que les données de recherche soient utilisées pour réaliser d'autres projets de recherche dans le même domaine ? Ces projets de recherche seront évalués et approuvés par un Comité d'éthique de la recherche de l'UQAM avant leur réalisation. Les données de recherche seront conservées de façon sécuritaire. Afin de préserver votre identité et la confidentialité des données de recherche, vous ne serez identifié que par le code aléatoire anonyme.

Acceptez-vous que les données de recherche soient utilisées dans le futur par d'autres chercheurs à ces conditions ?

Effacer la mention inutile :

☐ Oui ☐ Non

Participation volontaire et retrait

Votre participation est entièrement libre et volontaire. Vous pouvez refuser d'y participer ou vous retirer en tout temps sans devoir justifier votre décision. Si vous décidez de vous retirer de l'étude, vous n'avez qu'à aviser François Steffann (le chercheur) verbalement ou par courriel ; toutes les données vous concernant seront détruites.

Indemnité compensatoire

Aucune indemnité compensatoire n'est prévue.

Des questions sur le projet ?

Pour toute question additionnelle sur le projet et sur votre participation, vous pouvez communiquer avec les responsables du projet :

Directrice de projet : Lucie Ménard Ph. D., Département de linguistique, courriel : menard.lucie@uqam.ca

Étudiant chercheur : François Steffann Ba., étudiant à la maîtrise en linguistique, , courriel : steffann.francois@courrier.uqam.ca

Des questions sur vos droits ? Le Comité d'éthique de la recherche pour les projets étudiants impliquant des êtres humains (CERPE) a approuvé le projet de recherche auquel vous allez participer. Pour des informations concernant les responsabilités de l'équipe de recherche au plan de l'éthique de la recherche avec des êtres humains ou pour formuler une plainte, vous pouvez contacter la coordination du CERPE :

Julie Sergent, Agente de recherche et de planification, sergent.julie@uqam.ca, tél : (514) 987-3000 poste 3642

Remerciements

Votre collaboration est essentielle à la réalisation de notre projet et l'équipe de recherche tient à vous en remercier.

Consentement

Je déclare avoir lu et compris le présent projet, la nature et l'ampleur de ma participation, ainsi que les risques et les

inconvénients auxquels je m'expose tels que présentés dans le présent formulaire. J'ai eu l'occasion de poser toutes les questions concernant les différents aspects de l'étude et de recevoir des réponses à ma satisfaction.

Je, soussigné(e), accepte volontairement de participer à cette étude. Je peux me retirer en tout temps sans préjudice d'aucune sorte. Je certifie qu'on m'a laissé le temps voulu pour prendre ma décision.

Une copie signée de ce formulaire d'information et de consentement doit m'être remise.

Prénom Nom

Signature

Date

Engagement du chercheur

Je, soussigné(e) certifie

- (a) avoir expliqué au signataire les termes du présent formulaire; (b) avoir répondu aux questions qu'il m'a posées à cet égard;
- (c) lui avoir clairement indiqué qu'il reste, à tout moment, libre de mettre un terme à sa participation au projet de recherche décrit ci-dessus;
- (d) que je lui remettrai une copie signée et datée du présent formulaire.

____François Steffann
Prénom Nom



Signature

____31-10-2021
Date