Perception de parole sifflée par des auditeurs naïfs hors du laboratoire : Une exploration de l'impact du contexte expérimental.

Olivier Crouzet[1] Anna Marczyk Buklaha[2] Alain Ghio[3] Yohann Meynadier[3]
Maxwell Yeoman[3] Philippe Biu[4]
(1) Laboratoire de Linguistique de Nantes (LLING)– Nantes Université / CNRS, France
(2) Laboratoire de Neuropsycholinguistique (LNPL), UT2J Toulouse, France
(3) Aix Marseille Université, CNRS, LPL, Aix-en-Provence, France
(4) Université de Pau et des Pays de l'Adour, Pau, France
olivier.crouzet@univ-nantes.fr, anna.marczyk-buklaha@uiv-tlse2.fr,
alain.ghio@univ-amu.fr, yohann.meynadier@univ-amu.fr,
maxwell.yeoman@etu.univ-tlse2.fr, philippe.biu@univ-pau.fr

RÉSUMÉ

Nous avons collecté des données perceptives d' occitan béarnais sifflé (*shiular d'Aas*) auprès d'auditeurs et d'auditrices naïves dans le cadre d'un projet associant des problématiques scientifiques et de revitalisation de la langue et de la culture. Cette double composante nous a amenés à mettre en place une expérience de perception destinée à être mise en œuvre lors d'une fête populaire en plein air. Nous avons parallèlement répliqué cette expérience auprès de 2 groupes d'étudiants / étudiantes dans le cadre d'un cours. Nous présentons une analyse exploratoire de la comparaison des contextes de collecte en nous intéressant à l'impact potentiel sur les comportements de réalisation de la tâche, d'une part en termes de données manquantes, d'autre part en termes de performances de reconnaissance, en étudiant plus particulièrement le décours temporel de ces mesures à travers le déroulé de la tâche.

ABSTRACT

Whistled-speech perception by naive listeners during a field experiment : An exploration of the impact of the experimental situation.

Perceptual data from naïve listeners were collected in a perceptual experiment involving whistled speech (Whistled Occitan, *Shiular d'Aas*) within the framework of a research project also involving a language and culture revitalization component. This combination led us to collect speech perception data from a field experiment that took place during a people's fair. The same experiment was also conducted with 2 groups of students in their respective classrooms. The present paper provides an exploratory analysis of the comparison between collection conditions, focusing on their potential impact on participants behaviors in terms of missing data and recognition performance, through their evolution over the time course of the experiment.

MOTS-CLÉS : Perception de la Parole, Parole sifflée, Auditeurs naïfs, Expérimentation hors du laboratoire, Revitalisation.

KEYWORDS: Speech Perception, Whistled Speech, Naive listeners, Field Experimentation, Revitalisation.

1 Introduction

La parole sifflée constitue un cas limite particulièrement fécond pour l'étude des mécanismes perceptifs du langage, en ce qu'elle repose sur une réduction drastique des informations acoustiques tout en permettant néanmoins la transmission d'un message linguistique structuré (Busnel & Classe, 1976; Rialland, 2005; Meyer, 2015). Dans le cadre d'un projet ayant pour objet à la fois l'étude phonétique et phonologique d'une variété de parole sifflée (Busnel & Classe, 1976) et la revitalisation de l'usage correspondant auprès de la population, nous avons mis en place une collecte de données de production et de perception destinée à être recueillie hors du laboratoire dans le contexte d'un événement grand public en plein air. Une partie de ce projet qui porte spécifiquement sur l'occitan béarnais sifflé (*shiular d'Aas*) repose sur la collecte de données perceptives. Dans un contexte de revitalisation, la réduction des informations acoustiques liée à la transposition de la forme modale orale vers la forme sifflée soulève une question centrale qui dépasse la seule description des mécanismes de production : celle des indices acoustiques pertinents pour la différenciation des catégories phonologiques et leur décodage effectif par des auditeurs et auditrices non entraînées.

Plusieurs études antérieures, réalisées en laboratoire, ont permis de mettre en évidence la capacité qu'ont des auditeurs et auditrices naïves à apparier sans entraînement ou apprentissage intensif des formes sonores sifflées avec des catégories linguistiques de type voyelle ou consonne (Meyer *et al.*, 2017; Ngoc *et al.*, 2020a,b). Si les études qui se sont intéressées aux aspects phonétiques et phonologiques de ces modalités de production de la parole évoquent toutes le principe d'une *transposition* des formants avec des propositions diverses (2nd formant spécifiquement ou convergence de 2 formants contigüs entre F1 et F4, notamment) pour les variantes sifflées de langues non tonales (Busnel & Classe, 1976; Rialland, 2005; Meyer, 2008, 2015), les indices phonétiques précis utilisés pour caractériser ces formes linguistiques et servir de support aux contrastes phonologiques restent encore à explorer plus en détails.

Dans la perspective d'une collecte de données phonétiques combinée à la revalorisation de cette modalité sifflée du béarnais, nous avons engagé une démarche de science participative incluant la collecte de données perceptives auprès du grand public à l'occasion d'une fête populaire en plein air[1]. Le changement de contexte (recueil de données en plein air, au milieu d'une foule) mais aussi de population dont est issue l'échantillon peuvent influencer la manière dont les participants et participantes accèdent à l'information acoustique ou s'ancrent dans la tâche par rapport à un recueil de données en laboratoire. Les conditions expérimentales pourraient partiellement s'apparenter à des situations de collecte de données en ligne dans lesquelles les comportements des sujets diffèrent de ce qu'on peut observer en laboratoire du point de vue des interruptions de collecte (*breakoff*), des non-réponses ponctuelles (ou données manquantes) et de l'engagement dans la tâche (Peytchev, 2009, 2011) mais pourraient aussi renforcer la motivation des participants (dimension identitaire culturelle, investissement émotionnel…). Ces divers paramètres pourraient impacter à la fois la quantité de données exploitables recueillies mais aussi la validité *qualitative* des données (en jouant par exemple sur le degré d'attention que les participants affectent à la tâche).

Collecter des données hors du laboratoire permet aussi d'accéder à des échantillons qui se distinguent potentiellement des échantillons auxquels on a accès en laboratoire. Ils peuvent dans certains cas être plus largement représentatifs de la population dans son ensemble mais pourraient aussi être associés à des caractéristiques individuelles spécifiques des participants (age, catégories sociales...). Ceci peut donc constituer un apport non négligeable à la généralisation des résultats scientifiques en faisant

1. La « fête du fromage », Laruns, Pyrénées Atlantiques, 4-5 Octobre 2025.

porter les arguments expérimentaux sur des populations plus diverses ou en tout cas différentes.

En outre, à l'instar de ce qui est observé dans le domaine clinique (Marczyk *et al.*, 2022), l'évaluation de ce qui est effectivement compris par les auditeurs dans des conditions d'écoute aussi proches que possible des situations d'usage réelles se heurte à plusieurs contraintes méthodologiques. Les épreuves perceptives en laboratoire sont le plus souvent centrées sur le décodage du signal en tant que tel et recourent fréquemment à des stimuli contrôlés, tels que des pseudo-mots ou des segments isolés dans des situations expérimentales très spécifiques (cabine insonorisée, participants isolés hors d'un contexte communicationnel, relation sujet-scientifique verticale…). Si ces approches permettent un contrôle fin des différentes variables expérimentales, elles s'éloignent néanmoins des conditions écologiques de la pratique de la parole, laquelle s'inscrit généralement dans des contextes ouverts, spontanés et fortement contextualisés. Par ailleurs, la parole sifflée se pratique communément en extérieur, dans des environnements écologiques spécifiques (forêt, montagne) et dans des conditions de scène sonore particulières (vocalisations animales, bruits environnementaux comme le vent, la pluie…).

Dans cette perspective, il est utile d'examiner dans quelle mesure un public non spécialisé – occitanophone ou non– est en mesure d'identifier des stimuli de parole sifflée dérivée de l'occitan (*shiular d'Aas*) dans un contexte d'écoute plus naturel que celui des études antérieures (Meyer *et al.*, 2017; Ngoc *et al.*, 2020a,b) et dans lequel les contraintes environnementales et contextuelles sont susceptibles d'affecter l'engagement dans la tâche et les performances perceptives mais reflètent aussi mieux les conditions d'usage de cette modalité linguistique.

Les résultats présentés dans cette contribution se focalisent sur une analyse des données manquantes et des performances globales de reconnaissance en cherchant à décrire le comportement des participants en termes de changement des performances de réponses au cours de l'expérience. Cette analyse semble centrale pour planifier une procédure ultérieure de collecte de données la plus fiable possible ainsi que pour analyser les données actuelles à la lumière d'une réflexion sur le contexte de collecte. Dans cette perspective nous comparons les résultats collectés lors de cette fête populaire aux données obtenues auprès de deux groupes d'étudiants dans leur classe. Ces analyses pourraient aussi contribuer à la réflexion pour la diffusion d'une telle expérience perceptive en ligne via les réseaux sociaux.

2 Méthode

Les données perceptives ont été recueillies à travers un questionnaire en ligne (*Wooclap*©) pour qu'il puisse être accessible sur des téléphones portables et permettre de conduire l'expérience soit à travers la diffusion des stimuli sonores via des écouteurs branchés sur le dispositif personnel de chaque participant, soit via une diffusion amplifiée (que ce soit en plein air ou en intérieur).

2.1 Stimuli

Les stimuli sonores ont été produits par un locuteur expérimenté de sexe masculin, bilingue français-occitan enseignant le *shiular d'Aas* (transposition sifflée de l'occitan, le dernier auteur). Chaque stimulus a ensuite été intégré dans un questionnaire à choix multiple avec 4 réponses possibles pour chaque question. Afin de limiter les abandons en cours de recueil, nous avions pour contrainte de chercher à rendre l'expérience la plus conviviale possible et la plus courte possible. Nous avons donc

réduit le nombre d'items à une petite quantité (10 items). Les 10 items étaient composés pour partie de segments sans signification (4 items : respectivement 3 voyelles isolées et 1 consonne initiale de mot) et pour l'autre partie de mots ou de groupes de mots (6 items).

Pour chaque item, les choix possibles du questionnaire à choix multiple comportaient 4 réponses possibles. Au sein de chaque item, le nombre de syllabes / noyaux vocaliques de ces 4 choix possibles était proche mais pas systématiquement égal, ce qui pourrait potentiellement fournir des indices modifiant la probabilité de donner une bonne réponse au hasard. Pour les 4 items non-lexicaux, on compte 1 syllabe pour l'ensemble des solutions possibles. Pour chacun des 6 items lexicaux, les mots proposés comme solutions possibles comportent, dans la modalité orale de la langue (en Béarnais non sifflé), de 1 à 6 syllabes. Trois d'entre eux présentent des solutions qui ont le même nombre de syllabes (1 ou 2 selon l'item). Les 3 autres (items 5, 7 et 10) présentent des différences en nombre de syllabes qui sont plus ou moins marquées (3 ou 4 vs. de 2 à 4 vs. de 2 à 6). Ces derniers cas pourraient réduire l'incertitude associée et contribuer à de meilleurs résultats en termes de performance d'appariement, le taux théorique de bonnes réponses au hasard pouvant être amélioré par ce paramètre.

2.2 Participants

Les participants du groupe en extérieur ont tous participé sur la base du volontariat dans un contexte de fête populaire. Dans les 2 groupes d'étudiants, les participants avaient la possibilité de ne pas se connecter au dispositif mais étaient néanmoins plus « captifs ». Dans le groupe grand public (en extérieur), nous avons collecté les données associées à un effectif de 31 participants alors que les groupes en salle de cours (1 groupe de licence Orthophonie, 1 groupe de licence Sciences du Langage) ont donné lieu à l'enregistrement de données pour respectivement 28 et 19 participants.

2.3 Procédure

Dans les deux contextes, l'interface était affichée sur un grand écran et les sons étaient diffusés par le / la responsable de la collecte. Les stimuli étaient diffusés via des haut-parleurs afin de limiter la possibilité que de nombreux participants écoutent les stimuli avec le haut-parleur de leur téléphone, ce qui aurait pu conduire à un mélange sonore ajoutant des perturbations supplémentaires à celles qui sont inhérentes aux conditions d'expérimentation hors du laboratoire et aurait pu impacter massivement la perception des stimuli. Les participants se connectaient en parallèle via leur téléphone et remplissaient le questionnaire directement depuis leur appareil, ce qui permettait de collecter les données individuelles en parallèle de la diffusion collective.

Une fois la personne connectée sur l'interface, elle était amenée à répondre à quelques questions sur ses connaissances (musicales, du béarnais, du béarnais sifflé) et l'expérience commençait. Pour chaque item, les 4 choix étaient affichés à l'écran et les personnes pouvaient écouter la réalisation sonore sifflée à volonté avant de répondre, ce qui était contrôlé par l'expérimentateur / expérimentatrice en fonction des demandes de l'auditoire. Il n'était pas possible de modifier la réponse après l'avoir validée. En fin d'expérience, chaque participant pouvait revenir au début pour avoir un retour d'information sur les bonnes réponses tout en réécoutant les stimuli mais sans modifier ses réponses. Pour chaque item, les participants avaient la possibilité de ne pas donner de réponse, ce qui était enregistré comme une absence de réponse (donnée manquante). Les items étaient toujours présentés dans le même ordre. Les 3 premières questions portaient sur les voyelles isolées.

3 Résultats

Par manque de place, nous ne présentons pas les résultats de performance globaux mais nous concentrons sur l'évolution des réponses au cours de l'expérience. L'analyse des données a été réalisée avec R (R Core Team, 2025). Nous nous penchons d'abord sur l'analyse des données manquantes, qui correspondent à l'absence de réponse pour un item donné. Dans un second temps, nous explorons les performances de reconnaissance correcte.

Afin d'estimer le degré de généralisation des résultats observés, nous procédons à des analyses inférentielles fondées sur la méthode du *bootstrap* (DiCiccio & Efron, 1996; Good, 2005), qui est appropriée pour de petits échantillons[2] (Bibliothèque `boot`, Canty & Ripley, 2024). Les analyses ont pour objectif d'évaluer dans quelle mesure la pente de la droite de régression caractérisant respectivement l'évolution des données manquantes ou des réponses correctes est significativement supérieure à 0 (accroissement des données manquantes au cours du temps) ou inférieure à 0 (dégradation de la performance au cours du temps). Les décisions statistiques portent à la fois sur le seuil de significativité (*p-value* issue des simulations de bootstrap) et sur le fait que l'Intervalle de Confiance de bootstrap basé sur les percentiles inclut ou non la valeur 0. Nous considérons que la pente diffère significativement de 0 uniquement si les deux critères sont convergents. Notre objectif est d'évaluer :

1. Les conditions de collecte et caractéristiques des personnes constituant le groupe en extérieur influencent-elles les résultats et leur interprétation par comparaison aux groupes en salle ?
2. Les limites éventuelles liées à la quantité de données manquantes impactent-elles les résultats associés aux taux de reconnaissances correctes ?

3.1 Analyse des données manquantes

La distribution du nombre de données manquantes par participant / participante sur l'ensemble de l'expérience est très différente entre les deux contextes : si la plupart des participants en classe ont peu voire très peu de données manquantes, les données récoltées en extérieur font ressortir une grande quantité de données manquantes chez un nombre non négligeable de locuteurs (28.1% de données manquantes dans le groupe extérieur et seulement 10.4% et 2.6% pour les 2 groupes d'étudiants). On constate que dans les deux groupes d'étudiants, la plupart des participants ont moins d'1/3 de données manquantes. Si on choisit ce taux comme seuil maximal de proportion de données manquantes par participant, 13 participants (sur 31) dépassent ce seuil dans le groupe extérieur alors que seuls 2 et 0 (sur resp. 28 et 19) participants dépassent ce seuil dans les groupes en salle de cours.

Dans cette perspective, nous avons calculé pour chaque item de l'expérience, le pourcentage de données manquantes sur l'ensemble des participants et avons représenté graphiquement l'évolution de ce taux en considérant l'ordre d'apparition des items au cours de l'expérience (cf. figure 1). La figure 1(a) représente ces données pour l'ensemble initial des participants alors que la figure 1(b) correspond à l'échantillon de participants n'ayant pas plus d'1/3 de données manquantes.

Pour les analyses inférentielles, nous nous concentrons sur la pente de la droite de régression obtenue

2. En effet, dans tous les cas les réponses individuelles sont fusionnées sur l'ensemble des participants ou des items pour estimer les effectifs de données manquantes ou de réponses correctes *par unité expérimentale*, ce qui aboutit donc à des effectifs limités pour les données sur lesquelles portent les analyses de pente. Ainsi, pour 10 items répartis sur la durée de l'expérience, on dispose de seulement 10 valeurs de pourcentage de données manquantes ou de bonnes réponses pour l'étude de l'évolution au cours du temps.

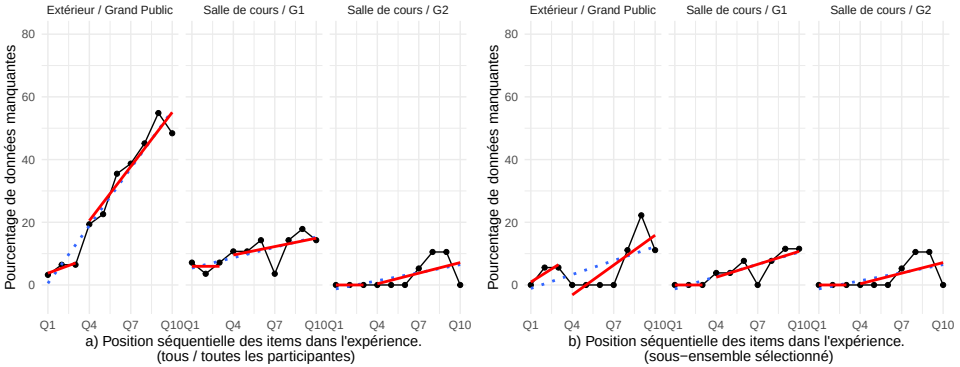

FIGURE 1 – Évolution du pourcentage de données manquantes par item en fonction de sa position séquentielle dans l'expérience. (a) Données de l'ensemble de l'échantillon. (b) Données restreintes aux participants qui n'ont pas produit plus d'1/3 de données manquantes. Chaque point représente un item. Pourcentage par item en fonction de sa position séquentielle. Droite de régression bleue en pointillés pour l'ensemble des points de mesure. En lignes continues rouges, les droites de régression séparées pour, resp., les voyelles isolées (items 1 à 3) et les autres items (4 à 10). Les analyses inférentielles présentées portent uniquement sur la droite de régression associée aux 7 derniers items.

dans la seconde partie de l'expérience (les 7 derniers stimuli, excluant donc les voyelles isolées) afin de mieux caractériser cette tendance à l'accroissement des données manquantes indépendamment de la nature des stimuli. Les paramètres de régression ont été estimés par un algorithme classique (*Ordinary Least-Squares*) sur 33000 [3] échantillons de bootstrap aléatoires. Pour chaque test statistique, nous fondons notre interprétation statistique à la fois sur la valeur du seuil statistique de significativité (*p-value*) et sur l'Intervalle de Confiance à 95% ($CI_{95\%}$).

D'un point de vue descriptif, la pente de la droite de régression dans l'échantillon initial est nettement plus élevée dans le groupe extérieur que dans les 2 autres groupes, ce qui pourrait illustrer une tendance à ce que les non-réponses deviennent plus fréquentes en cours d'expérience dans ce contexte. Une fois retirées les données des participants ayant plus d'1/3 de données manquantes, les pentes des droites de régression semblent plus similaires, ce qui pourrait indiquer que les participants concernés ont des comportements de désinvestissement potentiel comparables à ceux des groupes en classe.

Cependant, les analyses statistiques inférentielles font ressortir un comportement systématiquement différent du groupe en plein air du point de vue de l'évolution temporelle des données manquantes en comparaison aux groupes en salle de cours, et ce aussi bien si l'on considère l'intégralité des données qu'une sélection aux participants les plus assidus. En effet, ces analyses font ressortir que dans le groupe en extérieur, la pente de la droite de régression est significativement supérieure à 0 ($p < 0.001$, $CI_{95\%}$ = [3.47, 7.8]) alors que pour les groupes en salle de cours les conclusions sont moins nettes pour l'un des deux groupes. Les valeurs du seuil p sont contradictoires (G1 : $p < 0.05$, G2 : $p = 0.093$). Par contre, les intervalles de confiance de bootstrap incluent systématiquement la

3. Nombre de réplications aléatoires au-delà duquel les conclusions pour les *p-values* deviennent stables. Les conclusions associées aux Intervalles de Confiance se stabilisent quant à elles très rapidement pour des tailles d'échantillons beaucoup plus réduites, dès le nombre le plus bas d'échantillons de bootstrap testés –en l'occurrence 1000–, les conclusions que l'on peut tirer des Intervalles de Confiance ne changent pas. Pour des échantillons de 7 valeurs parmi 7 avec remise (selon les principes du bootstrap), il existe $7^7 = 823543$ échantillons distincts possibles.

valeur 0 pour la pente ($CI_{95\%}$ = [-0.67, 2.49] et [-0.51, 3.49], resp.).

Si l'on considère les données sélectionnées sur les participants qui présentent moins d'1/3 de données manquantes, on aboutit aux mêmes conclusions. Dans le groupe en extérieur, la pente de la droite de régression est significativement supérieure à 0 ($p < 0.05$, $CI_{95\%} = [1.48, 6.35]$). Pour les groupes en salle de cours, les valeurs du seuil probabiliste de significativité sont dans les deux cas inférieures à 0.05 et pourraient conduire à conclure à la présence d'une pente distincte de 0 dans chaque groupe. Néanmoins, les intervalles de confiance incluent toujours la valeur 0 pour la pente de la droite de régression ($CI_{95\%}$ = [-0.06, 2.62] et [-0.51, 3.49]), ce qui incite à pondérer la conclusion que l'on pourrait tirer du seuil de significativité.

Dans l'ensemble, que l'on supprime les participants présentant une proportion importante de données manquantes ou pas, la trajectoire temporelle du pourcentage de données manquantes reste similaire en termes inférentiels. Si l'on pondère les *p-values* par les conclusions dérivées des Intervalles de Confiance, il semble raisonnable de considérer qu'il n'y a pas de preuve suffisante que la pente soit distincte de 0 dans les deux groupes en salle de cours alors qu'au contraire elle est systématiquement positive pour les données du groupe en extérieur. Le pattern d'accroissement des données manquantes en cours d'expérience semble donc particulièrement marqué dans le groupe en extérieur et se distingue fortement des deux autres groupes.

3.2 Analyse des performances d'identification

Pour cette partie de l'analyse, afin d'adopter un calcul conservateur et réaliste, les données manquantes sont considérées comme des réponses incorrectes. Nous comparons de même l'ensemble de l'échantillon avec la sélection des participants ne présentant pas plus d'1/3 de données manquantes.

L'évolution des taux de reconnaissance correcte en fonction de la position des items dans la séquence expérimentale est représentée dans la Fig. 2. De manière générale, on n'observe pas de chute systématique des performances au cours de l'expérience mais les pentes des droites de régression qui sont associées au 7 derniers items de l'expérience font toutes ressortir une tendance assez marquée à une réduction de la performance au fur et à mesure du déroulement de l'expérience. Cette tendance pourrait être liée aux items 7, 8 et 9 qui sont systématiquement moins bien catégorisés. Concernant les 3 premiers items, leur nombre réduit rend peu pertinent l'estimation d'une régression sur ces données, mais la pente semble plus aléatoire, ce qui pourrait de toutes façons indiquer une relative indépendance vis à vis du décours temporel de l'expérience.

Pour l'étude de l'évolution des taux de reconnaissances correctes sur les 7 derniers items, les analyses inférentielles ont été réalisées sur 2000 échantillons de bootstrap (les conclusions sont stables pour toutes les tailles d'échantillons testées). Malgré l'apparente diminution des performances dans les 3 groupes, celle-ci n'est significative dans aucun : ($p = 0.059$, $CI_{95\%}$ = [-7.22, 1.06] ; $p = 0.069$, $CI_{95\%}$ = [-15.48, 4.22] ; $p > 0.1$, $CI_{95\%}$ = [-23.15, 9.59]). On aboutit strictement aux mêmes conclusions en restreignant les données aux participants les plus assidus ($p > 0.1$, $CI_{95\%}$ = [-9.72, 5.56] ; $p = 0.054$, $CI_{95\%}$ = [-16.67, 4.22] ; $p > 0.1$, $CI_{95\%}$ = [-23.15, 9.59]).

Par contre, il apparaît (cf. Fig. 2) une tendance très nette à ce que les performances des participants du groupe extérieur sur ces items lexicaux se situent presque systématiquement en dessous du seuil de réponses correctes au hasard alors que les participants en salle présentent des patterns de réponses plus variés, et ce que l'on prenne en compte tout l'échantillon ou seulement le sous-ensemble présentant peu de données manquantes.

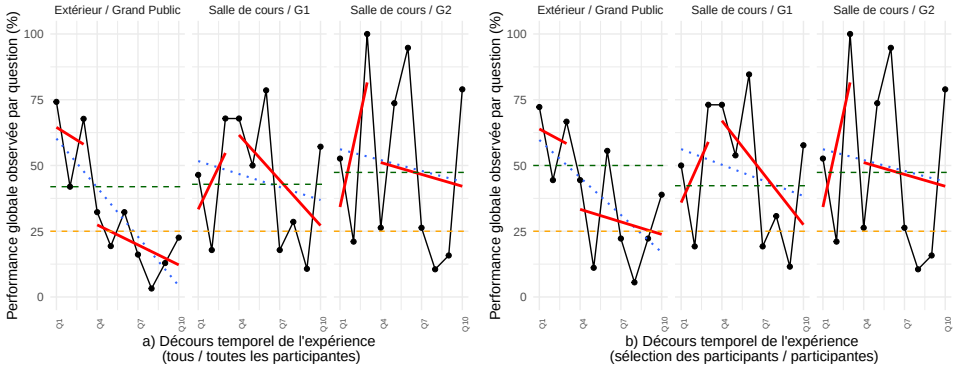

FIGURE 2 – Évolution de la performance de reconnaissance au cours de l'expérience. La ligne horizontale en pointillés orange correspond au taux théorique de réponses correctes au hasard (25%). La ligne horizontale vert foncé correspond au seuil au-dessus duquel le taux de bonnes réponses est significativement supérieur à ce seuil aléatoire pour un test binomial. (a) Données de l'ensemble de l'échantillon. (b) Données restreintes aux participants / participantes qui n'ont pas produit plus d'1/3 de données manquantes. Chaque point représente un item selon sa position séquentielle dans l'expérience. En pointillés bleus la droite de régression de l'ensemble des points de mesure. En lignes continues rouges, les droites de régression séparées pour, resp., les voyelles isolées (items 1 à 3) et les autres items (4 à 10).

4 Discussion

Sur la base des analyses présentées, il semble donc que le groupe pour lequel des données ont été collectées dans le cadre d'une fête en plein air présente un comportement de réponses qui diffère nettement des groupes d'étudiants en classe, se distinguant par (1) des données manquantes plus nombreuses, (2) une tendance significative à ce que le taux de données manquantes augmente au cours de l'expérience et (3) des performances de reconnaissances correctes quasi-systématiquement inférieures au seuil du hasard en dehors des voyelles isolées. Enfin, (4) ces conclusions se maintiennent si l'on fait porter les analyses uniquement sur les participants présentant une quantité limitée de données manquantes, ce qui semble mettre en évidence qu'une restriction des analyses à un sous-ensemble des participants ayant peu de données manquantes ne modifie pas les conclusions observées en termes de performance perceptive. Ces différences peuvent s'expliquer aussi bien par des conditions environnementales moins favorables en termes acoustiques ou attentionnels que par un investissement progressivement réduit dans la tâche mais aussi par le fait que les participants ne soient pas issus de la même population. Il conviendra de tenir compte de ces aspects et de les contrôler au mieux en collectant des informations individuelles complémentaires pour l'analyse finale des données.

Les observations décrites ici ont permis d'explorer certains points concernant les enjeux méthodologiques d'une expérience de perception de la parole conduite dans le cadre d'une approche participative de la science. Les tendances en termes de quantité de données manquantes et de leur évolution au cours du temps, et leur mise en perspective avec les performances de reconnaissance correcte tracent un portrait qui nous semble utile pour réfléchir aux choix à faire dans le cadre d'une expérience hors du laboratoire aussi bien en termes de planification de la méthode que de l'analyse des données.

Remerciements

Ce travail a été réalisé dans le cadre du projet ECO, soutenu par *Toulouse Initiative for Research's Impact on Society* (TIRIS), au titre de l'appel à projets « Co-Recherches avec la Société ». Les auteurs remercient également l'ensemble des participants, ainsi que les personnes ayant contribué à la collecte des données : Nina Roth, Théodose Peyusquet, ainsi que la direction et toute l'équipe du collège *Les Cinq Monts* de Laruns.

Références

BUSNEL R.-G. & CLASSE A. (1976). *Whistled Languages*, volume 13 de *Communication and Cybernetics*. Berlin : Springer-Verlag.

CANTY A. & RIPLEY B. D. (2024). *boot : Bootstrap R (S-Plus) Functions*. R package version 1.3-31.

DICICCIO T. J. & EFRON B. (1996). Bootstrap confidence intervals. *Statistical Science*, **11**(3), 189–228. DOI : 10.1214/ss/1032280214.

GOOD P. I. (2005). *Resampling Methods : A Practical Guide to Data Analysis*. Birkhäuser, 3rd édition.

MARCZYK A., GHIO A., LALAIN M., REBOURG M., FREDOUILLE C. & WOISARD V. (2022). Optimizing linguistic materials for feature-based intelligibility assessment in speech impairments. *Behavior Research Methods*, **54**(1), 42–53. DOI : 10.3758/s13428-021-01610-9.

MEYER J. (2008). Typology and acoustic strategies of whistled languages : Phonetic comparison and perceptual cues of whistled vowels. *Journal of the International Phonetic Association*, **38**(1), 69–94. DOI : 10.1017/S0025100308003277.

MEYER J. (2015). *Whistled Languages : A Worldwide Inquiry on Human Whistled Speech*. Berlin, Heidelberg : Springer Berlin Heidelberg. DOI : 10.1007/978-3-662-45837-2.

MEYER J., DENTEL L. & MEUNIER F. (2017). Categorization of Natural Whistled Vowels by Naïve Listeners of Different Language Background. *Frontiers in Psychology*, **08**. DOI : 10.3389/fpsyg.2017.00025.

NGOC A. T., MEYER J. & MEUNIER F. (2020a). Categorization of Whistled Consonants by French Speakers. In *Interspeech 2020*, p. 1600–1604 : ISCA. DOI : 10.21437/Interspeech.2020-2683.

NGOC A. T., MEYER J. & MEUNIER F. (2020b). Whistled Vowel Identification by French Listeners. In *Interspeech 2020*, p. 1605–1609 : ISCA. DOI : 10.21437/Interspeech.2020-2697.

PEYTCHEV A. (2009). Survey Breakoff. *Public Opinion Quarterly*, **73**(1), 74–97. DOI : 10.1093/poq/nfp014.

PEYTCHEV A. (2011). Breakoff and Unit Nonresponse Across Web Surveys. *Journal of Official Statistics*, **27**(1), 33–47.

R CORE TEAM (2025). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RIALLAND A. (2005). Phonological and phonetic aspects of whistled languages. *Phonology*, **22**(2), 237–271. DOI : 10.1017/S0952675705000552.