# Un changement de voix affecte t-il le processus de reconnaissance des mots parlés ?

*Sophie Dufour, Noël Nguyen*

Laboratoire Parole et Langage, CNRS et Université d'Aix-Marseille, Aix-en-Provence, France
5, Avenue Pasteur, 13604 Aix-en-Provence
sophie.dufour@lpl-aix.fr, noël.nguyen@lpl-aix.fr

## ABSTRACT

According to McLennan and Luce [1], variability in talker identity affects spoken word recognition when processing is slow and effortful. In the present study, we tested this hypothesis by manipulating the neighbourhood density of target words in a repetition priming experiment. Both for words with few and many phonological neighbours, the amount of priming for repeated words was not affected by a voice change. Such observation supports the claim that abstract representations exist and underlie spoken word recognition.

**Keywords**: Variability, abstract representations, episodic models, neighbourhood density.

## 1. INTRODUCTION

C'est avec rapidité et sans aucune difficulté que nous parvenons à reconnaître les mots et ceci malgré la forte variabilité présente dans le signal de parole. Un mot n'est jamais produit deux fois exactement de la même façon et présente des différences substantielles sur le plan phonologique et/ou phonétique selon le locuteur, le contexte phonologique (phénomène de co-articulation) ou encore le débit de parole. Chaque mot se matérialise ainsi par une infinité de formes sonores différentes que l'auditeur doit ramener à une entité lexicale unique. Un problème majeur auquel est confronté notre système de perception est donc de reconnaître une même production ou un même mot sous différents modes de réalisation.

Selon la théorie « abstractionniste », les mots dans le lexique mental seraient stockés sous la forme de séquences linéaires consistant en des traits [2], des phonèmes [3] ou des syllabes [4]. Le signal de parole serait dans un premier temps converti en une séquence de segments discrets écartant ainsi tous les détails acoustiques fins non pertinents pour l'identification, et serait ensuite projeté sur les représentations symboliques abstraites stockées en mémoire. Au contraire, selon les modèles « épisodiques » [5], les mots seraient stockés sous la forme de traces acoustiques détaillées encodant ainsi des informations fines liées par exemple à la voix du locuteur. Chaque mot serait alors associé à de multiples « tokens » et reconnaître un mot consisterait à trouver l'appariement le plus proche dans une vaste collection d'exemplaires.

Cette recherche fait suite à une étude récente conduite par McLennan et Luce [1] qui ont examiné l'impact de la variabilité acoustique liée à un changement de voix et de débit de parole (lent / rapide) sur le processus de reconnaissance des mots. Pour ce faire, ils ont utilisé le paradigme d'amorçage de répétition et manipulé la difficulté de discrimination entre des mots et des non-mots dans une tâche de décision lexicale ainsi que le format de réponse (immédiat/différé) dans une tâche de répétition de mots. Précisons que le paradigme d'amorçage de répétition consiste à présenter dans un premier temps un bloc de mots aux participants sur lesquels ils doivent réaliser une tâche (e.g. décision lexicale ou répétition). Dans un second temps, un second bloc de mots (bloc cible) leur est présenté, la moitié des mots ayant déjà été rencontré dans le premier bloc, l'autre moitié n'ayant jamais été rencontré. Typiquement, les mots répétés sont reconnus plus rapidement que les mots non répétés. Un tel effet résulterait de l'activation répétée de la même représentation lexicale en mémoire. L'atténuation de cet effet lors de la modification d'une dimension particulière (par exemple, un changement de voix) entre le premier et le second bloc indiquerait en accord avec les modèles « épisodiques » que le même mot prononcé par des voix différentes active différentes représentations lexicales et que des spécificités liées par exemple à la voix du locuteur seraient stockées en mémoire. Au contraire, aucune modulation de l'effet lors d'un changement de voix ou d'un débit de parole indiquerait en accord avec les théories « abstractionnistes » que le même mot prononcé de façon différente active la même représentation lexicale.

En tâche de décision lexicale, McLennan et Luce [1] ont montré une atténuation de l'effet d'amorçage de répétition lors d'un changement de voix ou de débit de parole lorsque la discrimination mots / non-mots était rendue difficile par l'utilisation de non-mots similaires à des mots (ex, bacov issue de bacon). Aucune atténuation dans l'effet d'amorçage de répétition n'a été observée lorsque la discrimination mots / non-mots était rendue facile par l'utilisation de non-mots ayant

peu de ressemblance avec des mots (ex, thushthudge). En tâche de répétition de mots, une atténuation de l'effet d'amorçage de répétition a été obtenue lorsque les participants devaient attendre l'apparition d'un signal pour donner leur réponse, mais pas lorsqu'ils devaient répondre immédiatement après l'apparition du mot. Suite à ces résultats, McLennan et Luce [1] en ont conclu que la variabilité dans le signal de parole liée à un changement de voix ou de débit affecte le processus de reconnaissance des mots parlés uniquement lorsque le traitement est lent et demande un certain effort. Notons que de tels effets sont compatibles avec des modèles dits hybrides [6] selon lesquels à la fois des informations spécifiques et abstraites seraient encodées en mémoire et où l'utilisation de l'une ou de l'autre type d'information dépendrait alors de la lenteur du traitement.

Dans cette étude, nous avons testé plus profondément l'hypothèse selon laquelle des effets liés à l'utilisation d'indices acoustiques émergeraient lorsque le traitement est lent et coûteux. Plutôt que de rendre difficile le traitement par le biais d'une manipulation de l'environnement lié à la tâche, nous avons directement manipulé la difficulté de traitement des mots eux-mêmes et utilisé comme McLennan et Luce [1] un paradigme d'amorçage de répétition dans lequel les mots étaient répétés soit par la même voix, soit par une voix de sexe différent. De façon à favoriser l'exploitation d'indices acoustiques liés à la voix du locuteur, les participants devaient réaliser une tâche de décision lexicale dans laquelle les non-mots ressemblaient fortement à des mots [1]. Comme, il est désormais bien établi que des mots ayant beaucoup de mots qui leurs sont phonologiquement proches (e.g. mots à forte densité de voisinage) sont reconnus plus lentement que des mots n'en ayant peu (e.g. mots à faible densité de voisinage) [7], la difficulté de traitement a été manipulée par le biais de la densité de voisinage phonologique. Notre hypothèse était que si les effets liés à l'utilisation d'indices acoustiques émergent lorsque le traitement est lent et coûteux une atténuation de l'effet d'amorçage de répétition devrait être observée au moins pour les mots difficiles à traiter et donc pour les mots ayant une forte densité de voisinage.

## 2. EXPÉRIENCE

### *2.1. Méthode*

#### 2.1.1. Participants

40 volontaires de l'Université de Provence ont participé à l'expérience. Tous étaient de langue maternelle française et n'ont rapporté aucun trouble de l'audition ou de la parole.

#### 2.1.2. Matériel

Quarante mots cibles de structure syllabique CVC ont été sélectionnés à partir de VOCOLEX [8]. La moitié d'entre eux résidaient dans une forte densité de voisinage et l'autre moitié dans une faible densité de voisinage. Le voisinage phonologique a été calculé en comptabilisant pour chaque mot cible le nombre de mots qui peuvent être générés par addition, délétion ou substitution d'un phonème quelle que soit sa position [7]. Les caractéristiques des mots cibles sont fournies dans le Tableau 1. 20 mots additionnels appariés aux mots cibles en fréquence, en nombre de phonèmes et en nombre de voisins phonologiques ont été également sélectionnés.

Afin que chaque mot cible soit vu dans la condition répétée et dans la condition non répétée, et qu'un même participant ne voit pas plusieurs fois le même mot cible, deux listes expérimentales ont été créées. Chaque liste était constituée de deux blocs de stimuli. Le premier (bloc 1) était constitué de la moitié des 40 mots cibles et des 20 mots additionnels. Parmi les 20 mots cibles, 10 étaient de forte densité et les 10 autres de faible densité de voisinage. Le second (bloc2) était constitué des 40 mots cibles, la moitié étant les mots présents dans le premier bloc et l'autre moitié étant alors des mots cibles contrôles non répétés. Les listes ont été contrebalancées de sorte à ce qu'un même mot cible serve à la fois en contrôle et en répétition. Pour les besoins de la tâche, 60 non-mots monosyllabiques de structure CVC ont été ajoutés dans chacune des listes et crées en changeant seulement le premier ou le dernier phonème de mots existants. 40 étaient présentés dans le bloc 1 et les 20 autres dans le bloc 2. Le bloc 2 comprenait également 20 non-mots du bloc 1.

De façon à manipuler le changement de voix, les 2 listes expérimentales ont été par la suite divisées en 4 sous listes chacune et se constituaient de la façon suivante : a) bloc 1 voix masculine, bloc 2 voix masculine, b) bloc 1 voix féminine, bloc 2 voix féminine, c) bloc 1 voix masculine, bloc 2 voix féminine, d) bloc 1 voix féminine, bloc 2 voix masculine.

**Table 1** : Caractéristiques des mots cibles.

| | Mots à faible densité | Mots à forte densité |
|---|---|---|
| Nombre de voisins phonologiques | 15 | 31 |
| Fréquence[1] | 58 | 55 |
| Nombre de Phonèmes | 3 | 3 |
| Durée[2] voix masculine | 565 | 565 |
| Durée[2] voix féminine | 565 | 565 |

Notes: [1] en nombre d'occurrences par million ; [2] en

millisecondes.

### 2.1.3. Procédure

Les stimuli ont été enregistrés par une locutrice et par un locuteur de langue maternelle française et ont été digitalisés à un taux d'échantillonnage de 22 kHz avec une résolution de 16 bits. Les participants munis d'un casque audio ont été testés individuellement dans une chambre insonorisée et les stimuli leur étaient présentés à un niveau sonore confortable. La présentation des stimuli était contrôlée par un ordinateur et les temps de réponse (TRs) étaient enregistrés à partir du début des stimuli. Pour chaque stimulus, les participants devaient indiquer le plus rapidement et le plus précisément possible si il constituait un mot ou non de la langue française, et devait fournir la réponse mot avec leur main dominante. La réponse du participant et le début de présentation du stimulus suivant étaient séparés par un délai de deux secondes. Les participants ont été testés sur une seule des sous listes expérimentales et ont commencé l'expérience avec 16 essais d'entraînement.

## *2.2. Résultats et Discussion*

Les temps de réaction obtenus dans le bloc 2 ont été analysés. Pour chaque participant, les temps de réaction supérieurs à 2,5 écart-types au-dessus et en-dessous de la moyenne des temps de réaction dans chaque condition ont été exclus des analyses. Adoptant ce critère seulement 1.06% des données ont été rejetées. Les réponses incorrectes ont été également supprimées des analyses. Les temps de réaction moyens obtenus en fonction du type de cible et du changement de voix sont représentés dans la Figure 1 pour les mots à forte densité et dans la Figure 2 pour les mots à faible densité. Les erreurs ayant été peu nombreuses (moins de 5%), les analyses ont été effectuées seulement sur les temps de réaction. Des analyses de variance (ANOVAs) par sujets ($F_1$) et par items ($F_2$) ont été conduites avec le type de cible (répétée, contrôle), la densité de voisnage (faible, forte) et le changement de voix (sans, avec) comme variables.

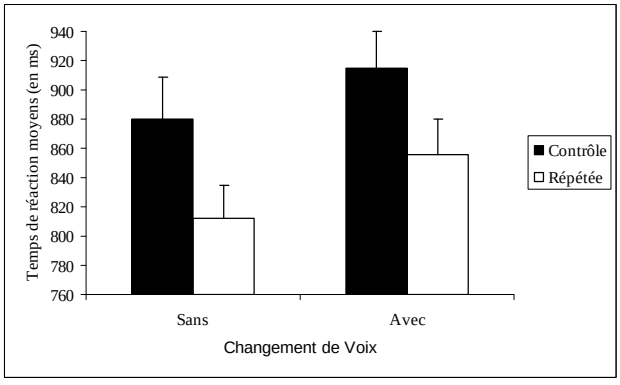


**Figure 1** : Temps de réaction moyens (en ms) en fonction du type de cible et du changement de voix pour les mots à forte densité (les barres représentent les erreurs standards).

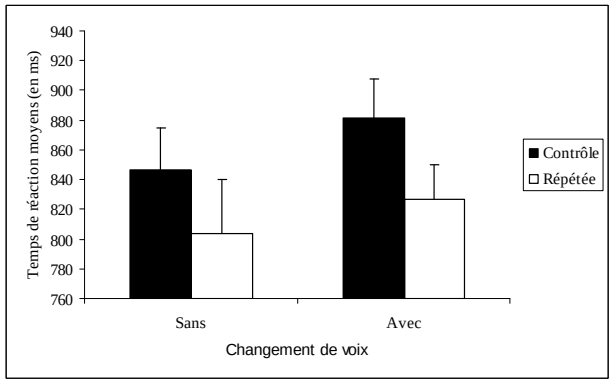


**Figure 2**: Temps de réaction moyens (en ms) en fonction du type de cible et du changement de voix pour les mots à faible densité (les barres représentent les erreurs standards).

Les temps de réponse étaient en moyenne plus rapides pour les mots cibles répétés (825 ms) que pour les mots cibles contrôles (881 ms). Cet effet était significatif à la fois par participants [$F_1(1, 38) = 66.30$, $p < .0001$] et par items [$F_2(1, 38) = 45.97$, $p < .0001$]. Les temps de réponse étaient en moyenne plus rapides pour les mots cibles résidant dans une faible densité de voisinage (840 ms) que pour ceux résidant dans une forte densité de voisinage (866 ms). Cet effet était significatif par participants [$F_1(1, 38) = 13.00$, $p < .001$] mais échouait à atteindre la significativité par items [$F_2(1, 38) = 1.82$, $p = .19$]. Les temps de réponse étaient en moyenne plus lents dans le cas d'un changement de voix (870 ms) que lorsque la voix restait identique (836 ms). Cet effet était signifiactif par items [$F_2(1, 38) = 18.28$, $p < .001$] mais pas par participants [$F_1(1, 38) = 0.87$, $p > .20$]. Que se soit pour les mots à forte ou à faible densité de voisinage, l'intéraction entre le type de cible et le changement de voix n'était pas significatif montrant ainsi aucune diminution de l'effet d'amorçage de répétition dans le cas d'un changement de voix [$F_1(1, 38) = 0.24$, $p > .20$; $F_2(1, 38) = 0.04$, $p > .20$ pour les mots à forte densité de voisinage; $F_1(1, 38) = 0.21$, $p > .20$; $F_2(1, 38) = 0.29$, $p > .20$ pour les mots à faible densité de voisinage.]

## 3. DISCUSSION GÉNÉRALE

L'hypothèse sous jacente à notre recherche était que si les effets liés à l'utilisation d'indices acoustiques émergent lorsque le traitement est lent et coûteux, une atténuation de l'effet d'amorçage de répétition devrait être observée pour des mots difficiles à traiter. De façon à manipuler la difficulté des mots, des mots à faible et à forte densité de voisinage ont été utilisés, les mots ayant beaucoup de voisins phonologiques étant généralement reconnus plus lentement que les mots ayant peu de voisins phonologiques [7]. Que se soit pour les mots à forte ou à faible densité de voisinage, aucune diminution dans la taille de l'effet d'amorçage de répétition n'a été observée lors d'un changement de

voix. Une telle observation argumente en faveur de l'existence de représentations abstraites et indiquerait qu'un même mot prononcé par différents locuteurs est susceptible d'activer la même représentation lexicale de base.

Comme nous l'avons vu précédemment, des modèles dit hybrides postulant la co-existence de représentations abtraites et détaillés ont été proposés [6]. En accord avec ce type de modèle, nous disposons dans la littérature de preuves expérimentales en faveur de l'un ou de l'autre type de représentations [9, 10]. L'existence conjointe de représentations abstraites et détaillées sous tendant la reconnaissance des mots parlés a clairement été mise en évidence dans l'étude de McClennan et Luce [1]. Comme ces auteurs nous avons favorisé l'exploitation d'indices liés à la voix des locuteurs en rendant difficile la discrimination entre les mots et les non-mots. Néanmoins dans notre étude aucun impact lié à un changement de voix sur le processus de reconnaissance des mots n'a été observé. Une différence entre notre étude et celle de McLennan et Luce [1] est relative à la durée de nos mots. En effet, ils ont été contrôlés de sorte à ce qu'il n y ait aucune différence de durée entre les mots prononcés par la voix masculine et ceux prononcés par la voix féminine. Cependant, bien que les auteurs précisent que leurs mots ont été jugés comme ayant été prononcés à un débit de parole normal, la durée moyenne des mots prononcés par la voix masuline et celle de ceux prononcés par la voix fémine différaient dans l'étude de McLennan et Luce [1]. Il se peut alors que certaines caractéristiques comme le débit de parole soient plus prépondérantes et aient plus d'impact sur le processus de reconnaissance des mots parlés. Notons en accord avec cette idée que Kittredge, Davis et Blumstein [11] ont récemment échoué à mettre en évidence une atténuation de l'effet d'amorçage sémantique lors d'un changement de voix avec des mots contrôlés en durée entre la voix masculine et féminine. D'avantage d'études sont donc nécessaires de façon à tester cette possibilité.

## BIBLIOGRAPHIE


[1] C.T. McLennan and P.A. Luce. Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31: 306-321, 2005.

[2] W. D. Marslen-Wilson and P. Warren. Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review*, 101: 653-675, 1994.

[3] J. L. McClelland and J. L. Elman. The TRACE model of speech perception. *Cognitive Psychology*, 18: 1 - 86, 1986.

[4] J. Mehler, J.Y. Dommergues, U. Frauenfelder, and J. Segui. The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, 20: 298–305, 1981.

[5] S.D. Goldinger. Echoes of echoes? An episodic theory of lexical access. *Psychological Review,* 105: 251-279, 1998.

[6] J. Pierrehumbert. Exemplar dynamics: Word frequency, lenition, and contrast. *In J. Bybee & P. Hopper (Eds.), Frequency effects and the emergence of lexical structure*. Amsterdam: John Benjamins, pages 137-157, 2001.

[7] P.A. Luce and D.B. Pisoni. Recognizing spoken words: The Neighborhood Activation Model. *Ear and Hearing*, 19: 1–36, 1998.

[8] S. Dufour, R. Peereman, C. Pallier and M. Radeau. VoCoLex : une base de données lexicales sur les similarités phonologiques entre les mots français. *L'Année Psychologique*, 102: 725-746, 2002.

[9] S.D. Goldinger. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22: 1166-1183, 1996.

[10] J.M. McQueen, A. Cutler and D. Phonological abstraction in the mental lexicon. *Cognitive Science*, 30: 1113-1126, 2006.

[11] A. Kittredge, L. Davis and S.E. Blumstein. Effects of Nonlinguistic Auditory Variations on Lexical Processing in Broca, *Brain and Language*, 97: 25-40, 2006.