

# Apport des informations audiovisuelles dans la perception et la production de la consonne /v/ par les apprenants thaïlandais du français

Thi Thuy Hien Tran<sup>1,\*</sup>, Supansa Tusnyingyong<sup>2</sup>, Coriandre Vilain<sup>1</sup>, Printemps Sordes<sup>1</sup> et Silvain Gerber<sup>1</sup>

<sup>1</sup>Grenoble Images Parole Signal Automatique (GIPSA-lab)

Univ. Grenoble Alpes, CNRS, Grenoble INP

<sup>2</sup>Thammasat University, Faculty of Liberal Arts

\*[thi-thuy-hien.tran@gipsa-lab.grenoble-inp.fr](mailto:thi-thuy-hien.tran@gipsa-lab.grenoble-inp.fr)

**Résumé.** Cette étude, s'inscrivant dans le contexte de l'enseignement et de l'apprentissage du français, explore l'une des difficultés récurrentes rencontrées par les apprenants thaïlandais lors du traitement de la consonne non native /v/ : la confusion fréquente avec /w/ en position initiale de syllabe.

Nous avons réalisé une étude pilote expérimentale impliquant huit apprenants thaïlandais de niveaux A1-A2 afin d'évaluer l'impact des informations audiovisuelles sur le traitement des consonnes /f v w/. Deux types de tâches (perception et production) portant sur ces trois consonnes dans deux structures syllabiques (CV, VCV) et quatre contextes vocaliques non arrondis (/i e ε a/) ont été menés dans deux modalités distinctes (auditive (A) et audiovisuelle (AV)). En perception, une tâche de discrimination demandait aux participants de juger la similitude entre les stimuli de test et les stimuli de référence. En production, ils répétaient les stimuli perçus précédemment, soit entendus seuls (modalité A), soit entendus et vus (modalité AV). Quatre participants ont été testés en modalité A et quatre en modalité AV.

Les résultats mettent en évidence une nette amélioration du taux de réussite dans la perception et la production de /v/ en modalité AV par rapport à la seule modalité A, quel que soit le type de tâche. De plus, la confusion entre /v/ et /w/ en perception et production est moins fréquente lorsque les apprenants bénéficient à la fois d'informations visuelles et sonores. Par ailleurs, le temps de réponse lors de la perception des consonnes est significativement réduit en modalité AV comparativement à la modalité A. Dans l'ensemble, ces résultats démontrent que les informations audiovisuelles facilitent et accélèrent le traitement de la consonne non native /v/ en position initiale de syllabe chez les apprenants thaïlandais.

## 1 Contexte

Les études sur l'acquisition des fricatives par les apprenants thaïlandais dans une langue étrangère ont unanimement mis en évidence les défis rencontrés lors de leur perception et/ou de leur production (Chunsuvimol & Ronnakiat, 2001; Ngammana, 2011; Le Corre, 2013; Promkesa, 2014; Sridhanyarat, 2017; Tusnyingyong & Tran, 2022 entre autres). Ces études ont mis en lumière la tendance des apprenants à assimiler ces consonnes à celles de leur langue maternelle (L1), engendrant des erreurs de prononciation liées aux interférences avec leur système phonologique d'origine. Par exemple, le dévoisement des consonnes sonores du français est une stratégie prédominante chez les apprenants thaïlandais, illustrée par des réalisations telles que [ʃwaziz] pour "choisir" ou [mɛzɔ̃] pour "maison". Ces observations rejoignent les études antérieures démontrant la fréquente prononciation des fricatives sonores /z/ et /ʒ/ en début de syllabe sous forme de [s] et [ʃ] (Le Corre, 2013) ou la réalisation de la plosive sonore /g/ par son homologue sourd [k] (Ngammana, 2011). Les différences phonologiques entre les deux langues expliquent en partie ces erreurs : les consonnes obstruantes en français sont distinguées principalement par le voisement, tandis qu'en thaï, la distinction se fait davantage par l'aspiration.

La stratégie de dévoisement diffère cependant pour la consonne /v/, absente en thaï. Contrairement aux autres fricatives sonores non natives /z/ et /ʒ/, les apprenants n'appliquent pas le même schéma pour /v/. Bien que la fricative sourde homologue /f/ soit présente en thaï, la confusion survient souvent avec /w/ en position initiale de syllabe, entraînant des déformations fréquentes par rapport à la cible (par exemple, "visa" prononcé [wi.sa], "vélo" prononcé [we.lo], "Covid" prononcé [ko.wit']) (Nacaskul, 1979; Tusnyingyong & Tran, 2022). Il est à noter que cette confusion entre /v/ et /w/ se manifeste exclusivement en position initiale de syllabe, jamais en finale. L'erreur de substitution de /v/ par /w/ semble provenir d'une association erronée entre le graphème latin < v > /v/ et le graphème < ʋ > /w/ du thaï, influencée par les processus de romanisation du thaï et de translittération des mots d'emprunt en caractères thaïs (Tusnyingyong & Tran, 2022). Par ailleurs, cette confusion est particulièrement observée lors de tâches impliquant une interaction directe ou indirecte avec l'orthographe, telles que la lecture ou l'entretien basé sur un texte préparé. En revanche, elle est moins fréquente lors d'exercices d'imitation où les apprenants sont simplement invités à écouter et répéter sans accès à un support écrit (Chunsuvimol & Ronnakiat, 2001; Le Corre, 2013; Ngammana, 2011; Sridhanyarat, 2017; Tusnyingyong & Tran, 2022).

De nombreuses recherches démontrent que la présentation audiovisuelle améliore le traitement des sons non natifs (Hazan et al., 2006; Wang et al., 2008, 2009; Burfin, 2015 entre autres). La performance dans la compréhension de la parole dans une langue étrangère s'améliore lorsque les apprenants sont exposés à la fois aux informations auditives et visuelles par rapport à une exposition uniquement auditive. Cependant, l'efficacité de l'exploitation des ressources audiovisuelles peut différer en fonction du système phonologique d'origine des apprenants, et la manière dont chaque phonème est traité différencie les apprenants d'une langue étrangère. Quelle utilité ces informations apportent-elles réellement et quels avantages spécifiques offrent-elles dans le traitement de la parole ?

## 2 La perception audiovisuelle dans le traitement de la parole

Dans nos interactions quotidiennes, le contact visuel avec nos interlocuteurs nous permet d'exploiter diverses sources d'informations pour améliorer la compréhension de la parole, intégrant des gestes, expressions faciales, postures et autres mouvements associés au discours. Cette capacité à unifier ces indices sensoriels est essentielle pour former une représentation unifiée du monde qui nous entoure (Stein & Meredith, 1993). La faculté à associer deux sources d'informations (visuel et auditif) se manifeste dès la naissance ou se développe assez tôt au cours de l'enfance. Les nourrissons démontrent une précocité à distinguer et imiter les expressions faciales (Meltzoff & Moore, 1977; Field et al., 1982), suggérant une correspondance entre les gestes observés et leurs propres mouvements musculaires. À trois ou quatre mois, ils peuvent associer un son entendu à une réalisation articulatoire correspondante en visionnant une vidéo : ils reconnaissent, par exemple, qu'une bouche ouverte correspond au son [a] et une bouche fermée au [i] (Kuhl & Meltzoff, 1982; Legerstee, 1990). Une aptitude similaire a été observée pour la perception des consonnes. À l'âge de six mois, les nourrissons hispanophones et anglophones identifient correctement la réalisation articulatoire correspondant à des sons auditifs tels que /ba/ ou /va/ (Pons et al., 2009).

Cette utilisation des indices visuels dans le traitement de la parole dès notre enfance suscite une question essentielle : quelle est l'équivalence visuelle du phonème dans le langage parlé ? Lorsque nous observons quelqu'un parler, que perçoit-on ? Les visèmes (Fisher, 1968, cité par Files et al., 2015) !

Un visème représente la plus petite unité distinctive de l'articulation visuelle et est considéré comme un "phonème visuel" (*visual phoneme*), selon la description de Dumont & Calbour (2002), englobant les phonèmes articulés à travers un même geste facial. En parlant, nous ajustons les caractéristiques de notre conduit vocal en coordonnant séquentiellement ou simultanément différents organes articulatoires et résonateurs, façonnant ainsi le signal acoustique émis. Cela implique divers mouvements articulatoires, dont certains sont plus perceptibles que d'autres. Par exemple, les consonnes /p b m/ partagent l'utilisation

des lèvres pour l'articulation. Leur articulation visuelle commune forme un visème. Ces trois consonnes /p b m/, ayant des similitudes labiales, sont considérées comme des "sosies labiaux" - un ensemble de phonèmes indiscernables uniquement par la lecture labiale (Istria et al., 1982; Borel et al., 2016, p.28). Néanmoins, certains phonèmes présentent une articulation "invisible", comme les consonnes /k g/ produites dans la partie postérieure de la cavité buccale.

Les catégories visémiques identifiables des consonnes du français, la langue cible de notre étude, peuvent varier en fonction de la position de la consonne dans la séquence (initiale, médiane, finale). Néanmoins, les fricatives labiodentales /f v/ se démarquent, avec les occlusives bilabiales /p b m/ et les fricatives post-alvéolaires protruses /ʃ ʒ/, comme des classes qui sont toujours visibles indépendamment de leur place dans la séquence (Istria et al., 1982). Les labiodentales font donc partie des phonèmes consonantiques les plus facilement distingués car articulés à l'avant du conduit vocal.

Diverses études mettent en lumière les bénéfices de l'utilisation de l'information audiovisuelle pour améliorer le traitement des sons non natifs, soulignant la nature multimodale de la parole où la combinaison des entrées auditives et visuelles accélère le processus de traitement des consonnes, qu'elles soient natives ou non. Hazan et ses collègues (2006) se sont intéressés à l'impact des indices visuels sur la distinction des lieux d'articulation (labial /p b/ vs labiodental /v/) chez des apprenants hispanophones et japonophones de l'anglais. Les participants ont été exposés à des pseudo-mots de structures CV, VCV ou VC contenant les consonnes /p b v/ et les voyelles /i a u/. Alors que les consonnes /p b/ existent dans les langues des deux groupes, /v/ n'est pas un phonème ni en japonais ni en espagnol. Les résultats mettent en évidence une nette amélioration de la perception en modalité audiovisuelle par rapport aux modalités auditive et visuelle seules. Cependant, les hispanophones surpassent les japonophones dans la perception du contraste /b/ ~ /v/, en raison de la présence du visème de la fricative labiodentale /f/ dans leur langue, alors qu'il est inexistant en japonais.

L'étude menée par Wang et ses collègues (2009) met en évidence que les apprenants de langues étrangères sont plus performants dans l'identification du lieu d'articulation des visèmes non natifs par rapport aux visèmes natifs lorsqu'ils sont présentés en modalité audiovisuelle. Les participants sinophones (mandarins) et coréens, ainsi qu'un groupe anglophone (groupe de contrôle), ont été testés en trois modalités : auditive, visuelle et audiovisuelle. Les résultats soulignent une meilleure performance pour le lieu interdental (visème non natif) en modalité audiovisuelle, mais une amélioration moins marquée pour le lieu labiodental (visème natif). De plus, aucune amélioration n'est constatée pour le lieu alvéolaire, suggérant que les apprenants sinophones et coréens tirent davantage parti des informations visuelles pour les visèmes non natifs que pour les visèmes natifs.

D'autre part, Burfin (2015) confirme l'impact des informations audiovisuelles dans le traitement des sons non natifs, notamment pour les visèmes absents de la langue maternelle. Cette étude se concentre sur la perception du contraste entre /f/ et /θ/ chez les francophones. Alors que /f/ existe en français, /θ/ et son visème (interdental) ne font pas partie de cette langue. Les participants ont été testés dans les modalités auditive et audiovisuelle. Les résultats indiquent une augmentation significative du pourcentage de réponses correctes pour la consonne /θ/ en modalité audiovisuelle par rapport à la modalité auditive chez les francophones. De plus, en perception, le temps de réponse en modalité audiovisuelle est significativement plus court qu'en modalité auditive, que les stimuli soient natifs ou non natifs.

Il est à noter que l'importance des informations visuelles dans la perception des langues étrangères semble varier selon le niveau des apprenants. Des études révèlent que les apprenants plus expérimentés tirent un meilleur parti des indices visuels. Par exemple, une étude de Wang et ses collègues (2008) compare trois groupes : des sinophones ayant vécu pendant 10 ans dans une zone anglophone (longue durée de séjour), d'autres y ayant séjourné pendant 2 ans (courte durée de séjour) et des anglophones canadiens. En utilisant des stimuli audiovisuels, les apprenants de longue durée de séjour obtiennent des résultats significativement

supérieurs, suggérant que l'utilisation des informations visuelles pourrait varier en fonction du niveau de maîtrise de la langue étrangère. De plus, selon Hazan et ses collègues (2006), un facteur culturel lié à la perception du regard (contact visuel) direct, présent dans certaines cultures, pourrait également influencer la capacité à exploiter ces indices visuels.

Malgré ces variations, les recherches soulignent que la présentation audiovisuelle facilite l'apprentissage des langues étrangères. Dans cette perspective, notre étude se focalise sur les apprenants thaïlandais du français, qui éprouvent des difficultés à distinguer la consonne /v/ de /w/ en début de syllabe, du fait de l'absence de /v/ dans la langue thaïe.

### **3 Etude expérimentale sur l'apport des informations audiovisuelles dans la perception et la production de /v/ chez les thaïlandais**

#### **3.1 Objectif et hypothèse**

L'objectif principal de cette étude pilote est d'évaluer l'impact des indices audiovisuels sur la perception et la production de la consonne /v/ en début de syllabe chez les apprenants thaïlandais.

Nous émettons l'hypothèse que l'introduction d'éléments visuels en complément des informations auditives pourrait améliorer à la fois la perception et la production de /v/, réduisant ainsi la confusion avec /w/. Cette amélioration devrait donc être plus marquée en modalité audiovisuelle par rapport à la seule modalité audio. En effet, étant identifiable visuellement grâce à la lecture labiale, la consonne /v/ devrait favoriser une meilleure perception et production lorsqu'elle est présentée en audiovisuel. De même, la consonne thaïe /w/, souvent confondue avec /v/, pourrait également tirer profit d'une présentation en audiovisuel, étant également identifiable visuellement par son articulation labiale (Istria et al., 1982).

Pour analyser cet impact, notre étude se base sur deux modalités de test, l'audio seule (A) et l'audiovisuelle (AV). Nous examinons spécifiquement les trois consonnes /f v w/. Ce choix est basé sur le fait que /f/ et /v/ partagent le même lieu d'articulation labiodentale, avec comme seule différence le voisement, et sur la confusion courante entre /v/ et /w/ chez les apprenants thaïlandais.

#### **3.2 Méthodologie**

##### **3.2.1 Participants**

Avant de sélectionner les participants pour l'expérience, un pré-test impliquant 14 étudiants thaïlandais a été mené où ils ont lu des mots et des phrases en français affichés sur les diaporamas. Les mots de test contenant /v/ en début (ex. *vase, voiture, venir, visa*) ou au milieu (ex. *télévision, travail, développer, rivière*) ainsi que les mots de contrôle servant de distracteurs (ex. *plat, cou, moi, poulet*), ont été présentés dans un ordre aléatoire. Les phrases incluent un ou plusieurs mots avec /v/ en initiale de syllabe (ex. *voici la photo des vacances, je vais à la mer, voudriez-vous les goûter ? posez les questions suivantes, écoutez les conversations avec les serveurs*).

Après l'analyse acoustique des enregistrements, huit étudiants (cinq hommes et trois femmes), âgés de 23 à 36 ans, montrant les plus grandes difficultés dans la prononciation de /v/, ont été choisis pour prendre part à l'expérience. Tous suivent des cours de français et poursuivent des études supérieures à xxx. Ils sont en France depuis moins d'un an au moment du test. Leur niveau du français est débutant (A1-A2). Le thaï « standard » du Centre est la L1 pour tous. Trois parlent d'autres langues comme L1 (laotien, malais, des langues régionales du Nord (lanna) et du Sud (dambro) de la Thaïlande), mais /v/ n'est pas présente dans leurs L1s respectives. L'anglais est leur première langue étrangère. Trois apprennent également le coréen,

le mandarin et l'arabe au niveau débutant. Deux portent des lunettes ou des lentilles de contact en raison de la myopie. Un participant est daltonien et réalise l'expérience en modalité audio seule. Aucun problème auditif n'a été identifié chez les participants. Ils sont répartis en deux groupes : quatre en modalité A et quatre en modalité AV.

### 3.2.2 Constitution du corpus

Tusnyingyong & Tran (2022) ont mis en évidence les similarités phonosymboliques entre <v> et [w], présentes dans des mots réels et intégrées dans le lexique mental des apprenants. Notre objectif est de garantir que les apprenants se concentrent uniquement sur l'aspect sonore des stimuli, sans être influencés par leur signification ou leur familiarité lexicale, comme cela pourrait être le cas avec des mots courants. Les habitudes articulatoires erronées pourraient conduire à une association par défaut, par exemple, du mot "visa" avec la forme sonore [wiza]. Ainsi, pour cette étude, nous avons sélectionné des pseudo-mots de deux structures syllabiques, CV (ex. [wi]) et VCV (ex. [iwi]), en raison de la confusion attestée entre /v/ et /w/ dans deux positions différentes (initiale absolue et médiane) au sein du mot (Le Corre, 2013 ; Tusnyingyong & Tran, 2022). Nous avons choisi des voyelles communes au français et au thaï (/i e ε a/). Les voyelles arrondies (/u o ɔ/) existant dans les deux langues, dont l'articulation labiale peut influencer la visibilité des consonnes, n'ont pas été incluses dans le test.

Au total, vingt-quatre stimuli de test ont été sélectionnés : 12 monosyllabes ([fi fe fe fa vi ve ve va wi we we wa]), 12 dissyllabes ([ifi efe efe afa ivi eve eve ava iwi ewe ewe awa]). De plus, vingt-quatre autres stimuli (de contrôle) de structure CV et VCV (contenant les consonnes communes aux deux langues /p b t/) ont été ajoutés afin de masquer l'objectif réel de l'étude. Pour la phase d'entraînement avant le test, douze stimuli comportant d'autres consonnes /m s k/ ont également été préparés.

Chaque item a été lu 5 fois par un locuteur français, âgé de 49 ans, originaire du Sud de la France. Les enregistrements audio et audiovisuels ont eu lieu dans la chambre anéchoïque du laboratoire xxx. Les vidéos ont été enregistrées avec la caméra Sony HDR-XR500E. L'enregistrement audio a été effectué simultanément avec l'enregistreur Marantz PMD 670 et le micro AKG C1000S. Les stimuli ont été numérisés au format .wav à 44,1 kHz et 16 bits.

Pour chaque item, un ou deux meilleurs exemplaires, débutant et se terminant avec la bouche fermée et le regard centré, ont été sélectionnés. Les deux entrées audio (enregistré avec le PMD 670) et vidéo (enregistré avec la caméra Sony) ont été synchronisées, ce qui a permis de remplacer le son de la vidéo par l'enregistrement audio pour obtenir une qualité optimale des stimuli audiovisuels utilisés pour le test. Au final, nous disposons de 78 stimuli audiovisuels fabriqués pour la modalité AV. Le même nombre de stimuli a été créé pour la modalité A où chaque stimulus sonore a été associé à une image statique (visage neutre) du locuteur.

Les interfaces de test (perception et production) ont été créées à l'aide de Presentation® (version 23.0) (NeuroBehavioral Systems, Inc., Berkeley, California). Ce logiciel de conception d'expérimentations, très couramment utilisé en neurosciences, fonctionne sur les PC Windows et administre des stimuli auditifs, visuels et multimodaux avec une précision temporelle de quelques millisecondes.

Dans la tâche de perception, les sujets sont exposés aux 4 séquences d'entraînement, 72 séquences de test (24 séquences × 3 répétitions) et 24 séquences de contrôle. L'ensemble des séquences de test et de contrôle a été mis en ordre aléatoire en respectant le schéma 3-1 : chaque série de 3 séquences de test est suivie d'une séquence de contrôle. Chaque séquence contient un item de référence (par exemple [vi]) présenté 6 fois et trois items de test dont un est identique à l'item de référence ([vi]) et deux sont différents ([fi] et [wi]) (cf. table 1). Il est à noter que, à l'intérieur de chaque séquence de test, les voyelles de chaque item sont inchangées. Pour diversifier la position des items de test, nous avons établi trois séquences pour chaque

item de référence, variant l'ordre de présentation des items de test. Ainsi, dans un bloc de trois séquences pour un item de référence (par ex. [vi]), chaque item de test (par ex. [vi], [wi], [fi]) est présenté une fois immédiatement après l'item de référence [vi]. Par conséquent, trois séquences distinctes ont été élaborées pour chacun des 24 stimuli de test.

**Table 1.** Exemples de séquences de test pour la tâche de perception.

| Séquences de test  |               |     |     |
|--------------------|---------------|-----|-----|
| Items de référence | Items de test |     |     |
| vi                 | vi            | fi  | wi  |
| vi                 | wi            | vi  | fi  |
| vi                 | fi            | wi  | vi  |
| awa                | awa           | ava | afa |
| awa                | afa           | awa | ava |
| awa                | ava           | afa | awa |

Pour la tâche de production, chacun des 24 stimuli de test (voir table 1) est présenté 5 fois en ordre aléatoire. Un stimulus de contrôle est inséré après tous les quatre stimuli de test. En somme, nous avons 4 stimuli d'entraînement ([ma], [eke], [si] et [εmε]), 120 stimuli de test et 30 stimuli de contrôle.

### 3.2.3 Déroulement de l'expérience

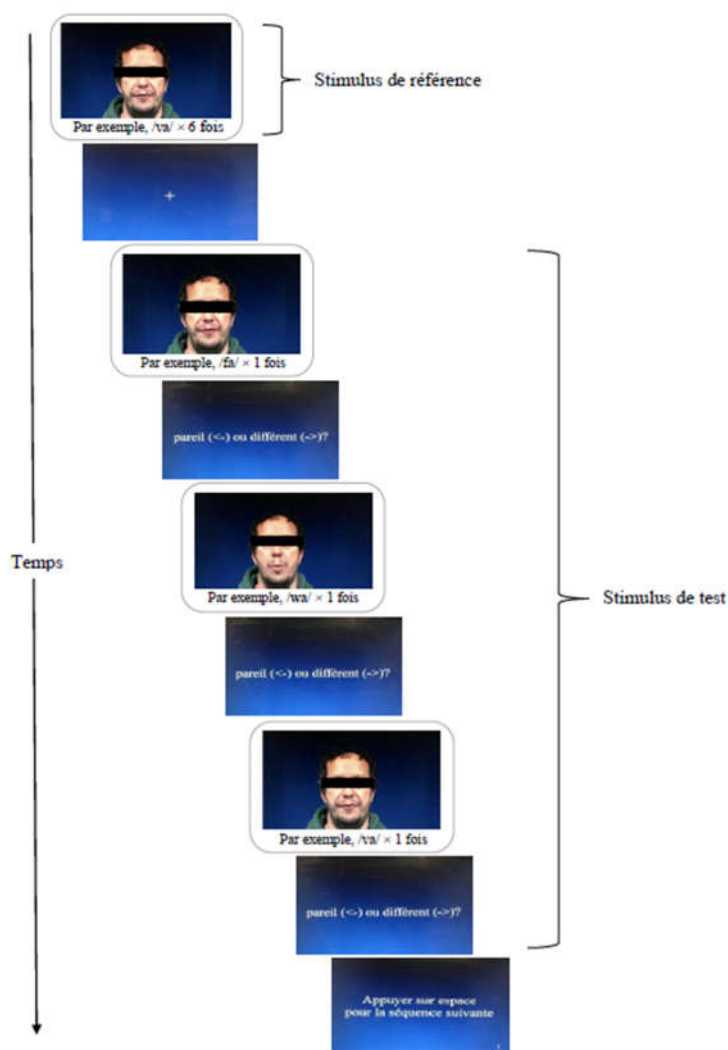
Chaque modalité de test (A et AV) implique quatre apprenants participant à deux sessions (perception et production). Chaque participant est évalué dans une seule modalité (audio seule ou audiovisuelle), avec un ordre de test varié pour équilibrer les séquences (perception-production ou production-perception). L'expérience se déroule dans la chambre anéchoïque du laboratoire xxx avec les stimuli présentés via le logiciel Presentation®. Les sujets sont assis à une distance de 0,5 mètre de l'écran, le son étant diffusé par un haut-parleur derrière celui-ci. Le niveau sonore est individuellement ajusté pour chaque participant. Avant les tests, tous les participants remplissent un questionnaire sociolinguistique, les détails étant exposés dans la section 3.2.1.

#### **Test de perception**

Avant de commencer, une phase préliminaire permet aux participants de se familiariser avec l'interface à l'aide de stimuli différents de ceux utilisés dans le test. Ensuite, les sujets observent un stimulus de référence répété six fois. Par la suite, trois stimuli de test sont présentés. Les sujets doivent déterminer si chaque stimulus est identique ou différent du référentiel en utilisant les touches ← (pour "pareil") et → (pour "différent") du clavier. Ils passent à la séquence suivante en appuyant sur la touche "espace". Le temps de réponse est mesuré depuis la fin de la présentation du stimulus jusqu'au moment où le sujet appuie sur la touche de réponse.

En modalité AV, les sujets visionnent une vidéo du locuteur natif prononçant les stimuli, tandis qu'en modalité A, seuls les sons sont diffusés avec une image statique du visage du locuteur à l'écran. La phase de test comprend 4 séquences d'entraînement et 96 séquences de test, entrecoupées de cinq pauses toutes

les 16 séquences, totalisant une durée d'environ 30 minutes pour la tâche de perception. Une représentation schématique de cette tâche est illustrée à la figure 1.



**Figure 1.** Représentation schématique d'une séquence de test dans la tâche de perception en modalité AV

### **Test de production**

Pour la tâche de production, les sujets ont pour consigne de répéter les sons perçus. En modalité AV, ils visionnent des vidéos, tandis qu'en modalité A, seuls les sons sont présentés accompagnés d'une image fixe du visage neutre du locuteur (bouche fermée). Chaque stimulus est présenté une seule fois, suivi de l'apparition du mot « Répétez » à l'écran pour inciter les sujets à reproduire les sons. L'expérimentateur contrôle la progression en passant au stimulus suivant. L'enregistrement est effectué à l'aide d'un enregistreur Marantz PMD 670 et d'un microphone AKG C1000S à directivité cardioïde. La phase de production comprend 4 stimuli d'entraînement et 150 stimuli dans la phase de test, avec des pauses tous les 25 stimuli. La durée totale de la tâche de production est d'environ 10 minutes.

### 3.2.4 Traitement de données

À partir des fichiers de résultats obtenus par le logiciel Presentation®, l'extraction des données du test perceptif a été faite à l'aide du logiciel Matlab. Des analyses quantitatives et statistiques ont été menées par la suite avec Microsoft Excel et R. Les données acoustiques du test de production ont été traitées à l'aide des deux logiciels Phon (Rose et al., 2006) et Praat (Boersma & Weenick, 1992-2022), en nous basant sur les descriptions acoustiques des consonnes de Calliope (1989) et de Vaissière (2006).

Les segments de la fricative labiodentale [v] sont identifiés par la présence du bruit de friction sur le signal acoustique. Des pics diffus de faible intensité vers 1,5 kHz et une barre de voisement dans les très basses fréquences doivent également être repérés sur le spectrogramme. Une prononciation correcte de [v] est donc confirmée par la présence du bruit de friction et de la barre de voisement. En cas de réalisation de [w] à la place de [v], les segments de cette approximante labiovélaire sont identifiés par l'absence du bruit de friction et par la présence d'une structure formantique de faible amplitude, avec une variation constante en fonction du contexte vocalique environnant. En cas d'incertitude, nous avons fait appel à trois locuteurs natifs pour écouter les sons et juger quelle consonne ils entendent.

Nous souhaitons étudier l'influence de 4 variables explicatives et de leurs interactions sur 2 variables de réponse « scores » et « temps de réponse ». Les 4 variables explicatives sont les modalités de test (2 modalités : A et AV), les consonnes de référence (3 modalités : f, v, w), les contextes vocaliques (4 modalités : a, e, ε, i) et les structures syllabiques (2 modalités : CV et VCV).

La variable réponse « scores » dans les tâches de perception et de production suit une distribution binomiale (soit 1, soit 0). Compte tenu qu'un participant est sollicité à plusieurs reprises, nous introduisons la variable « sujets » comme effet aléatoire dans le modèle. Nous avons donc choisi d'utiliser une régression linéaire mixte avec distribution binomiale. La méthode a été réalisée à l'aide de la fonction *glmer* du package *lme4* du logiciel R. Quant à la variable « temps de réponse » dans la tâche de perception, notre protocole ayant permis de recueillir plusieurs valeurs de cette variable pour un même sujet, il ne nous garantit donc pas l'indépendance des observations. Notre choix s'est porté sur le modèle linéaire à effets mixtes (avec la fonction *lme* du package *nlme*) pour permettre de respecter l'hypothèse selon laquelle les résidus suivent une loi normale (condition d'application des modèles mixtes), nous avons choisi de transformer la variable réponse en son logarithme.

Pour analyser la différence entre les deux modalités (A et AV) à l'intérieur de chaque modalité de consonnes, voyelles, structures syllabiques, nous appliquons la méthode de Hothorn et al. (2008) qui permet de réaliser les comparaisons multiples de moyennes avec le modèle mixte. La méthode a été appliquée aux données avec la fonction *emmeans* du package *emmeans* du logiciel R.

## 3.3 Résultats

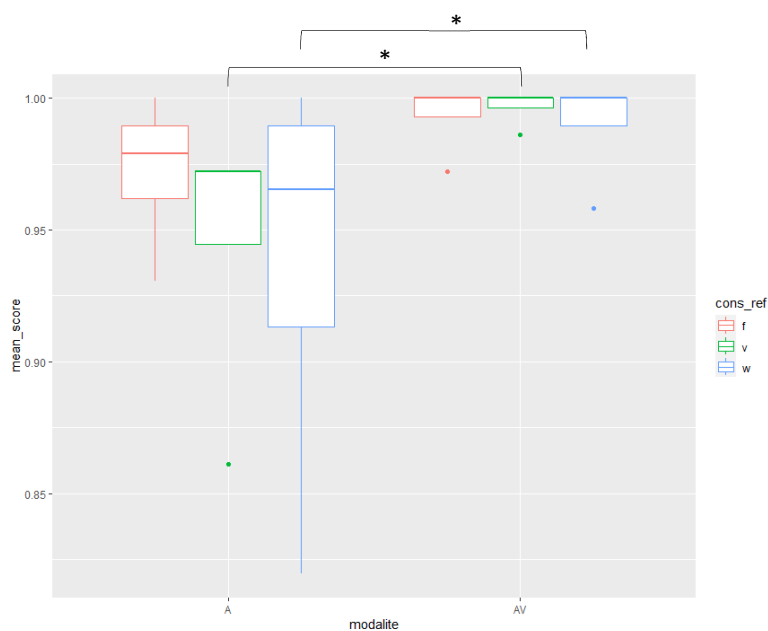
### 3.3.1 Tâche de perception

#### Taux de réussite

Sur l'ensemble des 1728 réponses des huit participants, le score de réponses correctes en modalité AV (99,30 %) est significativement plus élevé qu'en modalité A (95,14 %) ( $z = -2.75$ ,  $p < .01$ ). Le taux de non-réussite est de 2,8 % avec 48 réponses incorrectes parmi lesquelles 42 réponses sont en modalité A et seules 6 réponses en modalité AV.

Les consonnes /f v w/ sont toutes mieux perçues en modalité AV qu'en modalité A lorsqu'elles sont présentées comme référence. En moyenne, les taux de réponses correctes sont très élevés et présentent une faible variabilité en modalité AV (cf. figure 2). En revanche, une plus grande dispersion des réponses est

constatée en modalité audio seule, particulièrement en ce qui concerne le son /w/. Les participants perçoivent /v/ significativement mieux en modalité AV (98,6 %) qu'en modalité A (94,4 %) ( $z = -2.49$ ,  $p < .01$ ). La tendance similaire est également observée pour /w/ (97,9 % vs 93,8 %) ( $z = -2.24$ ,  $p < .05$ ), tandis que pour /f/, la différence des taux de réussite entre les deux modalités n'est pas significative (99,3 % vs 97,2 %) ( $z = -1.53$ ,  $p = 0.12$ ). A l'intérieur de chaque modalité, que ce soit en audio seul ou en audiovisuel, aucune différence significative n'est relevée entre les taux de réussite de chaque consonne ( $p > .05$ ).



**Figure 2.** Pourcentage des réponses correctes dans la tâche de perception en fonction de la modalité (A vs AV) et de la consonne de référence (/f v w/)

En fonction du contexte vocalique, globalement, les taux de réponses correctes sont plus élevés en modalité AV qu'en modalité A pour chacune des 4 voyelles : 98,6 % vs 95,3 % en contexte /a/ ; 99,5 % vs 96,7 % en contexte /e/ ; 99 % vs 93,5 % en contexte /ɛ/ ; 100 % vs 94,9 % en contexte /i/. Cependant, seul dans le contexte /ɛ/, le taux de réponses correctes est significativement meilleur en modalité AV ( $z = -2.18$ ,  $p < .05$ ). Les comparaisons entre les deux modalités pour les autres contextes vocalique n'ont pas révélé de différences significatives ( $p > .05$ ). A l'intérieur de chaque modalité, que ce soit en A ou AV, les comparaisons entre les taux des voyelles n'ont pas fait ressortir de différence significative ( $p > .05$ ).

Quel que soit le type de structure syllabique, les taux de réussite sont significativement plus élevés en modalité AV qu'en modalité A : pour la structure CV, respectivement 99,8 % vs 96,2 % ( $z = -2.47$ ,  $p < .05$ ) ainsi que pour la structure CVC, 98,8 % vs 93,9 % ( $z = -2.29$ ,  $p < .05$ ). Les scores semblent plus élevés avec la structure CV, que ce soit en modalité A ou AV. Cependant, les comparaisons des scores entre les différentes structures (CV vs VCV), à l'intérieur chaque modalité, n'ont pas révélé de différence significative ( $p > .05$ ).

### Temps de réponse

Les participants testés en modalité audio seule ont présenté des temps de réponse significativement plus longs en moyenne (1,11 seconde) que ceux évalués en modalité audiovisuelle (0,47 seconde) ( $t = 6.99$ ,  $p <$

.001). Pour chacune des trois consonnes de référence, le temps moyen de réponse en modalité AV est significativement deux fois plus court qu'en modalité A : /f/ respectivement 0,47 s vs 1,11 s ( $t = 7.37, p < .001$ ) ; /v/ 0,47 s vs 1,16 s ( $t = 7.78, p < .001$ ) ; /w/ 0,47 s vs 1,09 s ( $t = 7.1, p < .001$ ). En outre, la variabilité des temps de réponse en AV est notablement réduite comparée à celle observée en A.

Cette tendance est également observée lors de l'analyse des contextes vocaliques, avec un temps deux fois plus court et une très faible variabilité en modalité AV, indépendamment de la voyelle considérée : /a/ respectivement 0,47 s vs 1,11 s ( $t = 7.53, p < .001$ ) ; /e/ 0,47 s vs 1,06 s ( $t = 7.38, p < .001$ ) ; /ɛ/ 0,46 s vs 1,15 s ( $t = 8.28, p < .001$ ) ; /i/ 0,48 s vs 1,14 s ( $t = 7.65, p < .001$ ).

De la même manière, pour les deux structures syllabiques analysées, le temps de réponse est significativement plus court en modalité AV par rapport à la modalité A, pour CV (0,46 s vs 1,15 s) ( $t = 8.51, p < .0001$ ) ainsi que pour VCV (0,48 s vs 1,08 s) ( $t = 7.34, p < .005$ ). Cependant, à l'intérieur de chaque modalité (A ou AV), aucune différence significative n'est constatée dans les temps de réponse entre les trois consonnes de référence (/f v w/), entre les quatre contextes vocaliques (/a e ε i/), ni entre les deux structures syllabiques analysées (CV et VCV) ( $p > .05$ ).

### Réponses incorrectes

Les erreurs sont bien plus présentes en modalité A (42) qu'en modalité AV (6) (cf. table 2). En modalité A, les erreurs surviennent principalement lors de la distinction entre les consonnes /v/ et /w/, que ces dernières soient positionnées en consonne de référence ou de test. En revanche, en modalité AV, les erreurs ont une répartition relativement homogène. Il est à noter qu'il y a une tendance à davantage confondre /w/ lorsqu'elle est présentée en référence et /v/ comme consonne de test.

**Table 2.** Erreurs de perception dans les deux modalités en fonction des consonnes de référence et de test

| Modalité A  |     |           |           | Modalité AV |     |          |     |
|-------------|-----|-----------|-----------|-------------|-----|----------|-----|
| ref // test | /f/ | /v/       | /w/       | ref // test | /f/ | /v/      | /w/ |
| /f/         | 1   | 5         | 2         | /f/         | 1   | 1        | -   |
| /v/         | 2   | 1         | <b>13</b> | /v/         | -   | 1        | -   |
| /w/         | -   | <b>15</b> | 3         | /w/         | -   | <b>3</b> | -   |

En analysant les erreurs en fonction des structures syllabiques, il apparaît que les participants commettent davantage d'erreurs lorsque la consonne est située dans une position intervocalique VCV (31 erreurs, dont 26 en modalité audio seule et 5 en audiovisuel), par rapport à une position initiale absolue CV (17 erreurs, dont la majorité se produit en modalité audio seule (16) plutôt qu'en audiovisuel (1)).

Les participants font plus d'erreurs avec les voyelles moyennes /ɛ e/ (24) qu'avec /a/ (13) ou /i/ (11). Cependant, il est à noter que les distributions des erreurs sont approximativement similaires (plus fréquentes en modalité audio seule et dans la structure VCV) lorsque l'on observe chaque contexte vocalique.

### 3.3.2 Tâche de production

#### Taux de réussite

Sur l'ensemble des 960 items analysés, les apprenants ont correctement prononcé 87,2 % des cas, soit 837 items bien produits. Les scores sont significativement plus élevés en modalité AV (92,5 %) qu'en modalité A (82,05 %) ( $z = -2.34, p < .05$ ). Globalement, les apprenants parviennent à mieux produire les consonnes

cibles en modalité AV qu'en modalité A. C'est particulièrement notable pour la consonne /v/, mieux prononcée par les sujets testés en modalité AV (86,25 %) que par ceux en modalité A (69,37 %) ( $z = -2.11$ ,  $p < .05$ ). Une tendance similaire est observée pour /f/ (99,3 % vs 95,6 %) et /w/ (91,82 % vs 81,13 %) mais la différence n'est pas significative pour /f/ ( $z = -1.72$ ,  $p = .08$ ), ni pour /w/ ( $z = -1.78$ ,  $p = .07$ ). A l'intérieur de chaque modalité, /f/ est significativement mieux produite que /v/ et /w/, en modalité A ( $p < .001$ ) et aussi en AV ( $p < .05$ ). La production de la labiovélaire /w/ semble plus réussie que celle de /v/ en modalité AV (91,82 % vs 86,25 %), mais cette différence n'est pas statistiquement attestée ( $z = -1.6$ ,  $p = .2$ ), probablement en raison de la grande dispersion des scores de /w/.

En général, les scores de prononciation correcte dans chaque contexte vocalique est supérieur en modalité AV qu'en modalité A. Cependant, seules les différences significatives sont attestées dans les contextes /a/ (90,9 % vs 75,8 % ;  $z = -2.21$ ,  $p < .02$ ) et /e/ (93,3 % vs 81,7 % ;  $z = -1.97$ ,  $p < .05$ ), tandis que celles dans les contextes /i/ (94,1 % vs 89,9 % ;  $z = -0.92$ ,  $p = .35$ ) et /ɛ/ (91,7 % vs 80,8 % ;  $z = -1.7$ ,  $p = .07$ ) ne le sont pas. Une grande dispersion des scores est constatée en modalité A pour tous les contextes vocaliques.

En ce qui concerne les structures syllabiques, les scores de production correcte sont sensiblement meilleurs en modalité AV, que ce soit en CV (95,4 % vs 84,6 %) ou en VCV (89,6 % vs 79,5 %). Cependant, seule la différence en CV est statistiquement établie ( $z = -2.57$ ,  $p < .01$ ), contrairement à VCV ( $z = -1.77$ ,  $p = .07$ ). Il est à noter qu'une moindre variabilité est observée dans les scores lorsque les apprenants ont accès aux stimuli audiovisuels. Par ailleurs, les apprenants semblent également mieux réussir lorsque la consonne cible est en position initiale absolue (CV), avec des scores plus élevés qu'en VCV. Cette tendance est statistiquement significative en AV ( $z = 2.3$ ,  $p < .02$ ), mais non en modalité A ( $z = 1.5$ ,  $p = 1.3$ ).

### **Analyse des erreurs**

Parmi les 122 items mal produits, les difficultés rencontrées sont nettement plus fréquentes en modalité A (86 items) qu'en modalité AV (36 items). Les erreurs se répartissent entre les deux structures syllabiques, mais quelle que soit la modalité, elles sont moins fréquentes en CV (48 items) qu'en VCV (74 items). Concernant les contextes vocaliques, les erreurs sont beaucoup moins fréquentes lors qu'il s'agit de la voyelle fermée [i] (19 items). Les erreurs sont plus prévalentes en contexte [a] (40 items), mais moins fréquentes en modalité AV (11 items) qu'en modalité A (29 items).

Lorsque les apprenants ne parviennent pas à produire correctement les consonnes cibles, trois types d'erreurs apparaissent : la substitution, l'épenthèse et la suppression. Quelle que soit la modalité de test (A ou AV), la substitution est la plus couramment utilisée par les apprenants (93 occurrences, dont 73,1 % en A et 26,9 % en AV). L'épenthèse, moins fréquente (28 occurrences, dont 60,7 % en A et 39,3 % en AV), est appliquée uniquement lorsque la cible contient /v/ ou /w/. Les apprenants ajoutent [w] à la cible /v/ (ex. [vi] > [vwi], [ivi] > [iwvi]) et vice versa (ex. [ewe] > [ewve]). Quant à la suppression, elle n'est observée que dans un seul cas en modalité A, lorsque la cible [vi] est prononcée [i].

Indépendamment de la modalité, /v/ représente toujours la consonne posant le plus de difficultés chez les apprenants (58,2 % de l'ensemble des erreurs analysées), avec 49 occurrences en audio seule et 22 en audiovisuel. Dans la majorité des cas, cette consonne est substituée par /w/ dans 61,22 % des occurrences en modalité A (soit 30 occurrences sur 49) et dans 40 % des occurrences en modalité AV (soit 9 occurrences sur 22). Elle est également remplacée par son homologue sourd [f] (14 fois), la consonne battue [v] (7 fois) et l'approximante [ʋ] (4 fois).

En ce qui concerne /w/, elle est substituée par /v/ dans la majorité des cas (53,5 %), avec 23 occurrences sur 43, principalement en modalité A (15 occurrences) plutôt qu'en modalité AV (8 occurrences). Les erreurs les plus courantes incluent les séquences [vw] (39,5 %, soit 17 occurrences sur 43), avec une prévalence plus marquée en modalité A (12 occurrences) qu'en AV (5 occurrences). Les apprenants

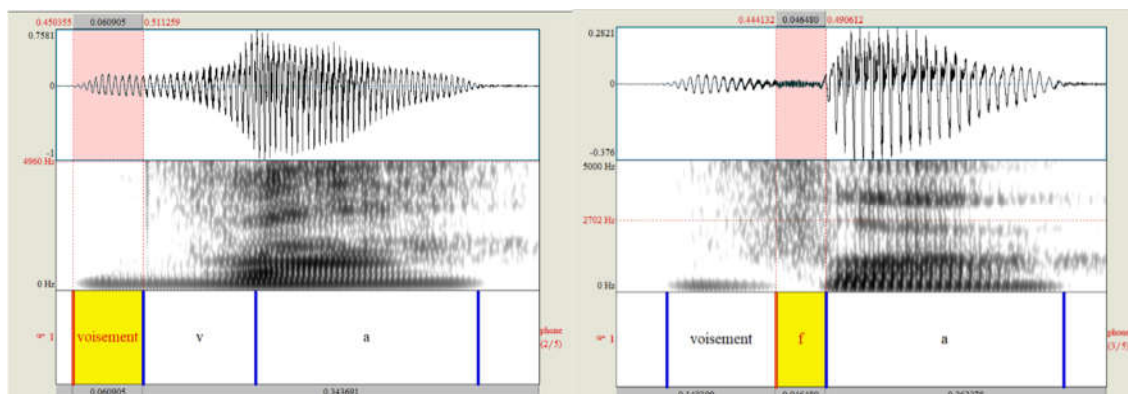
montrent une tendance à se tromper davantage entre [v] - [w] et [w] - [v] (dans les deux sens) dans les contextes de /ε/ (17 occurrences) et de /a/ (13 occurrences).

Il est à noter qu'une tendance se dégage : en modalité A, lorsqu'on présente /v/ comme référence, les apprenants commettent davantage d'erreurs avec /w/ (30 confusions) que lorsque /w/ est la référence (15 confusions).

### Stratégies de prononciation de /v/

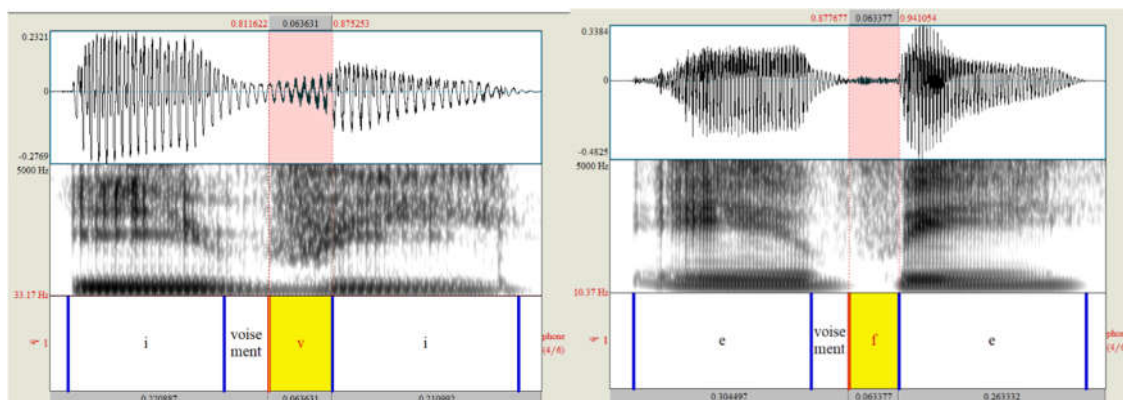
Sur les 320 items cibles contenant la consonne /v/, les participants prononcent correctement [v] dans 77,8 % des cas, soit 249 items. Leur performance est meilleure en modalité AV (86,25 %, soit 138 items sur 160) qu'en modalité A (69,37 %, soit 111 items sur 160). Parmi les 249 items où [v] est bien prononcé, deux stratégies distinctes ont été observées chez les participants : le pré-voisement et la décomposition des traits articulatoires (voisement + friction sourde) (cf. figures 3 & 4).

La première stratégie implique la mise en vibration des plis vocaux avant la réalisation de [v]. Cette stratégie est la plus fréquemment utilisée parmi les prononciations considérées comme « correctes » de [v] par les participants, que ce soit dans une structure syllabique CV ou VCV, bien que son utilisation soit majoritaire en CV. Il est à noter que lorsque les apprenants ne réalisent pas le pré-voisement avant [v], le bruit de friction est nettement moins intense.



**Figure 3.** Exemples de la réalisation du stimulus de type CV ([va]), avec deux stratégies, le pré-voisement avant [v] (à gauche) et le voisement suivi du bruit de friction sourd (à droite)

La deuxième stratégie consiste à produire le voisement avant la réalisation d'un bruit de friction sourd de type [f]. Cette méthode est moins courante que celle du pré-voisement mais utilisée par la plupart des participants. Nous avons fait écouter ces stimuli à trois locuteurs natifs et leur avons demandé d'identifier la consonne, sans leur montrer le spectrogramme. Les natifs ont tous identifié cette consonne comme étant /v/. En conséquence, ces items sont considérés comme étant une prononciation correcte de /v/.



**Figure 4.** Exemples de la réalisation des stimuli de type VCV, avec deux stratégies, le pré-voisement avant [v] (à gauche) et le voisement suivi du bruit de friction sourd (à droite)

### 3.4 Discussions

Les résultats de cette étude pilote permettent de confirmer l'hypothèse concernant l'impact des informations visuelles. Ils mettent, en effet, en lumière l'effet bénéfique des informations visuelles, démontrant une amélioration significative du taux de réussite de la consonne /v/ en modalité audiovisuelle, tant en perception (98,6 %) qu'en production (86,25 %) par rapport à la modalité auditive (respectivement 94,4 % et 69,37 %). Ces résultats corroborent les études précédentes d'Hazan et al. (2006), Wang et al. (2008, 2009) et Burfin (2015), qui mettent en avant la meilleure reconnaissance des phonèmes non natifs en modalité audiovisuelle.

La consonne native /w/ est également mieux perçue en modalité AV par rapport à la modalité A, avec un taux de réussite significativement supérieur en AV (97,9 %) comparé à la modalité A (93,7%). Bien que le score de prononciation correcte semble supérieur en AV (91,8 % vs 81,1 %), cette différence n'est pas statistiquement significative, probablement due à la dispersion des scores en modalité A. Les participants rencontrent des problèmes avec /w/, bien qu'elle soit une consonne native, particulièrement en modalité A où seules les informations auditives sont accessibles, engendrant davantage des confusions avec /v/, point sur lequel nous reviendrons plus loin.

Le temps moyen de réponse dans la tâche de perception est significativement plus court en modalité AV qu'en modalité A, indépendamment du type de stimuli (natifs /f w/ ou non natif /v/), du contexte vocalique (/i e ε a/) et de la structure syllabique (CV ou VCV). Ces résultats rejoignent ceux de Burfin (2015), montrant que le temps de réaction est nettement réduit en modalité AV par rapport à la modalité A, que ce soit pour les stimuli natifs ou non natifs. Le traitement des phonèmes, qu'ils soient natifs ou non, est plus rapide lorsqu'ils sont présentés avec une double information (audio et vidéo).

Quant au /f/, les résultats indiquent que les participants parviennent à le discerner et à l'articuler avec des scores élevés et comparables dans les deux modalités testées, sans différence significative entre elles. Ces résultats sont en conformité avec les deux études de Wang et al. (2008, 2009), montrant une capacité de reconnaissance élevée des fricatives de l'anglais, indépendamment de la modalité utilisée, chez les locuteurs natifs. Comparativement au phonème non natif /v/, les apprenants sont plus performants dans le traitement du phonème natif /f/, quel que soit le type de tâche. Ces résultats sont alignés avec l'Hypothèse de l'Analyse Contrastive (Lado, 1957), stipulant que les éléments similaires entre la langue étudiée et la langue maternelle seront plus simples à appréhender, tandis que les différences poseront davantage de difficultés. L'existence de /f/ dans les deux langues, thaï et français, explique la facilité des participants à le percevoir et à le prononcer, contrairement au phonème non natif /v/ étudié.

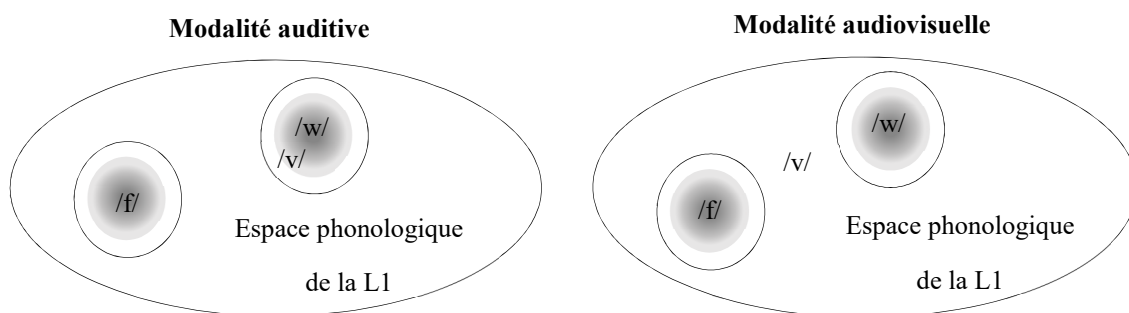
### Confusion entre /v/ et /w/

La confusion fréquente entre /v/ et /w/, rapportée dans d'autres recherches (Chunsuvimol & Ronnakiat, 2001 ; Ngammana, 2011 ; Le Corre, 2013 ; Sridhanyarat, 2017 ; Tusnyingyong & Tran, 2022), diminue considérablement en modalité AV. Tant en perception (passant de 28 à 3 occurrences) qu'en production (de 41 à 15 occurrences), cette réduction est notable indépendamment de la consonne utilisée en référence (que ce soit /v/ ou /w/), du contexte vocalique et de la structure syllabique.

Il est possible que cette diminution de confusion s'explique par la visibilité des visèmes de /v/ et /w/, présents dans la langue thaïe, facilitant ainsi leur distinction lorsqu'ils sont présentés à la fois en audio et en vidéo. Nous avons pu noter, lors du passage de l'expérience, que les participants ayant accès aux stimuli AV ont tendance à focaliser leur attention sur la bouche du locuteur natif à l'écran. Ils répètent également en chuchotant le stimulus de référence immédiatement après sa diffusion avant de produire la cible.

Le modèle *Perceptual Assimilation Model* (PAM) (Best et al., 1988; Best, 1995) se concentre sur la perception des phonèmes au sein d'un contraste par des individus inexpérimentés. Selon le PAM, en modalité audio seule, le phonème non natif /v/ est plus susceptible d'être confondu avec la catégorie native /w/ car l'information acoustique seule de /v/ ne suffit pas à le distinguer de /w/, induisant ainsi un taux de confusion plus élevé. Les deux, difficilement différenciables, sont assimilés à une seule catégorie /w/, suivant le schéma du type « *Single Category* ». Cependant, en modalité audiovisuelle, la confusion entre les deux est nettement moins fréquente. Les mouvements visuels des lèvres sont exploités, permettant aux participants de les distinguer comme deux catégories distinctes, suivant le type « *Two Categories* ». Ainsi, /v/ pourrait être perçu différemment de la catégorie native /w/ (cf. figure 5).

Il est pertinent de noter que la confusion entre ces consonnes est plus fréquente en position intervocalique VCV (56 occurrences) qu'en position initiale absolue CV (31 occurrences), indépendamment du type de tâche (perception ou production) ou de la modalité (audio ou audiovisuelle). Ce constat est en accord avec des études antérieures démontrant que la prononciation des phonèmes varie en fonction de leur emplacement dans la syllabe ou le mot (Fougeron & Keating, 1997; Keating et al., 1999 entre autres). En position intervocalique, les consonnes présentent une articulation complexe et moins distincte, ce qui altère leur perception et leur production, les rendant plus sujettes à confusion. La transition avec les voyelles atténue leurs caractéristiques acoustiques, réduisant le contraste phonétique et affectant la précision de leur perception et de leur production.



**Figure 5.** Représentation de /v/ dans l'espace phonologique de la L1 chez les apprenants thaïlandais lors de l'exposition à l'input auditif (à gauche) et à l'input audiovisuel (à droite)  
(adapté du modèle PAM de Best, 1995)

Concernant la perception de la consonne /v/, son taux de réussite demeure élevé, atteignant 99 % en modalité AV. Même en utilisant uniquement l'audio (modalité A), ce taux s'élève à 94 %. Ces résultats corroborent les études de Hazan et al. (2006) et de Burfin (2015) concernant la perception de /v/ et /b/ chez les locuteurs hispanophones. Bien que /v/ ne soit pas un phonème distinct en espagnol mais plutôt un allophone de /f/ en position finale, précédé d'une consonne voisée comme dans /afganistán/ [avɣanistán] (Hazan et al., 2006), les locuteurs hispanophones montrent de bonnes performances en perception, même en modalité audio. Ces études suggèrent que la présence de [v] comme allophone de /f/ en espagnol, combinée à la caractéristique labiodentale, permet aux hispanophones de différencier les lieux d'articulation entre /v/ et /b/. De façon similaire, dans notre étude, le taux élevé de réussite en perception de /v/ pourrait être attribué à l'existence de la fricative labiodentale /f/ en thaï. Ceci offre aux participants un avantage notable grâce à la caractéristique acoustique de la friction (en A) et au visème labiodental (en AV) pour percevoir les différences entre la labiodentale /v/ et la labiovélaire /w/.

Cependant, bien que les participants perçoivent correctement la consonne /v/ en modalité A (94 %), ils ne la prononcent correctement que dans 69 % des cas. Il semblerait que cette consonne non native est filtrée par le crible phonologique (Troubetzkoy, 1939), lequel perturbe sa production correcte chez les participants, principalement en la confondant avec /w/. Il est également important de noter que les fricatives labiodentales sont parmi les moins utilisées dans les langues à travers le monde, ne représentant que 3,3 % des occurrences dans la base UPSID (Vallée et al., 1999). En ce qui concerne le trait de voisement, les fricatives sourdes prédominent largement par rapport aux fricatives sonores : les langues présentent presque deux fois de plus de /f/ que de /v/. La rareté de /v/ est donc bien établie dans les langues du monde. Selon l'Hypothèse de la Différence de Marquage (Eckman, 1977), les éléments moins fréquents ou moins prévalents posent des difficultés d'acquisition. La marque influence ainsi la phonologie de l'interlangue. En tant que fricative peu répandue dans les langues du monde, /v/ représente un défi non seulement pour les apprenants thaïlandais, mais également pour ceux ayant d'autres langues maternelles (Silverman, 1992; Detey et al., 2005; Endarto, 2015).

L'analyse acoustique de la production révèle que la plupart des participants initient la vibration de leurs plis vocaux avant d'articuler le son [v]. Lorsque la synchronisation entre le voisement et l'articulation labiodentale au niveau supra-glottique n'est pas obtenue, les traits articulatoires se scindent en deux phases distinctes : le voisement suivi du bruit de friction sourd caractéristique de [f]. Les apprenants doivent ainsi déployer des stratégies spécifiques pour tenter de produire la fricative labiodentale sonore [v], une réalisation phonétique rare dans les langues à travers le monde.

## 4 Conclusion

Cette étude, s'inscrivant dans le contexte de l'enseignement et de l'apprentissage du français, met en lumière l'impact positif des informations audiovisuelles dans le traitement du phonème non natif /v/. Les résultats indiquent clairement que ces informations facilitent et accélèrent le traitement de la consonne /v/ en position initiale de syllabe chez les apprenants thaïlandais. Ces conclusions pourraient contribuer aux méthodes de remédiation pour la perception et la production de /v/ chez les thaïlandais, les aidant à déconstruire les associations erronées entre le graphème < v > et le son [w] acquis dans leur langue maternelle.

Dans une perspective future, étant donné la taille restreinte du groupe de participants dans cette étude préliminaire, élargir l'échantillon à une taille plus importante en incluant des niveaux de compétence variés, serait bénéfique pour consolider les conclusions sur l'impact des indices visuels dans l'acquisition du /v/. En outre, le modèle *Speech Learning Model* de Flege (1995) suggère que la même consonne dans une position syllabique différente peut être classée dans des catégories distinctes. Ainsi, il serait pertinent d'explorer la perception de la consonne /v/ en position finale de syllabe pour mieux comprendre dans quelle catégorie native elle est assimilée.

## Références bibliographiques

- Best, C. (1995). A direct realist view of cross-language speech perception. In *Speech Perception and Linguistic Experience : Issues in Cross-Language Research* (p. 171-204).
- Best, C., McRoberts, G., & Sithole, N. (1988). Examination of Perceptual Reorganization for Nonnative Speech Contrasts : Zulu Click Discrimination by English-Speaking Adults and Infants. *Journal of experimental psychology. Human perception and performance*, *14*, 345-360. <https://doi.org/10.1037//0096-1523.14.3.345>
- Boersma, P., & Weenick, D. (1992). *Praat : Doing phonetics by computer [Computer program]* [Version 6.2.06, retrieved 23 January 2022 from <https://www.praat.org>].
- Borel, S., Vaissiere, J., Lavrut, M., Noiret, A., Ambert-Dahan, E., & Sterkers, O. (2016). *Quels sosies labiaux pour les voyelles nasales du français ?* 120.
- Burfin, S. (2015). *L'apport des informations visuelles des gestes oro-faciaux dans le traitement phonologique des phonèmes natifs et non-natifs : Approches comportementale, neurophysiologique* [Thèse de doctorat, Université Grenoble Alpes (ComUE)]. <https://www.theses.fr/2015GREAS002>
- Calliope. (1989). *La Parole et son traitement automatique*. Masson.
- Chunsuvimol, B., & Ronnakiat, N. (2001). (v) is Really a Problem Sound for Thai Speakers. *Thammasat Review*, *6*(1), 177-195.
- Detey, S., Durand, J., & Nespoulous, J.-L. (2005). Interphonologie et représentations orthographiques. Le cas des catégories /b/ et /v/ chez des apprenants japonais de Français Langue Etrangère. *Revue PAROLE*, *34-35-36*, 140.
- Dumont, A., & Calbour, C. (2002). *Voir la parole : Lecture labiale, perception audiovisuelle de la parole*. Elsevier Masson.
- Eckman, F. R. (1977). Markedness and the Contrastive Analysis Hypothesis. *Language Learning*, *27*(2), 315-330. <https://doi.org/10.1111/j.1467-1770.1977.tb00124.x>
- Endarto, I. (2015). *Comparison between English Loanwords in Thai and Indonesian : A Comparative Study in Phonology and Morphology*.
- Field, T. M., Woodson, R., Greenberg, R., & Cohen, D. (1982). Discrimination and imitation of facial expressions by neonates. *Science*, *218*(4568), 179-181. <https://doi.org/10.1126/science.7123230>
- Files, B. T., Tjan, B. S., Jiang, J., & Bernstein, L. E. (2015). Visual speech discrimination and identification of natural and synthetic consonant stimuli. *Frontiers in Psychology*, *6*. <https://doi.org/10.3389/fpsyg.2015.00878>
- Flege, J. (1995). Second language speech learning : Theory, findings and problems. In *Speech perception and linguistic experience : Issues in cross-language research* (York Press, p. 229-273).
- Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728-3740. <https://doi.org/10.1121/1.418332>
- Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America*, *119*(3), 1740-1751. <https://doi.org/10.1121/1.2166611>
- Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346-363. <https://doi.org/10.1002/bimj.200810425>
- Istria, M., Nicolas-Jeantoux, C., & Tamboise, J. (1982). *Manuel de lecture labiale : Exercices d'entraînement* (Masson).
- Keating, P., Wright, R., & Zhang, J. (1999). Word-level asymmetries in consonant articulation. *UCLA Working Papers in Phonetics*, *97*, 157-173.
- Kuhl, P. K., & Meltzoff, A. N. (1982). The Bimodal Perception of Speech In Infancy. *Science*, *218*(4577), 1138-1141. <https://doi.org/10.1126/science.7146899>

- Lado, R. (1957). *Linguistics Across Cultures : Applied Linguistics for Language Teachers*. University of Michigan Press.
- Le Corre, C. (2013). *Etude de la prononciation des étudiants thaïlandais en apprentissage du français et majeure de français à l'université de KHON KAEN, Thaïlande*.
- Legerstee, M. (1990). Infants use multimodal information to imitate speech sounds. *Infant Behavior and Development*, 13(3), 343-354. [https://doi.org/10.1016/0163-6383\(90\)90039-B](https://doi.org/10.1016/0163-6383(90)90039-B)
- Meltzoff, A. N., & Moore, M. K. (1977). Imitation of Facial and Manual Gestures by Human Neonates. *Science, New Series*, 198(4312), 75-78.
- Nacaskul, K. (1979). *A note on English loanwords in Thai*. 151-162.
- Ngamma, P. (2011). *Les problèmes de la prononciation du français chez les lycéens thaïlandais* [Mémoire de Master]. Université Chulalongkorn, Thaïlande.
- Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26), 10598-10602. <https://doi.org/10.1073/pnas.0904134106>
- Promkesa, S. (2014). *Étude des problèmes de prononciation des consonnes fricatives du français par des apprenants thaïlandais et propositions de correction phonétique* [Mémoire de Master 2 Recherche en Sciences du langage]. Université Stendhal, Grenoble.
- Rose, Y., MacWhinney, B., Byrne, R., Hedlund, G., Maddocks, K., O'Brien, P., & Wareham, T. (2006). Introducing Phon : A Software Solution for the Study of Phonological Acquisition. *30th Annual Boston University Conference on Language Development*, 489-500. <https://doi.org/10.1016/j.nihms.2006.10.003>
- Silverman, D. (1992). Multiple Scansions in Loanword Phonology : Evidence from Cantonese. *Phonology*, 9(2), 289-328.
- Sridhanyarat, K. (2017). The Acquisition of L2 Fricatives in Thai Learners' Interlanguage. *3L The Southeast Asian Journal of English Language Studies*, 23, 15-34. <https://doi.org/10.17576/3L-2017-2301-02>
- Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. The MIT Press.
- Troubetzkoy, N. S. (1939). *Principes de phonologie (traduction française par J. Cantineau en 1949)*. C. Klincksieck.
- Tusnyingyong, S., & Tran, T. T. H. (2022). Pourquoi les apprenants thaïlandais confondent-ils souvent /v/ et /w/ du français ? Une étude pilote sur l'effet de l'orthographe. *SHS Web of Conferences*, 138, 08007. <https://doi.org/10.1051/shsconf/202213808007>
- Vaissière, J. (2006). *La phonétique*. Presses universitaires de France.
- Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America*, 124(3), 1716-1726. <https://doi.org/10.1121/1.2956483>
- Wang, Y., Behne, D. M., & Jiang, H. (2009). Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics*, 37(3), 344-356. <https://doi.org/10.1016/j.wocn.2009.04.002>